Probabilistic Graphical Models for the Analysis of Omics Heterogeneity

by

Sahand Khakabimamaghani

M.Sc., Iran University of Science and Technology, 2010
B.Sc., University of Tabriz, 2007

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy

in the
School of Computing Science
Faculty of Applied Sciences

© Sahand Khakabimamaghani 2019
SIMON FRASER UNIVERSITY
Summer 2019

Copyright in this work rests with the author. Please ensure that any reproduction or re-use is done in accordance with the relevant national copyright legislation.

Approval

Name:	**Sahand Khakabimamaghani**
Degree:	**Doctor of Philosophy (Computer Science)**
Title:	**Probabilistic Graphical Models for the Analysis of Omics Heterogeneity**
Examining Committee:	**Chair:** Gregory Baker Senior Lecturer **Martin Ester** Senior Supervisor Professor **Ryan Morin** Supervisor Associate Professor **Leonid Chindelevitch** Supervisor Assistant Professor **Anoop Sarkar** Internal Examiner Professor **Fabio Vandin** External Examiner Associate Professor Information Engineering University of Padova
Date Defended:	**July 10, 2019**

Abstract

One of the biggest challenges in diagnosis, prognosis, and treatment of complex diseases like cancer is the heterogeneity of underlying disease mechanisms. This challenge has rendered the conventional and evidence-based medicine ineffective as a common remedy does not cure every patient with the same complex disease. The new paradigm in medicine, called precision or personalized medicine, is aimed at utilizing the new data collection technologies, such as high-throughput DNA sequencing, together with computational resources and algorithms, such as machine learning, to enable the scientists and physicians to understand the specifics of diseases for individuals and provide treatment strategies based on their personal characteristics.

In this thesis, we provide probabilistic graphical models to decipher the heterogeneity of diseases with an emphasis on cancer, using the recently available omics data from patients. We model the heterogeneity at two levels. First, we propose unsupervised and supervised biclustering methods for detecting heterogeneity at the level of a population of patients based on their genomic, transcriptomic and clinical characteristics. The provided frameworks are also theoretically applicable to other omics data types. Second, we provide a phylogenetic analysis method to analyze the heterogeneity of a population of cells of a tumor, i.e. intra-tumor heterogeneity, based on genomic data. By transferring the evolutionary information across different tumors, this method leverages the inter-tumor heterogeneity information to infer the intra-tumor heterogeneity of individual tumors with more certainty.

The proposed methods have promising performance when compared with the-state-of-the-art using both synthetic and real data.

Keywords: Probabilistic Graphical Models; Patient Stratification; Transcriptomic Heterogeneity; Tumor Heterogeneity; Bayesian Biclustering; Phylogenetic Analysis

Acknowledgements

I would like to express my deep appreciation to my senior supervisor Dr. Martin Ester for his extensive supports and guidance. He is not only a great supervisor, but also a kind and cheerful friend with whom I enjoyed spending time. I would also thank my supervisors, Dr. Leonid Chindelevitch and Dr. Ryan Morin. Dr. Chindelevitch is one of the most knowledgeable scholars that I have met and had the opportunity to learn from. His genuine supports, including his helps towards my internship, makes me indebted. Dr. Morin has always provided critical points and I have learned a lot from him during our fruitful conversations and collaborations. I would also thank Dr. Daniel Ziemek, who supervised my internship project at Pfizer, for his sincere supports.

I would also use this opportunity to appreciate my other great collaborators and coauthors: Dujian Ding, Bruno Grande, Dr. Yogeshwar Kelkar, Dr. Stacho Ladislav, Salem Malikic, Mehrdad Mansouri, Dr. Masoud Masjoodi, Oliver Snow, Jeffrey Tang and Olga Zolotareva. This thesis contains three out of seven projects accomplished through these collaborations.

During these years, I enjoyed the company and helps of many friends. Especially, I wish to thank (sorted alphabetically) Ali Arab, Raquel Aoki, Qinqyuan Feng, Tong He, Marten Heidemeyer, Hadi Moein, Boshra Nabaei, Hossein Sharifi and Beidou Wang.

Most importantly, I would like to thank my family, especially my parents, without whom I would not be able to come this far.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

In this chapter we motivate the use of probabilistic graphical models as the computational technique and discuss the importance of the analysis of omics heterogeneity as the biological problem considered in this thesis.

1.1 Probabilistic Graphical Models

Mathematical models reflect our understanding of systems. They describe the system elements and their relationships with each other. Given observed data about a system, models can be trained and used for making predictions.

Different types of systems require different modeling approaches [50]. Systems can be categorized into two classes. Deterministic systems can be clearly described in terms of the relationships between their elements subject to enough knowledge about the system. On the other hand, stochastic systems have randomness embedded in the relationship between their elements. Probabilistic models can help capturing a partially unknown deterministic system, a known deterministic system with noisy observed data, or a probabilistic system. Because biological systems are often complex, partially known and associated with noisy measurements, probabilistic modeling is a reasonable approach for studying these systems [120].

In probabilistic modeling, observations are modeled as random variables following a particular distribution. The values of some variables may influence the distribution of others. A Probabilistic Graphical Model (PGM) is a probabilistic approach for modeling the systems. It is a graph with nodes representing the variables of a system and edges representing the relationships between them. This representation increases the interpretability of the model and paves the way for applying graph theory concepts to the modeling problem [59].

One class of PGMs is Bayesian Networks (BNs). BN is a directed acyclic graph. The direction of edges indicate the flow of influence between variables. The variable at the tail of the edge influences or is a parameter of the distribution of the variable at the head of the edge. The structure of a BN determines the conditional decomposition of the joint probability of all model variables, i.e. the probability of a state of a model based on the values of the model variables in that state. In the joint probability decomposition, there is a factor per variable. Each factor consists of the corresponding variable and its parents (those variables that affect the corresponding variable and, thus, there is an edge from those to the variable). So, each factor is the conditional probability of the corresponding variable given its parents. The joint probability is equal to the product of all factors.

Parameters of variable distributions in a BN model are themselves considered as random variables with statistical distributions. This assumption provides the following advantages [21]:

- It allows to incorporate modeler's uncertainty about the parameters, and makes the application of rules of probability possible for inferring the parameters.

- Since the prior distributions that define the parameters are subjective, the modeler can feed his/her knowledge or "belief" about the parameters to the model. These beliefs about parameters are then revised after observing the data using Bayes' theorem.

- Another advantage is the ability to deal with the nuisance parameters, which are those parameters that are not of modeler's interest for inference and should not interfere with inferring other parameters. These parameters are integrated out from the joint probability.

- Bayesian statistics is predictive and provides the possibility to compute the conditional probability of one observation given the sample data.

However, Bayesian models are sensitive to the type of prior distributions that a modeler uses, i.e. modeler's beliefs. If the selected distributions are dissimilar to true distributions, the model will learn wrong parameters and make wrong predictions. Hyper-parameters of the prior distributions are also influential on the final inferences and should be selected with care. Hyper-parameter tuning methods (e.g. [119]) can be useful for this purpose. Moreover, computational costs associated with parameter learning makes BNs computationally more expensive than many deterministic models. Faster inference algorithms (e.g. variational methods) mitigate this problem to a degree, however there is sometimes a trade-off between speed and accuracy. These limitations should be considered when using BNs.

In this thesis, we use BNs for modeling the data measured for biological systems, known as "omics" data, to uncover patterns within these data to gain insight into the underlying biological mechanisms.

1.2 Omics Data and Heterogeneity

Perhaps, sequencing almost the whole human genome for the first time in 2001 [136] was the most crucial step towards the advancement of medicine. There has been a burst of omics data availability since the invention of DNA sequencing technologies. DNA sequencing that would initially cost billions and then millions can now be done for less than the cost of a single colonoscopy or magnetic resonance imaging (MRI) scan [84].

Reminiscent of "blind men and the elephant" story, each type of omics data captures a specific aspect of an individual's status [13]. Some of the most popular types of omics data are:

- Genome: "A genome is an organism's complete set of DNA, including all of its genes. Each genome contains all of the information needed to build and maintain that organism. In humans, a copy of the entire genome–more than 3 billion DNA base pairs–is contained in all cells that have a nucleus [132]." Exome, is the portion of genome that provides instructions for making proteins. The neucleotide content of exome is important as most known disease causing mutations occur in this area [133]. The following research areas have substantially benefited from whole genome sequencing (WGS) and whole exome sequencing (WES) at bulk or single cell levels [27]: (1) Cancer: A large number of cancer genomes have been sequenced through individual or collaborative efforts, such as the International Cancer Genome Consortium (http://www.icgc.org/) and The Cancer Genome Atlas (http://cancergenome.nih.gov/). Sequencing identifies somatic mutations that occur as a cancer develops. These mutations are not inherited or passed. (2) Hereditary genetic diseases: association between germ-line mutations (i.e. mutations that occur within a germ cell from either of the parents) and hereditary diseases can be studied when genetic information is available for a population. (3) Pharmacogenomics: Genetic information can be used to assign drug doses and reduce side effects [27].

- Epigenome: This refers to the potentially heritable chemical modifications to DNA and histone proteins that modulate chromatin structure and genome function. These modifications affect how the genome is expressed during different developmental stages and disease states or across different tissues [20]. Epigenetic alterations can be used as markers for cancer detection, diagnosis and prognosis. The enzymatic processes controling the epigenome provides therapeutic opportunities to reverse transcriptional abnormalities that are inherent to the cancer epigenome [12].

- Transcriptome: High-throughput whole transcriptome (cDNA) sequencing, abbreviated as RNA-Seq, has become a powerful tool for disease studies [27]. This technology provides the abundance of RNA from each gene as well as the genetic material of those RNA segments. The abundance of RNA indicates its actual activity and is closer to the real profile compared to the genomic sequence. The genetic material of RNA segments reveals more complex aspects of the transcriptome such as splicing isoforms and editing events, some of which are associated with cancer diagnosis and prognosis [27].

- Proteome: Proteome contains the sequences of proteins of an individual. This information is gathered using mass spectrometry technology, which can now quantify thousands of proteins in a single sample [27]. Proteome contains data that is even closer to phenotype compared to the transcriptome. Expressed mutations and editing events are among the information that can be extracted from proteome. The limitation of proteome, which decreases its popularity, is the low diversity of proteins that are quantified in a study.

- Metabolome: Mass spectrometry also generates metabolome profiles. Metabolome is the collection of small molecules known as metabolites [124]. This information is important for precision medicine as it reflects the real-time energy status as well as metabolism of the living organism. Also, some metabolites bind and directly regulate the activity of other biomolecules like kinases [78]. So, they can be targeted for therapy or measured for diagnosis and prognosis.

- Microbiome: Microbiome refers to the genome of the microbes living in individuals body. These microbes have essential functions in regulating growth and homeostasis and contribute to a significant fraction of our metabolome [13, 37]. "Emerging evidence suggests that the composition of a person's microbiome is a combination of innate immunity, introduction to organisms early in life, diet, and exposure to antibiotics and other environmental factors" [71]. The microbiome is associated to some brain diseases as well as response to therapies [13]. Researchers have been examining the microbiome in obesity, cardiovascular disease, cystic fibrosis, inflammatory bowel disease, skin disorders, cancer risk, and autism [71]. The limitation of this type of information is the size of microbiome data that is required to be gathered if an individual is going to be monitored throughout his/her lifetime. Given the dynamic plasticity and complexity of microbiome, this data might be orders of magnitude larger than genome data [13].

- Envirome: Sometimes the data types mentioned so far do not provide enough information for detecting the causal factors of a phenotype. For example, a study on three pairs of identical twins, one of each pair having multiple sclerosis, a disease known to have genetic components, failed to identify genomic, epigenomic or transcriptomic contributors [9]. In those cases, environmental factors (e.g., physical or psychological factors) could be involved in the causation of the disease. So, envirome is one of the data types that should be considered in precision medicine.

- Clinical Data: Clinical data are usually recorded in the form of electronic health records. These records might contain clinical test results, demographic data, life style data (e.g., nutrition, exercise, stress control and sleep [118]), etc. As will be discussed later, these data can be used for different types of prognosis and diagnosis contributing to treatment and prevention.

The high-throughput omics data, that are collected for individual patients can have two main contributions when analyzed with computational algorithms [27]: advancing our understanding of diseases and biological processes with unknown mechanisms, and when the mechanisms are clarified, helping to provide individualized health care through health monitoring, preventative medicine, and personalized treatment.

The remarkable heterogeneity of omics profiles poses challenges for understanding the disease mechanisms and treating them. At the higher level, the heterogeneity exists between the omics profiles of individual samples from patients, which we call *inter-sample* heterogeneity. For example, levels of expressions of a subset of genes might be significantly different between two samples resulting in transcriptomic heterogeneity. As another example, studies on sequenced genomes have revealed that every tumor is different with respect to the mutation profile and "driver" mutations.

At the lower level, heterogenity is also found between the cells in a sample taken from a patient. We call this *intra-sample* heterogeneity. For example, in a sample taken by tumor biopsy, tumor cells might constitute sub-populations with similar genomic or transcriptomic profiles [58]. These heterogeneous sub-populations complicate the treatment as some drugs might be effective only on part of these sub-populations, leaving others to proliferate even faster in the absence of rivalry. Methods that can detect this heterogeneity can inform the treatment process and provide insights into the evolution of these sub-populations from the cancer stem cell, i.e. the founding cancerous cell of the tumor.

Omics heterogeneity implies the existence of diverse disease mechanisms. Because this type of heterogeneity was previously unknown, conventional symptoms-oriented disease diagnosis and treatment was associated to several significant limitations including ignorance of preclinical risk factors, neglecting the underlying mechanisms of the symptoms, and broad disease descriptions which might include multiple disease with similar symptoms [27]. This approach over-simplifies the complex nature of most diseases [83]. Later, evidence-based medicine allowed for departure from that classic empirical paradigm. Although powerful and widely used, practice of evidence-based medicine also has limitations. In evidence-based medicine, data are collected from populations or large cohorts, from which mean values or figures are derived to infer recommendations [29]. Then these recommendations will be applied to all patients which is the "one size fits all" scenario which essentially ignores the outliers [13].

Recent technological, scientific, and social developments are likely to change the paradigm of medicine. Emergence of revolutionary, high-resolution, high-throughput data generating technologies, continuous innovations in information sciences and clinical bioinformatics, and empowering individuals by the proliferation of social media, ensure that we are living in perhaps one of the most profound periods of advancement in biology and medicine [13].

The ability to study biological phenomena at omics levels in turn is expected to make it possible for patients to be treated according to their own specific molecular characteristics [27]. For example, recently, "testing for specific genetic abnormalities has been transforming the classification and treatment of cancer. For example, in lung cancer, the traditional classification that is based on anatomic and histologic criteria is being augmented by molecular testing of EGFR, MET, RAS, ALK, and other genetic markers [71]." Moreover, it is now known that the same drug may have different effects on different individuals due to their personal genomic background and living habits [6, 85]. All of these direct to a shift from evidence-based medicine towards "precision medicine".

Precision medicine is defined as "treatments targeted to the needs of individual patients on the basis of genetic, biomarker, phenotypic, or psychosocial characteristics that distinguish a given patient from other patients with similar clinical presentations. Inherent in this definition is the goal of improving clinical outcomes for individual patients and minimizing unnecessary side effects for those less likely to have a response to a particular treatment" [71]. PM is aimed at providing quick, efficient, and accurate course of action for a patient [7].

1.3 Contributions

The goal of this thesis is to propose methods that leverage omics data to facilitate precision medicine. The first two methods discussed in this thesis are endeavors to model the disease heterogeneity at the level of a population of samples, i.e. inter-sample omics heterogeneity. This is done by grouping of samples into subtypes with similar omics characteristics. Assuming one sample per patient, this approach is also called "patient stratification". The first method, called B2PS (Bayesian Biclustering for Patient Stratification), simultaneously models multiple omics data types. To the best of our knowledge, this is the first unsupervised Bayesian approach that utilizes integrative biclustering for patient stratification.

Unsupervised detection of disease subtypes might produce subtypes that are irrelevant to any known phenotype [3]. This happens because clustering methods are usually driven by the strongest signal in the data which might correspond to undesired phenotypes (e.g., gender). Therefore, the second method, called SUBSTRA (Supervised Biclustering for Patient Stratification), uses a single omics data as the core, however incorporates clinical data to supervise the processes of patient stratification and provide information about the features that are relevant to the phenotype and subtypes. This provides interpretability and produces patient strata based on the relevant omics characteristics. The method can be used both for descriptive analysis (patient stratification, gene clustering and feature weighting) as well as predictive analysis (predicting the phenotype of a new patient).

While the first two projects capture omics similarities to discover subtypes of patients, the focus in the third method is on detecting similarity groups or subclones of cancer cells within a tumor of a patient. In other words, we are interested in heterogeneity at a lower level, i.e. heterogeneity among the cells within a sample rather than heterogeneity among the samples. This is known as intra-tumor heterogeneity in the literature (e.g. [52, 32, 76]). In addition, knowing the evolutionary relationships between the detected subclones is also desired. In the third method discussed in this thesis, we connect the two levels of heterogeneity by detecting the omics similarities among subsets of tumors in terms of evolution, and using them to more confidently model the heterogeneity within each tumor. The proposed method, called HINTRA (Collaborative Intra-Tumor Heterogeneity Detection), finds common evolutionary patterns of tumors of a specific disease using their genomic profiles. Then, it uses these patterns to resolve ambiguity for those tumors for which the genomic profiles imply the intra-tumor heterogeneity with less confidence. The novel approach used in HINTRA for modeling the evolutionary trees as well as its new Bayesian parameter learning method provide advantages over similar state-of-the-art methods.

1.4 Organization of this Thesis

In chapter 2, we briefly review the literature around the three problems discussed in this thesis: unsupervised patient stratification, supervised patient stratification and intra-tumor heterogeneity detection. We also discuss some of the existing gaps and sketch general ideas for filling these gaps.

In chapter 3, we define the problem of unsupervised patient stratification and introduce our solution, B2PS, for that. Then, we provide experimental results about using B2PS for finding suitable omics data types for patient stratification. Moreover, we compare the performance of B2PS with a popular patient stratification method.

Chapter 4 is dedicated to our solution to the problem of supervised patient stratification, called SUBSTRA. After defining the problem of supervised patient stratification, we describe SUBSTRA and provide experimental results using both synthetic and real data and evaluating the method from both descriptive and predictive aspects. We show that SUBSTRA achieves comparable predictive performance and superior descriptive results compared with existing supervised and unsupervised methods.

In chapter 5, we present HINTRA, a method for collaborative intra-tumor heterogeneity detection. First, the problem is defined followed by the description of the method. We then describe the results of experiments with synthetic data indicating that HINTRA outperforms the state-of-the-art methods. The results for real data are shown later, which are consistent with the existing domain knowledge.

Chapter 6 summarizes the presented methods and findings. We conclude the thesis by a discussion on limitations of our methods and future works.

Chapter 2

Literature Review

In this chapter, we review and categorize the existing methods related to each of the problems considered in this thesis: unsupervised patient stratification, supervised patient stratification and intra-tumor heterogeneity detection. While the goal in the first two problems is to capture omics heterogeneity among different samples to discover disease subtypes, the output of the third problem is the heterogeneity among the cells within a sample and the evolutionary relationship between the subclones of cells. However, in our collaborative approach to solve the third problem, as in the first and second problem, we also take into account the similarities of samples in terms of evolutionary trajectories. In this chapter, we do not limit our review to the collaborative methods as both collaborative and stand-alone approaches are relevant to us and we will later considers the merits of both in chapter 5.

In the next sections, a general definition is provided for each problem. Then, we explain the aspects with respect to which existing solutions differ. In particular, we include, among others, those aspects that are relevant to our contributions. The publications are listed according to their characteristics and some of them are briefly described. Finally, the gaps in the existing approaches are discussed.

2.1 Unsupervised Patient Stratification

Groups of patients with the same disease can be subdivided into different categories depending on the underlying mechanism of disease [140]. The disease mechanisms can be studied using omics data which provide us with the omic aberrations of individuals. When the subtypes and their specific characteristics are discovered, these can be used to design subtype-specific treatments. For example, four transcriptomic subtypes were detected for breast cancer in [97], each of which were associated with different genomic aberrations and druggable mutations.

Different methods are proposed in the literature for leveraging omics data to discover the diseases subtypes corresponding to different mechanisms. Most of the methods used

Reference	Probabilistic	Technique	Input	Non-param.
Verhaak et al. [137]	No (HC)	Clustering	Expression	No
Hochreiter et al. [56]	Yes (FA)	Biclustering	Expression	No
Shen et al. [116, 117]	Yes (FA)	Clustering	Multiple	No
Zhang et al. [146]	No (NMF)	Biclustering	Multiple	No
Hofree et al. [57]	No (NMF)	Biclustering	Mutation	No
Cho and Przytycka [28]	Yes (PGM)	Clustering	Multiple	No
Sun et al. [123]	No (SVD)	Biclustering	Multiple	No
Raykov et al. [106]	Yes (PGM)	Clustering	Clinical	Yes
Liu et al. [81]	No (CC)	Clustering	Multiple	No

Table 2.1: Existing patient stratification methods sorted by year of publication. Abbreviations used in this table: HC (Hierarchical Clustering) - FA (Factor Analysis) - NMF (Non-negative Matrix Factorization) - SVD (Singular Value Decomposition) - PGM (Probabilistic Graphical Model) - CC (Consensus Clustering).

for this tasks, which is also known as patient stratification, use an unsupervised approach. These methods can be categorized from different aspects:

- Modeling approach: Some works use non-probabilistic methods such as Singular Value Decomposition (SVD) [47] and Non-negative Matrix Factorization (NMF) [73]. Others use probabilistic models such as Plaid [72] and SAMBA [128].

- Unsupervised technique: Some of these methods use clustering, which only clusters patients based on the observed features, and others use biclustering, which performs clustering on both patients and features simultaneously or finds subsets of patients with similar values across a subset of genes.

- Inputs: Some methods use a single input type, such as transcriptomic profiles, while others integrate multi-omics data.

- Detection of the number of clusters: The methods that do not need the number of clusters as input and automatically detect that during the clustering process are called non-parametric.

Table 2.1 lists some of the existing patient stratification methods. Next, we briefly describe each of these methods. Then, we discuss their advantages and disadvantages.

2.1.1 Verhaak et al. [137]

This research uses a set of existing methods to provide one of the first stratifications of Glioblastoma Multiforme using TCGA data for 202 patients. First, the authors use multivariate analysis [90] to integrate gene expression data for the same patients from three different platforms by assuming that the three platforms as samples from the same distribution with a latent parameter and deriving that parameter as the true expression. Then, they use average-linkage hierarchical clustering as the basis for consensus clustering method provided in [95] to cluster the patients. They evaluate the clustering stability for different numbers of clusters between 2 and 10 and choose the number with the most stable clusters, which is 4 in this case. Then, they use copy number variation data to annotate each of the detected subtypes.

2.1.2 Hochreiter et al. [56]

The authors present FABIA (Factor Analysis for Bicluster Acquisition), a generative multiplicative model tailored to the special characteristics of gene expression data. In their model, they consider that real microarray datasets are not Gaussian distributed and have heavy tails after prefiltering. Therefore, they choose multiplicative modelling over additive modeling (as in previous biclustering methods) to account for heavy tails and and be able to model the multiplicative effects of real conditions as well as artificial preprocessing on gene expression levels. The authors define a generative model as follows:

$$X = \sum_{i=1}^{p} \lambda_i z_i^T + \epsilon,$$

where X is the input expression matrix, p is the number of biclusters, λ_i and z_i are the sparse prototype vector and the sparse factors vector of bicluster i which constitute the multiplicative model, and ϵ is the additive Gaussian noise. The above equation is very similar to factor analysis. However, they use sparse Laplacian priors for λ and z in contrast to commonly used Gaussian distribution, to account for heavy tails in gene expression distributions. This makes the likelihood analytically intractably. Therefore, they use a variational expectation maximization approach for learning the parameters of the model.

A novel consensus score is proposed to consider overlapping biclusters. This score is used for evaluating FABIA against other biclustering methods for 100 synthetic datasets. The results of experiments with three real gene expression datasets indicate the relevance of detected subtypes and gene clusters. The method is also applied to a drug design dataset to find compounds with similar effects on gene expression.

2.1.3 Shen et al. [116, 117]

This reference presents an integrative clustering method called iCluster. The model is based on factor analysis and is defined as follows:

$$\begin{aligned}
X_1 &= W_1 Z + \epsilon_1 \\
X_2 &= W_2 Z + \epsilon_2 \\
&\vdots \\
X_m &= W_m Z + \epsilon_m,
\end{aligned}$$

where X_1 to X_m are the available omics data types with dimensions $p_1 \times n$ to $p_m \times n$ for n samples, W_1 to W_m are the corresponding coefficient matrices of dimensions $p_1 \times K$ to $p_m \times K$ with K being the number of factors and Z is the latent variable indicating the sample cluster memberships and it is shared across all data types. The authors define priors for these variables as follows:

$$\begin{aligned}
Z &\sim \mathcal{N}(0, I) \\
\epsilon &\sim \mathcal{N}(0, \Psi), \Psi = \text{diag}(\psi_1, .., \psi_{\sum_i p_i}) \\
X = (X_1, .., X_m)' &\sim \mathcal{N}(0, \Sigma), \Sigma = WW' + \Psi,
\end{aligned}$$

where ψ_j is variance of feature j and $W = (W_1, .., W_m)'$. Then, an expectation maximization method is proposed. The E-step involves computing expected values of Z and ZZ' given X under the current parameter estimates $(W^{(t)}, \Psi^{(t)})$. The M-step uses those expected values to compute the parameters by maximizing expected value of the data log-likelihood plus a lasso regularization term promoting a sparse W. The final clusters are computed by applying K-means to the inferred Z. The authors also provide a simple method for selecting K by measuring the 'perfectness' of cluster separability based on $Z'Z$ matrix's block structure. iCluster is applied to breast and lung cancer subtype discovery using joint copy number and gene expression data.

2.1.4 Zhang et al. [146]

The authors propose a method for factorizing multiple data matrices simultaneously. They generalize the Non-negative Matrix Factorization (NMF) method proposed by Lee and Seung [75], which involves a multiplicative update rule for learning the basis and coefficient matrices. Then, a method for extracting the modules/biclusters with correlation across features of different data types is proposed. The method evaluates the statistical significance of local Pearson correlation (over the samples included in the bicluster) among the features of biclusters from different datatypes. They call these vertically correlated biclusters 'multi-dimensional modules'. The method is applied to ovarian cancer to extract significant features and subtypes based on multi-dimensional modules across DNA methylation, gene expression and miRNA expression data.

2.1.5 Hofree et al. [57]

The motivation behind this work is to reduce the extensive heterogeneity in the somatic mutation data to make it applicable to patient stratification. For this purpose, the authors propose Network-Based Stratification (NBS) which uses a gene interaction network to smooth the genomic profiles before applying stratification algorithms. First, somatic mutations for each patient are represented as a binary profile indicating single-nucleotide base changes or the insertions or deletions of bases. These profiles are projected onto a human gene interaction network (such as Pathway Commons [26] or STRING [127]). Then network propagation [135] is used to spread the influence of each mutation over its network neighborhood. NMF [73] is used to cluster the resulting matrix of 'network-smoothed' patient profiles into a predefined number of subtypes $k = 2, 3, ..12$. The process is repeated 1000 times for each k and final clustering is computed using consensus clustering [95]. NBS in applied to ovarian, uterine and lung cancer cohorts from TCGA to identify subtypes correlated with clinical outcomes such as patient survival, response to therapy or tumor histology. The authors also identify characteristic network regions of the subtypes based on NMF outputs.

2.1.6 Cho and Przytycka [28]

The authors present a method for integrating data at different levels of central dogma for patient stratification. The define two types of data: phenotypic data and the underlying causative features. The use gene expression as phenotypic data and mutation, copy number varaiation and miRNA expression as feature explaining that phenotypic data. First, a binary network of patient-patient similarity is constructed based on the pair-wise correlation between the gene expression profiles. Then a generative probabilistic model is defined in which both similarity network and explaining features are generated in a subtype-specific fashion. In the probabilistic model, the subtype indexes of patients follow a Dirichlet-Categorical distribution. The explaining features are generated based on categorical distribution with subtype-specific parameters. Finally, the patient similarity network links are generated based on the similarity between the subtype assignment latent vectors of each pair of patients. The authors provide a parameter learning method based on Gibbs sampling. The method is applied to TCGA Glioblastoma Multiforme dataset to derive the subtypes. The authors discuss the agreement between their subtypes and those provided by Verhaak et al. [137]. More interestingly, their model provides the explaining genetic features of each subtype.

2.1.7 Sun et al. [123]

This paper proposes a multi-view matrix decomposition approach that integrates clinical features with genetic markers to detect disease subtypes by maximising within-subtype consistency between the clinical and genetic dimensions of data. The method simultaneously identifies the clinical features that define the subtype and the genotypes associated with the subtype. It is based on sparse singular value decomposition (SSVD) [77] which is a single-view method with the following mathematical model:

$$\begin{aligned} \min_{\sigma,u,v} \quad & \|M - \sigma u v^T\|_F^2 + \lambda_u \|\sigma u\|_0 + \lambda_v \|\sigma v\|_0 \\ \text{s.t.} \quad & \|u\|_2 = 1, \|v\|_2 = 1, \end{aligned}$$

where M is the data matrix with rows and columns respectively corresponding to objects and features, u and v are singular vectors, σ is the corresponding singular value, $\|.\|_F$ indicates the Frobenius norm, $\|.\|_0$ and $\|.\|_1$ are the 0-norm and 1-norm respectively indicating the number of non-zero elements and the sum of absolute element values. This method identifies one bicluster at a time and the next bicluster can be found by removing the rows corresponding to the first bicluster and solving the above model again.

In this paper, the authors generalize the above single-view model to a multiple-view model using a shared cluster membership vector z as follows:

$$\begin{aligned} \min_{z,\sigma_i,u_i,v_i,i=1,\ldots,m} \quad & \sum_{i=1}^{m} \|M_i - \sigma_i (z \odot u_i) v_i^T\|_F^2 + \lambda_z \|z\|_0 + \sum_{i=1}^{m} \lambda_{v_i} \|\sigma_i v_i\|_0 \\ \text{s.t.} \quad & \|u_i\|_2 = 1, \|v_i\|_2 = 1, i = 1, ..., m, \ z \text{ is binary}, \end{aligned}$$

where $\odot$ is the element-wise vector product (Hadamard product). The authors design a fast optimization algorithm that alternates over the latent variables and learns one variable at a time by fixing the others.

The method is first evaluated on synthetic data indicating the proposed approach identified hypothesized subtypes and associated features outperforming five other biclustering and multi-view data analytics. Moreover, experiments with real-life disease data about cocaine use and related behaviors, the proposed approach identified clinical subtypes of a disease that differed from each other more significantly in the genetic markers.

2.1.8 Raykov et al. [106]

The authors provide a non-parametric Dirichlet process mixture model for clustering to overcome the disadvantages of K-means algorithm. Unlike K-means, the proposed algorithm named MAP-DP (maximum a posteriori Dirichlet process mixtures), automatically infers the natural number of clusters based on a Chinese restaurant process approach. It also can handle any type of data unlike K-means which is specific to continuous data. MAP-

DP can also separate the outliers and deal with missing data. The proposed expectation maximization approach for parameter learning is much faster than Gibbs sampling and produces slightly better results. They applied the algorithm to a cohort of ParkinsonâĂŹs disease patients using their clinical data.

2.1.9 Liu et al. [81]

This paper present a patient stratification method based on consensus clustering called Entropy-based Consensus Clustering (ECC). The authors formulate consensus clustering as an optimization problem with objective function

$$\max_{\pi} \sum_{v=1}^{r} U(\pi, \pi^{(v)}),$$

where π is the consensus partition, $\pi^{(v)}$ are the basic partitions, and U is the utility function measuring the similarity between the consensus and the basic partitions. They employ an entropy-based utility function for its fast convergence and high quality. They transform the above optimization problem into a modified K-means clustering problem, in which they construct a feature vector for each data point based on its membership status in basic partitions and use that feature vector for clustering. The distance between each data point and K centroids is measured using KL-divergence to account for the entropy based utility function. The authors also provide methods for handling missing data without imputation. ECC is tested using 110 synthetic and 48 real datasets and shows superior performance against the included benchmark panel.

2.1.10 Discussion

Although biclustering is proven useful for patient stratification [102, 100] there has not been enough attention paid to this approach in the current literature with only half of the methods using that technique ([56, 146, 57, 123]). More research on the applicability of this technique for patient stratification is required.

Integrating multiple data sources is an important direction towards more robust patient stratification and requires further investigation. Most of the attempts to integrative patient stratification has been focused on generalizing matrix factorization approach ([116, 117, 146, 123]). Exploring alternative approaches is guaranteed. Moreover, there is a lack of systematic validation approach in the literature. For example, Shen et al. [116] and Sun et al. [123] do not compare with existing methods. More importantly, the merits of the integrative approach compared to single-input patient stratification is not demonstrated in the literature. Although Sun et al. [123] compares multi-view and single-view SSVD methods, we believe that their results are not an indicator of superiority of the integrative method, but are the natural result of their experimental setup because they use a dataset for single-view SSVD that is not the basis for defining true subtypes. More research is required to answer this question.

In terms of the popularity of the data types, gene expression has been the most frequently used data type for the purpose of stratification. This is a natural choice due to closeness of gene expression data to the disease phenotype and larger coherence of these data compared to extremely heterogeneous genomic data such as point mutations. Gene expression can be seen as a result of genomic variations such as point mutations and copy number variations and already contains information about those variations. There are cases that genomic data can become of potential contribution: 1) in absence of gene expression data, where certain pre-processing (such as the one proposed in [57]) might be needed and 2) as addition to gene expression or other data, where additional supervision (such as using patient similarity network as in [28]) might be required. More research is required to investigate the applicability of heterogeneous genomic profiles, as they are, for patient stratification.

Probabilistic method usually return a probabilistic assignment of objects to clusters. This is more desirable for patient stratification, because first, it provides a model-based (rather than ad-hoc) approach to predict subtypes for new patients with unknown subtypes, and second, patients in one subtype often share features with patients in other subtypes and probabilistic assignments to subtypes capture these similarities and are more informative than strict assignments [28]. Stochastic methods used for training probabilistic are less prone to getting stuck in local optimums. In addition, probabilistic models allow for the introduction of prior knowledge into model. Finally, non-parametric probabilistic methods automatically detect the number of subtypes. Despite all these advantages, only half of the discussed methods use probabilistic modeling ([56, 116, 117, 28, 106]) and there is a need for more investigation on probabilistic stratification methods.

In chapter 3, we address some of these open issues and propose a novel Probabilistic Graphical Model (PGM), which we call B2PS (Bayesian Biclustering for Patient Stratification). To the best of our knowledge, B2PS is the first integrative Bayesian biclustering method for patient stratification. We also briefly investigate the applicability of different data types for patient stratification in that chapter.

2.2 Supervised Patient Stratification

One important challenge for precision medicine is to improve patient treatment based on molecular markers while simultaneously ensuring interpretability of the resulting signatures. In the previous section, we discussed unsupervised patient stratification methods. Many of these methods provide interpretable stratification relating the strata to their characteristic omics features. However, unsupervised methods only provide us with descriptive results indicating the strongest signals in the features and the produced strata are influenced by

Reference	Direction	Specifications	Phenotype	Interpretable
Gönen and Kaski [48]	S	PBMP	Binary (multi)	Yes
Graziani et al. [51]	P	PCSN	Survival	Yes
Ammaduddin et al. [4]	S	PBMP	Binary (multi)	Yes
Gligorijevic et al. [46]	S	DBMP	Binary (multi)	Yes
Duan et al. [38]	P	PCSN	General	Yes
Ross et al. [109]	P	PCSN	Image composition	Yes
Ahmad and Fröhlich [3]	S	PCSN	Survival	Yes

Table 2.2: Existing supervised patient stratification methods sorted by year of publication.

these stronger signals. As an example, transcriptional data is a popular and widely available data type to reveal underlying disease mechanisms and derive predictive or diagnostic signatures. In general, however, many of the thousands of measured transcripts will not be related to the desired phenotype (e.g. metastasis) directly but rather fulfill other biological functions. As the number of samples is generally small compared to the number of transcript, it is difficult to distinguish irrelevant measurements from relevant ones. This problem has led to irreproducible and noisy predictors in the past. Consequently, a key task is to reliably identify and weight transcriptional features based on their relevance to the target phenotype and use these weights for patient stratification in a predictive setting. We call this *supervised patient stratification.*

Multiple recent methods have been proposed for supervised patient stratification. These methods can be categorized with respect to the following aspects:

- Integration direction: whether they incorporate phenotype data into patient stratification or the other way around. We indicate the former by S and the latter by P.

- Modeling specifications: this aspect consist of general approach (**P**robabilistic or **D**eterministic), technique (**C**lustering or **B**iclustering), input type (**S**ingle input or **M**ultiple inputs) and clustering prior (**P**arametric or **N**on-parametric). We code the combination of these four characteristics within four letters. For example, PBSN indicates a **P**robabilistic **B**iclustering **S**ingle input **N**on-parametric method.

- Generality: whether they are specific to a particular phenotype or can be generalized to other phenotypes.

- Interpretability: whether they provide information about the relevance of omics features to the phenotype and detected strata.

Table 2.2 lists some of the recent relevant methods. Next, we briefly describe each of these methods and discuss their advantages and disadvantages.

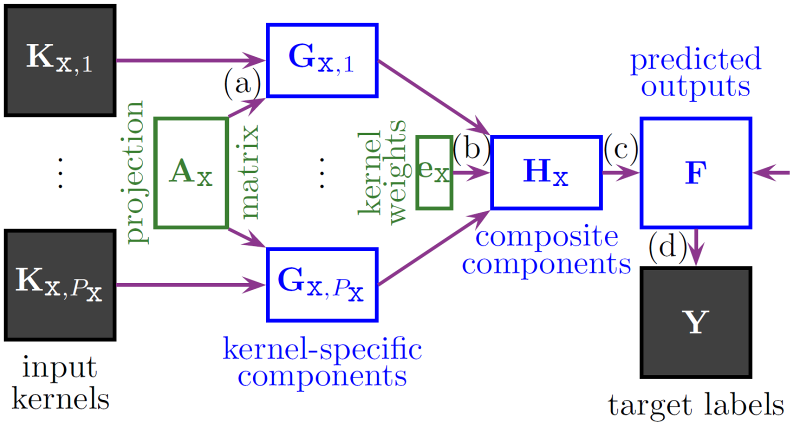


Figure 2.1: The model of KBMF [48] ©2014 IEEE. Latent variables are shown with white rectangles.

2.2.1 Gönen and Kaski [48]

The authors provide a kernelized Bayesian matrix factorization KBMF) method that can use multiple side information about the objects (both rows and columns). In other words, KBMF is a fully Bayesian extension to kernelized matrix factorization that can work with multiple side data in form of kernels. The kernels are computed as similarities based on either different data views or different notions of similarity between the objects.

The proposed model is shown in figure 2.1. N_X rows and N_Z columns are assumed. All P_X different kernels for the same rows indicated by $K_{X,i}$, $1 \leq i \leq P_X$, are first transformed to a lower dimensional representation $G_{X,i}$ with dimensions $N_X \times R$, R is the number of latent factors, after multiplication by a projection matrix A_X. The assumed distribution is $G_{X,i} \sim \mathcal{N}(A_X^T K_{X,i}, \sigma_g^2)$. Then transformed kernels $G_{X,i}$ are combined with each other with weights $e_X \in \mathbb{R}^{P_X}$ to generate the final composite row components $H_X \sim \mathcal{N}(\sum_{m=1}^{P_X} e_{X,m} G_{X,m}, \sigma_h^2)$. A similar process is applied to column kernels $K_{Z,j}$, $1 \leq j \leq P_Z$, to generate the final composite column components H_Z. Then, the predicted interaction matrix F is generated with distribution $F \sim \mathcal{N}(H_X^T H_Z, 1)$, which corresponds to factorizing F into two low-rank matrices. Finally, the observed interactions are defined as $Y \sim \delta(Y \odot F > \nu)$, where δ is the Kronecker delta function and ν is the margin parameter to remove ambiguity in the scaling and place a low-density region between the two classes (i.e. interacting and not interacting).

The authors propose an efficient variational approximation method for parameter learning. They evaluate their method on one toy dataset, two drug-protein interaction datasets, and 14 multi-label classification datasets.

2.2.2 Graziani et al. [51]

The authors present a subtype-specific method for predicting the clinical outcome. The method was used for predicting the efficacy of a targeted agent (e.g., a drug or an engineered T-cell). The assumption was that the effect of agent on outcome is mediated, at least in part, through some biomarkers (e.g., expression values). In this research, the clustering is performed based on the difference between pre- and post-treatment expression values of a selected set of genes (only *p-PDGFR* in this study on prostate cancer) indicated by X_i and Y_i for each patient i. X_i and Y_i are measured respectively n_i and m_i times for each patient i. Each instance of X_i or Y_i is assumed to follow a normal distribution with parameters specific to that instance. These parameters are assumed to be generated by a Dirichlet process specific to that patient. So, each patient has two Dirichlet processes associated with X_i and Y_i. The base distributions of the patient-specific Dirichlet processes are also assumed to follow a Dirichlet process, which forms a hierarchical non-parametric Dirichlet process.

For the predictive part, they first transformed X_i and Y_i into a single value $P(X_i < Y_i)$ using the parameters of the patient-specific Dirichlet processes (distributions of X_i and Y_i) inferred during the clustering. This is done using the vertical quantile comparison function, which is related to the Receiving Operating Characteristic (ROC) curve. This value is interpreted as the probability that the targeted agent (e.g. drug) affects the biomarkers (e.g., expression of certain genes). Then, they used this transformed value beside other covariates (hemoglobin and prostate-specific antigen levels) to train a Bayesian parameterized regression model for predicting the clinical outcome (overall survival time in this study). The descriptive and predictive parameters are trained simultaneously using Gibbs sampling.

2.2.3 Ammaduddin et al. [4]

This work proposes an extension to Kernelized Bayesian Matrix Factorization (KBMF) by Gönen and Kaski [48] discussed earlier. The method is called component-wise KBMF (cwKBMF). In KBMF, each kernel could contribute to the hidden representation with a kernel-specific weight $e_{X,i}$ (see figure 2.1). In this study, KBMF is extended by defining these kernel weights in a component-wise way to specify the extent that each kernel affects each component of the composite row and column components H_X and H_Z. Therefore, e_X is a matrix instead of a vector with e_X^s, $1 \leq s \leq R$ (R is the number of components) indicating the effect of different kernels on component s. Similar to [48], this variable follows a normal distribution with mean zero and a variance with a gamma distribution with the same dimensionality as e_X.

In this study, the rows of the input data matrix correspond to cell-lines and the columns corresponds to drugs. The values registered inside matrix indicate whether the corresponding drug has been effective on the corresponding cell-line. This study only uses kernelized side-data for cell lines (and not for drugs). The side-data used in this study consists of different views of the gene expression data of cell-lines. Each view corresponds to the expression of genes in a pathway from a set of selected pathways. The method can process several kernels for row and/or column side of the matrix to be factorized. Again, variational inference was used in this study.

2.2.4 Gligorijevic et al. [46]

This work proposes a framework based on graph-regularised non-negative matrix tri-factorization. This technique can be used for co-clustering heterogeneous datasets. Th method integrates somatic mutation profiles and drug-target interaction data using matrix tri-factorization regularized by transcript interaction and drug similarity data. This method simultaneously discerns patient strata and gene and drug clusters. The results can be used to perform driver gene prediction and drug re-purposing based on the identified strata.

The authors use the method originally proposed in [115] to simultaneously decompose both relation matrices into a product of three non-negative low-dimensional matrices. Assuming n_1 patients, n_2 genes and n_3 drugs, given are a patient-gene mutation matrix $R_{12} \in \{0,1\}^{n_1 \times n_2}$, a gene-drug interaction matrix $R_{23} \in \{0,1\}^{n_2 \times n_3}$, a network of gene interactions and a matrix of chemical similarities between the drugs. The objective for the graph-regularized non-negative matrix tri-factorization becomes:

$$\min_{G_i > 0, 1 \leq i \leq 3} [\|R_{12} - G_1 H_{12} G_2^T\|_F^2 + \|R_{23} - G_2 H_{23} G_3^T\|_F^2 + tr(G_2^T L_2 G_2^T) + tr(G_3^T L_3 G_3^T)],$$

where $L_2 \in \mathbb{R}^{n_2 \times n_2}$ and $L_3 \in \mathbb{R}^{n_3 \times n_3}$ are graph Laplacians of the gene interactions and drug similarity matrices, and G_1, G_2 and G_3 are respectively the latent factors of the patients, genes and drugs. Please not that because G_2 is shared between the first two terms of the objective function, G_1 and G_3 will be dependent. The method is applied to ovarian cancer and the identified patient subtypes are shown to be more related to the clinical data than those of the method proposed in [57]. Also, potential new driver genes are obtained and validated through enrichment analysis and literature. Finally, potential candidate drugs are identified for repurposing and validated through other dataset and literature.

2.2.5 Duan et al. [38]

The authors provide a decision tree ensemble method, called Bayesian Ensemble Trees (BET), for phenotype prediction. Data is partitioned first and each partition of data is used to construct a separate decision tree in this research. The partitions are learned such that the prediction preformance is optimized. The trees and partitions are learned simultaneously using a Bayesian approach. The authors define a probability distribution over possible decision trees consisting of probability distributions over the topology, decision features at each node, decision values for each decision feature, and the distribution of labels/values at each leaf. The observed data is assumed to follow a Dirichlet process with the distribution over trees as its base distribution. Stick-breaking process is used to form the Dirichlet process and the decision averaging weights. They provide methods for inferring the number of trees/partitions.

Gibbs sampler was used for posterior simulation to perform three steps: 1) tree fitting (given the data partitions, optimizing trees for each partition), 2) data reassignment (modifying the partitions given the optimized trees): they used a slice sampler in this step to gain faster convergence and reduce the number of trees, and 3) weight updating (the weights used for decision averaging are updated using stick-breaking sampler). The modeling also allows for variable importance inference following a method similar to random forests but considering the averaging weights. The authors applied BET to simulated and cystic fibrosis data. Using much smaller number of trees, they gained comparable accuracy to other popular methods like random forests.

2.2.6 Ross et al. [109]

This method incorporates phenotype data captured from images into disease subtyping based on clinical data to improve the performance. The authors introduce a Bayesian non-parametric model for subtyping. In their study, disease trajectories were acquired by transforming fractions of 6 different lung tissue types identified on mutiple CT scans of a patient in different points of time. The assumption was that each subtype of the disease exhibit distinct disease trajectory and these trajectories can be computed as a linear combination of clinical features.

The proposed probabilistic graphical model is shown in figure 2.2. The phenotypes Y are D-dimensional vectors and each element is predicted with specific feature weights for each of M features. Moreover, the feature weights applied to the clinical data for predicting the phenotypes are subtype-specific. Thus, the weights form a matrix $W \in \mathbb{R}^{M \times D \times \infty}$, where ∞ refers to the number of subtypes in the non-parametric setting. Clustering assignment is indicated by $z \in \{0,1\}^{N \times \infty}$. A set of constraints of type "must-link" guide the clustering, i.e. data instances that represent the same patient at different time points are forced to be in the same cluster. Stick-breaking process is used as part of the Dirichlet process mixture model for subtypes. Y and W follow Gaussian distributions and λ, the parameter for the prior of Y, is generated by a gamma distribution.

Variational inference is used for variable inference and the predictive performance of the model is evaluated for model selection. Also, the authors compare the predictive performance of their model with multivariate ordinary least squares regression.

2.2.7 Ahmad and Fröhlich [3]

Motivated by the fact that unsupervised clustering methods for subtyping might produce subtypes that are irrelevant to any known phenotype, the authors proposed a survival-based Bayesian clustering method. They incorporated survival data into patient stratification to

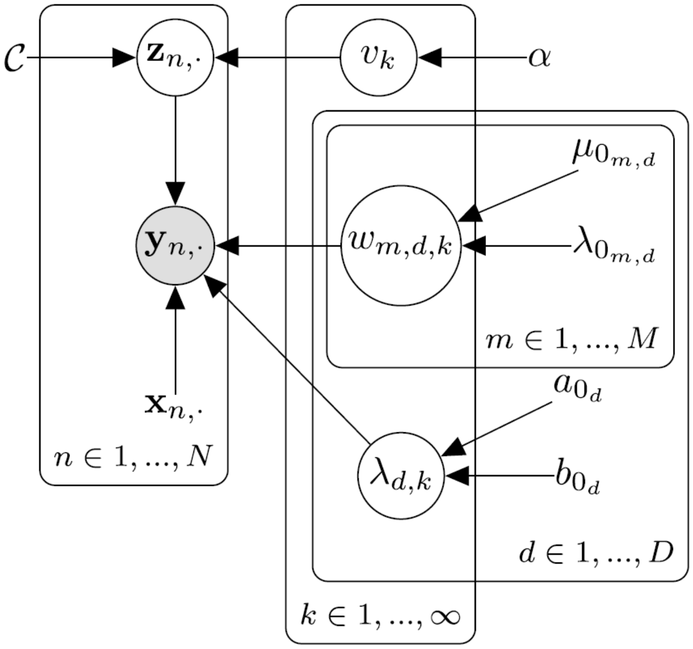


Figure 2.2: The probabilistic graphical model of Ross et al. [109] ©2017 IEEE.

improve the separability of disease subtypes with regard to their survival curves. They introduced a novel Hierarchical Bayesian Graphical Model, termed Survival-based Bayesian Clustering (SBC), which combines a Dirichlet Process Gaussian Mixture Model(DPGMM) with an Accelerated Failure Time (AFT) model to simultaneously cluster heterogeneous genomic, transcriptomic and time-to-event data.

The proposed model is shown in figure 2.3. Their DPGMM for modeling the expression data is as below:

$$X_i|c_i = j \sim \mathcal{N}(\mu_j, S_j^{-1})$$
$$\mu_j|S_j, \zeta, \rho \sim \mathcal{N}(\zeta, (\rho S_j)^{-1}),$$

where X_i indicates a D-dimensional vector of measurements (expression values) for patient i, μ_j and S_j are mean and precision of cluster j, and ζ is a hyper-parameter. For modeling the survival data, they use Accelerated Failure Time model with the log-normal assumption as below:

$$w_i|\beta_{0c_i}, \beta_{c_i}, X_i, \sigma^2 \sim \mathcal{N}(\beta_{0c_i} + \beta_{c_i}X_i, \sigma^2),$$

where $w_i = log(t_i)$ if survival time t_i is not censored, i.e. $I_i = 1$, and otherwise w_i is drawn from a left truncated normal distribution. Please note that the regression parameters β_{0c_i} and β_{c_i} are assumed to be cluster-specific.

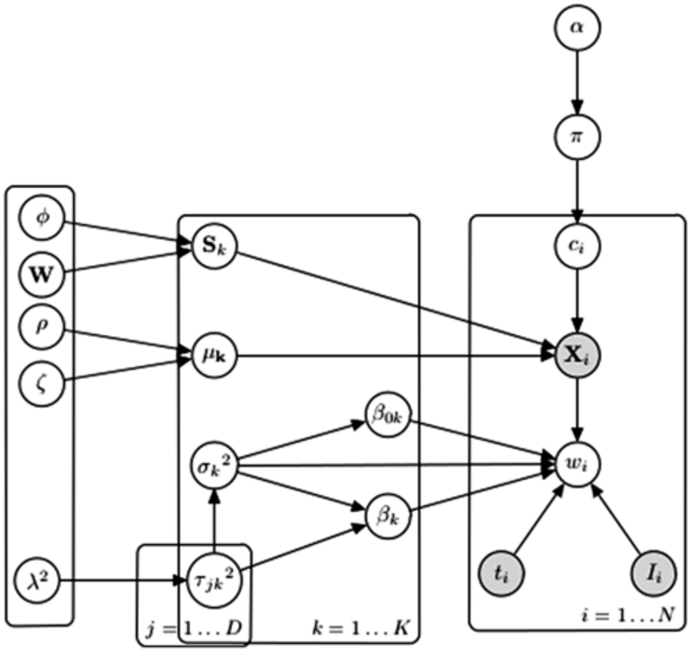


Figure 2.3: The probabilistic graphical model of SBC [3]

Based on the above priors, the authors use Gibbs sampling for inference. For example, the following conditional probability is used for sampling the cluster membership:

$$P(c_i = j|c_{-i}, \mu_j, S_j^{-1}, \beta_{0j}, \beta_j, \sigma_j^2, \alpha) \propto \frac{n_{-i,j}}{(N-1+\alpha)}\mathcal{N}(w_i|\beta_{0j} + \beta_j^T X_i, \sigma^2)\mathcal{N}(X_i|\mu_j, S_j^{-1})$$

The above conditional probability includes terms related to both expression data and survival data which indicates the influence of both data on clustering. The authors also provide methods for inferring the feature significance for identifying the features that discriminate only a pair of clusters. They also provide methods to predict cluster membership and survival time for an unseen test case. Experiments on simulated data as well as Breast Cancer and Glioblastoma Multiforme data indicate superior clustering and survival prediction accuracy of the proposed model compared with the state-of-the-art methods.

2.2.8 Discussion

Most of the discussed methods assume subtype-specific prediction models and feature importance ([51, 38, 109, 3]). The assumption in some of these methods ([38, 109]) is that the target phenotype is influenced by unknown variables. Therefore, the subtypes are defined to separate the patients into groups with similar values for the unknown variables, but not necessarily similar observed features, although the predictive relationship between the observed features and the phenotype is shared among the patients. However, this might deviate from the definition of subtype as a group of patients with similar observed features. Therefore, the focus in these methods is more on the predictive performance than subtyping. Moreover, having subtype-specific feature weights increases the number of inferred variables and might result in over-fitting. This methods use non-parametric clustering and their model is often limited to a specific type of phenotype, e.g. survival. This restricts the applicability of these methods.

On the other hand, the rest of the methods concentrate on subtype detection and use the phenotype data to improve the subtyping ([48, 4, 46]). All of these methods use integrative matrix factorization to simultaneously analyze omics data and a binary matrix of multiple phenotypes, e.g. response to different drugs. This type of phenotype data is not available in all settings, e.g. prediction of transplant rejection or survival. This results in limited applicability of these methods.

Addressing these issues, we provide a method called Supervised Bayesian Patient Stratification (SUBSTRA) in chapter 4. SUBSTRA uses biclustering, which is more appropriate for detecting local patterns in omics data [100] as genes work in groups and biclustering provides a better picture of these patterns. It is a non-parametric method and assumes one categorical phenotype, which makes it more general. The distributions of the proposed Bayesian framework can be tailored to other types of data.

2.3 Intra-Tumor Heterogeneity Detection

In this section we move one level deeper and look at the heterogeneity within individual tumors. Cancer is the result of a gradual accumulation of somatic genetic mutations. While most of the acquired mutations are putatively neutral and have no significant effect on a cell's phenotype, some confer a selective advantage to the host cell; they are known as *driver mutations*. Consequently, individual tumors are heterogeneous and typically consist of multiple populations of cells (subclones), each harbouring a distinct set of driver mutations and possessing a distinct phenotype, a phenomenon known as intra-tumor heterogeneity (ITH). Detecting the subclones and the order in which they have evolved, which we call ITH detection, helps identify the key events initiating the development of the disease or leading to metastasis, and allows for the determination of a tumor's subclonal composition.

This area has attracted remarkable attention recently and there is a rich literature about this topic. In this section, we categorize and review some of the existing ITH detection methods. These methods can be classified with respect to the following aspects:

- Compatible DNA sequencing technology: Whether the method is designed for bulk or single-cell sequencing data or it leverages both. Currently, most of the existing datasets (e.g. TCGA [1]) contain bulk sequencing data. Bulk sequencing data provides a collective picture of alterations that has occurred in all cells in a taken sample. From this data, one can roughly extract the portion of cells harbouring a particular mutation. At a lower granularity, some methods consider only the existence of mutations in a sample based on the strength of the corresponding signal, e.g. if a large portion of cells have the mutation. Single-cell sequencing is a recent technology and is not as available, partly, due to considerable associated expenses. This data provides a much higher resolution compared to bulk sequencing and facilitates the ITH detection. However, it still suffers some sources of noise (e.g. allelic drop-out). The quality of both forms of data depends on the sequencing coverage, which is the average number of reads that cover each base, i.e. maps to the portion of DNA that contains the base.

- Type of input data: Whether the method works with single nucleotide variations (SNVs), copy number variations (CNVs) or both. For each of this mutation types, different input formats can be considered. Binary format only indicates the existence (1) or absence (0) of the mutation in a single sample or cell, depending on the sequencing technology. Most of the existing methods working with single-cell data use this representation. Variant allele frequency (VAF) data indicates the portion of reads (small portions of DNA with known contents), out of total number of reads that map to a specific locus on genome, that contain a mutation. For a heterozygous SNV in a diploid region (a region with only two copies with one of them mutated), this value is half the fraction of cells that contain the SNV. VAFs can also provide information about CNVs. Read count data contains the exact number of total and variant reads corresponding to a specific locus with a mutation. This higher granularity adds another dimension to the data indicating the certainty of the signals. Larger numbers of reads indicate higher certainty about the information. For example, assume two loci a and b with variant read counts $v_a = 10$ and $v_b = 1000$ and total read counts $t_a = 100$ and $t_b = 10000$. Both loci have a VAF of 10%, however, the certainty of VAF for locus b is higher because if v_a and v_b are varied by 1, VAF will have smaller shift for b compared to a. Accordingly, larger read counts, or equivalently more sequencing depth, reduces the effect of noise.

- Level of input data: Whether the method uses the information from all individuals (i.e. population/ensemble level methods) to infer the evolutionary model or it uses only one individual's sequencing data. If multiple samples are available for an individual and the sequencing quality is high, it might be possible to infer reliable evolutionary models. Otherwise, more information can be acquired from other individuals profiles. Although, as discussed earlier, inter-sample heterogeneity exists between different individuals and complicates the understanding and treatment of disease, one can still expect partially similar evolutionary patterns for subsets of individuals. Finding and leveraging these similarities when studying the disease evolution might result in less uncertainty. Accordingly, some methods use the mutation profiles of the whole population as input.

- Assumed evolutionary model: Evolutionary models show the different subclones of a tumor and their evolutionary relationships with each other. These can be represented as directed graphs. Often, subclones are shown by nodes and mutations that relate them together are illustrated by directed edges. The tail of each edge is a subclone that is transformed to the subclone at the head of the edge with the corresponding mutation of the edge. Accordingly, mutations on that edge are the only differences between the tail subclone (parent) and the head subclone (child). An alternative approach is to use nodes as the mutations and edges indicating the order of mutations. We call the former model the subclone-based model and the latter the mutation-based model. As default, we assume a subclone-based model in our discussions. The evolutionary models can take one of the following forms: 1) a chain of events or a directed path, which is a linear model of mutations, 2) a tree, which can be seen as a combination of different linear models and allows for modeling parallel evolution of different subclones, and 3) a directed acyclic graph (DAG), which, unlike trees, allows for one subclone to have multiple parents. Tree models are often based on *infinite site assumption (ISA)*, which limits the number of occurrences of a mutation to one, i.e. each mutation can appear only in one edge of the tree and is present (conserved) in all the descendants of the subclone in which it first occurs. On the other hand, DAGs relax this assumption and allow for modelling different possible orders of mutations that result in the same subclone.

- Level of output models: Whether the computed evolutionary model refers to only an individual, a group of individuals (sub-population) or it is a general model for the whole population. Obviously, if the input data is for one individual, the output will correspond to only that individual. However, for population-level methods (methods that use population information as input), the output models can be any of the mentioned generality levels. Population-level methods that compute individualized models use more information than individual-level methods and, at the same time, provide higher resolution personalized models.

Table 2.3 lists several methods for ITH detection. Although there are relevant methods that only detect the subclones without inferring their evolutionary relationships (e.g. [141, 113, 76, 144, 112]), we are only interested in and list the methods that detect both. In the next paragraphs, some of the listed methods that are representative of a wide range of different approaches are briefly described.

2.3.1 Popic et al. [101]

The authors provide a method for deriving mutation-based evolutionary trees. They take the VAF profile of a population of samples as input. In the first step, they binarize these profiles and use the binary profiles to get a rough grouping of mutations. Then, they use

Reference	Technology		Input Type		Input Level		Output Model			Output Level		
	Bulk	SC	SNV	Both	Individual	Population	Linear	Tree	DAG	Individual	Sub-population	Population
[41, 52, 86, 101]	●		V		●			●		●		
[36, 61]	●		R		●			●		●		
[33]	●			R	●			●		●		
[60, 68, 108, 145]		●	B		●			●		●		
[87, 88]	●	●	B,R		●			●		●		
[105]	●			B		●	●*					●
[34, 35]	●			B		●		●				●
[30, 44]	●			B		●			●*			●
[8, 14, 18, 19, 43, 55, 103, 111, 114]	●			B		●			●			●
[64]	●		V			●	●*				●	
[16, 17, 131, 82]	●			B		●		●			●	
[23]	●		V			●		●		●		

Table 2.3: Existing intra-tumor heterogeneity detection methods sorted by their input and output levels."Both" in the "Input Type" column means both SNV and CNV data. "B", "V" and "R" in the "Input Type" column respectively stand for binary, VAF and read count data. Asterisk * indicates evolutionary model at the level of pathways, i.e. pathways on edges instead of genes.

Gaussian mixture models to further cluster each group based on their VAF values. These finer clusters are then organized into a DAG by connecting each mutation to the other mutations that have smaller VAFs (with some tolerance) across all patients. Then, they search for all spanning trees of that DAG that satisfy the sum rule for evolutionary trees, which requires the sum of VAFs of children of a node not to be larger than VAF of the node itself. The allow for some deviation from the sum rule and rank the spanning trees based on the amount of deviation from that rule and select the top tree as the solution. The method is evaluated on several synthetic and real datasets and performed reasonably. The advantage of this method, called LICHeE, is its relatively low computational complexity that allows it to handle hundreds of mutations.

2.3.2 Donmez et al. [36]

This paper introduces a method called CTPsingle that is designed for finding evolutionary tree models from single-sample low-coverage bulk sequencing data. The input to the method consists of read count data from heterozygous SNVs in diploid regions. First step of CTPsingle is clustering the mutations. The authors use a beta-binomial mixture model for this aim. Let y_i and n_i be the variant and total read count data for mutation i. Then, y_i is assumed to have a binomial distribution with an unknown success probability p_i:

$$y_i|(n_i, p_i) \sim Binom(n_i, p_i)$$

The parameter p_i is generated from a Dirichlet process with concentration parameter α: $p_i|(\alpha, G_0) \sim DP(\alpha, G_0$. The baseline distribution G_0 is defined as $G_0 = beta(a_1, b_1)$. Given the conjugacy of beta and binomial, Markov Chain Monte Carlo (MCMC) approach is used for parameter learning. The above non-parametric clustering provides the number of subclones and their cellular prevalence. These information are then provided to CITUP, a method introduced in [86], to detect the evolutionary tree. CITUP employs a mixed integer linear program to detect the optimal tree out of all possible trees.

CTPsingle is evaluated on synthetic and prostate cancer data and has achieved satisfactory results compared to other methods such as [101].

2.3.3 Deshwar et al. [33]

The authors improve upon a previously existing method for analyzing SNV data, called PhyloSub [61], by considering CNV data when inferring the sub-clones and phylogenetic relationships between them. In PhyloSub, the generative process is as follows. First, the phylogenetic tree is generated using a tree-structured stick-breaking process prior. Then SNV cellular prevalence values are generates based on the tree topology and following a Dirichlet process. Then at each locus, the genotype is sampled following a categorical distribution. The genotype identifies the total as well as the reference and variant allele copy numbers. Finally, the observed data are generated. The data is the number of reads mapped to the reference alleles at each loci. This variable follows a binomial distribution and is dependent on the SNV cellular prevalence, genotype, total number of reads, proportion of reads from normal cells mapped to the reference allele and proportion of reads from tumor cells mapped to the reference allele for each loci.

The method proposed in this research [33] is called PhyloWGS. In this method, the CNV data is included in phylogeny inference. The input CNV data includes the value of copy number together with the proportion of cell population having that alteration. We note that it is not in practice easy, if possible, to compute these values from bulk sequencing data. Assuming that these data are accessible, to extend PhyloSub by incorporating these data, the authors provide different solutions considering whether the CNV locus overlaps an SNV or not. If they do not overlap, the copy number alterations are converted into pseudo-SNVs, which are represented in the model as a heterozygous, binary somatic mutations happening in the cell population containing the CNV. If CNV and SNV overlap, there are different scenarios regarding the order of occurrence of CNV and SNV, for each of which the authors provide modeling instructions. The inference process finds the scenario that fits the best to the data.

2.3.4 Ross and Markowetz [108]

The authors provided a probabilistic score for evolutionary trees as well as a search strategy for finding a locally optimum tree. The proposed algorithm, called OncoNEM, has three steps: 1) initial search for building a cell tree, in which they find a locally optimum phylogenetic tree between cells, 2) expansion, in which they add possibly extinct or unobserved cells to the cell tree if it increases the tree score more than a given threshold, and 3) clustering, in which the (adjacent) cells are merged into clusters/sub-clones to form new trees with higher scores.

The probabilistic model used for scoring the trees consists of these variables: binary cell SNV profiles indicated by D, the tree topology indicated by T, and the first clone in which each mutation happens indicated by θ. Then the joint probability is:

$$P(D, T, \theta) = P(D|T, \theta)P(\theta|T)P(T)$$

T and θ have uniform prior distributions. Data probability distribution is defined as below:

$$P(D|T, \theta) = \prod_{l=1}^{m} \prod_{k=1}^{n} p(\omega_{kl}|\delta_{kl}),$$

where the probability $p(\omega_{kl}|\delta_{kl})$ relates the observed value for the lth SNV of the kth cell, i.e. ω_{kl}, to its predicted value, i.e., δ_{kl}. This probability distribution is defined based on the false-positive and false-negative rates in SNV detection experiments. Finally, the tree score is computed by marginalizing over the SNV-clone assignment parameter θ as below:

$$score(T) = P(D|T) = \int_{\theta} P(D|T,\theta)P(\theta|T)d\theta$$

They provide heuristic algorithms for each of the three mentioned steps to maximize the above tree score. Specifically, in the initial search step, they move in the tree topology space by either assigning a new parent to one of the nodes or swapping two nodes that are connected by an edge. In the expansion step, they insert additional nodes under a branching node, i.e. a node with more than one child, if the insertion increases the joint probability. Finally, in the clustering step, they merge two neighbor nodes if it improves the joint probability.

A very similar method called SCITE is proposed in [60]. The difference between SCITE and OncoNEM is that OncoNEM infers a subclone-based tree while SCITE infers a mutation-based tree, which provide equivalent information. The only difference is the way the tree nodes are labeled (subclones versus mutations). Accordingly, the likelihood in SCITE is defined as below:

$$p(D|T,\theta,\sigma) = \prod_{i=1}^{m}\prod_{j=1}^{n} p(D_{ij}|E_{ij}),$$

where E_{ij} is the predicted mutation matrix defined by T and σ, σ indicates the attachment of each cell to each node of mutation tree T, and θ is sequencing errors (false-positive and false-negative rates). Similar to OncoNEM, SCITE uses an MCMC scheme for searching the tree topology space to estimate the posterior distribution of tree topologies.

2.3.5 Malikic et al. [87]

This method is one of a few methods that leverage the advantages of both bulk and single-cell sequencing data based on the fact that these two data types complement each other when inferring evolutionary trees. The method, called B-SCITE, is based on joint likelihood of both bulk and single-cell data.

For bulk data, binomial distribution is used to model the read count data. Because a high coverage is assumed, the binomial distribution is approximated by a Gaussian distribution. The variance of this distribution is fixed to the value observed in the input data. However, the mean is considered a latent variable. Then, the log-likelihood of bulk data from a sample given the tree T is defined as below:

$$S_{\text{bulk}}(T) = \max_{y_1,\ldots,y_{s+1}} \sum_{i=1}^{n} \frac{-t_i}{8.\frac{z_i}{2}(1-\frac{z_i}{2})}.(z_i - y_i)^2,$$

where s is the number of subclones (the number of tree nodes considering the root is $s+1$), n is the total number of mutations, y_i is the true cancer cell fraction (CCF) for mutation i

and z_i is the observed CCF for mutation i computed using the read count data. The values of y_i $(1 \leq i \leq s)$ are constrained by the tree structure T.

For single-cell sequencing data, the probabilistic model proposed in [60] and discussed briefly in the previous section is used. The authors modify the original likelihood to incorporate noises originating from doublets, a sequencing noise which happens when two cells are captured together and their mixed mutation profile is reported as a single-cell profile. The final data likelihood is defined as the sum of the two log-likelihoods of bulk and single-cell data. The parameters, including the tree structure and error rates of single cell sequencing, are learned using Metropolis-Hastings sampling as in [60]. The results of experiments on synthetic data indicates superiority of B-SCITE to OncoNEM [108] and ddClone [112] and its robustness to doublets. The method is also applied to childhood leukemia data from two patients.

2.3.6 Desper et al. [35]

This paper presents one of the earliest methods of evolutionary modelling. The method assigns distances between each pair of mutations and uses phylogeny construction algorithms to reconstruct the evolutionary model, which the authors call oncogenetic tree. The input is a binary matrix indicating the occurrence of mutations in set L in k samples. The assumption is that the samples are generated by a distribution p over $2^|L|$ possible combinations of the mutations and p is defined by a tree T which has the L mutations as its leaves. The problem then becomes finding a tree that has an associated distribution or $2^|L|$ possible data that is close to observed p.

For this, the authors reduce the problem to numerical taxonomy problem, which is based on a distance matrix between the entities. To compute the distance between each pair of mutations x and y, a path metric is defined as $d_T(x, y) = \sum_{e \in P_{xy}} d(e)$, where P_{xy} is the path in tree T between x and y and e indicates an edge. Considering $d(e) = -\log p(e)$, where $p(e)$ is the probability of edge e in T, the distance between the two events becomes:

$$d_T(x, y) = -2 \log p_{xy} + \log p_x + \log p_y,$$

where p_{xy} is the probability that x and y happen together estimated as the proportion of observed samples having both of the mutations and p_x and p_y are probabilities of each event estimated as the proportion of samples having the mutation. Then a tree fitting algorithm can be used for finding a tree that has an associated metric close to the estimated d_T. The method is used for renal cancer and the results are consistent with the existing domain knowledge and suggests new findings.

2.3.7 Cristea et al. [30]

The authors provide a probabilistic graphical model that captures mutual exclusivity to discover pathways and it simultaneously derive a complex DAG evolutionary model among those pathways. It generalizes both TiMEx [126], a method that finds mutually exclusive sets of genes, and Conjunctive Bayesian Networks (CBN) [19, 43], which identifies partial orders of mutations. It is also a generalization of the method proposed in [105], as it considers DAG among the pathways instead of a linear progression model as in [105].

The method consists of the following steps. Given a binary matrix of mutations, first, the mutually exclusive groups of genes are discovered by TiMEx as the current pathways. Then, the progression among the current pathways is inferred with CBN. Starting with this initial solution, an iterative approach consisting of two steps is used for optimizing the solution. In the first step, given the fixed progression among pathways, the assignment of genes to pathways is optimized through an MCMC approach. In the second step, given the fixed pathways, progression among pathways is optimized using simulated annealing. The joint optimization is repeated until convergence of both aspects.

The objective function for the above optimization is the marginal likelihood of the observed data Y given the model consisting of the assignment of genes to pathways and their evolutionary relationship in form of a DAG. In the probabilistic graphical model underlying the marginal likelihood, the latent variables include $T = (T_1, ..., T_n)$, the waiting time to alteration of n genes, $U = (U_1, ..., U_p)$, the waiting time to alteration of p pathways, and $X = (X_1, ..., X_n)$ the true noiseless mutation statuses of genes. T_{obs} is the time of biopsy relative to the tumor onset. Because this time is not known, it is modeled as the exponential distribution $T_{\text{obs}} \sim \text{Exp}(1)$.

The waiting time for a pathway to be altered is equal to the waiting time to the first mutation in the pathway. A graphical model indicates the dependencies between the pathways implying the order of pathway mutations. For example, if pathway i gets mutated before pathway j, i.e. $U_i < U_j$, then U_i is the parent of genes that belong to pathway j and the T variables of those genes are the parents of U_j in the graphical model. Moreover, the time of gene alteration for gene g is defined as:

$$T_g \sim \max_{Q \in pa(P_g)} U_Q + \text{Exp}(\lambda_g),$$

where P_g is the pathway containing g, $pa(P_g)$ is the set of direct parents of P_g in the partially ordered set of pathways, and λ_g is the timing parameter for gene g. The above equation means that the gene can be mutated with a exponential distance after all parents of the corresponding pathway are mutated. The prior distributions for X is a Bernoulli distribution, which is then simplified to an "if" statement that sets $X_g = 1$ if $T_g < T_{\text{obs}}$ and g is the only mutated gene in P_g. For the timing variables U and T and exponential distribution is used as described before. The values of X also depends on the timing of genes within the same pathway. Y depends on X through an error parameter ϵ.

2.3.8 Beerenwinkel et al. [19]

Multiple methods on evolutionary models (e.g. [14, 19, 43, 111, 114]) are based on a theoretical framework called Conjunctive Bayesian Network (CBN) originally proposed in [18, 19]. CBN is a DAG that indicates the relationships between mutations in terms of dependency of a mutation to others. CBN relaxes the ISA and only assumes non-reversible mutations. In other words, in a CBN a mutation can appear on multiple edges, however, when it occurs, the corresponding gene remains mutated. Therefore, unlike the tree evolutionary model, a subclone can have multiple parents. For example, a subclone with mutations $\{a, b\}$ can be the result of b happening after a or vice versa, which are modeled as two different paths resulting in the same subclone in the CBN.

A CBN is defined as a triplet $(\xi, \leq, \theta)$, where ξ is a set of n genomic events, $(\xi, \leq)$ is a partially ordered set or poset over ξ, and $\theta = (\theta_1, ..., \theta_n)$ is a vector of parameters for the events. A relationship $e_1 < e_2$ between two events in ξ indicates that e_1 must happen before e_2 can. θ_e is the conditional probability that event e will happen given that all of its predecessors in $(\xi, \leq)$ have occurred. Accordingly a CBN is a distributive lattice of order ideals in ξ. An order ideal is a genotype $g \subseteq \xi$ such that if $e_2 \in g$ and $e_1 < e_2$, then $e_2 \in g$.

Given the parameters θ, the probability of a genotype g can be defined as below:

$$P_g(\theta) = \prod_{e \in g} \theta_e \prod_{e \in \min(g^c)} (1 - \theta_e),$$

where $\min(g^c)$ is the set of events that are not in g but can happen next according to poset $(\xi, \leq)$ given that events in g have already occurred. According to this, a CBN defines a distribution over the genotypes. So, the task is to find the CBN that describes the observed genotypes.

For this, the authors provide a maximum likelihood estimator for the parameter θ given the number of observations of each genotype g denoted as u_g and poset $(\xi, \leq)$:

$$\hat{\theta}_e = \frac{\sum_{g: e \in g} u_g}{\sum_{g: below(e) \subseteq g} u_g},$$

where $below(e)$ is the set of events that happen before e in $(\xi, \leq)$. This equation is the number of genotypes that contain e divided by the number of genotypes that contain all predecessors of e. The poset is then formed by including a relationship $e < f$ if and only if $g \cap \{e, f\} \neq f$ for all observed g. The authors prove that a CBN defined in this way will be associated with the maximum log-likelihood.

Although CBN is applied to HIV genetic data in the original paper, later Gerstung et al. [43] apply it to different types of cancer after extending it to hidden CBN (H-CBN) by modifying the prior distributions and adding a layer of hidden variables to capture the noise. H-CBN is also used for infer evolution at the level of pathway in [44] using known pathways. In another direction, Shahrabi Farahani and Lagergren [114] generalize CBN by defining more complex models of dependency between an event and its parents which allowed for dependency on a subset of the parents with a tolerance. CBN is a special case of this general model with zero tolerance.

2.3.9 Attolini et al. [8]

The authors propose a method named RESIC based on the principles of population genetics. They assume single cell per person and study the evolutionary dynamics of individuals accumulating the mutations leading to cancer. At steady state, the population is distributed across all possible states and this distribution is assumed to be close to the observed distribution of samples. The parameters of the mathematical model are estimated by minimizing the difference between the prediction and the observed frequencies.

2.3.10 Khakabimamaghani et al. [64]

This is another work that addresses evolution at the level of pathways instead of genes. As in [30], this method also generalizes the model of [105], but in another direction. In [30], the previous model is generalized with respect to the progression model from linear to DAG. Differently, in this paper, the authors assume linear progression, however the progression is modeled in a subtype-specific way. The method, called SPM, infers the mutual exclusivity pathways simultaneously with subtypes (in this work, groups of patients with similar progression orders) and their linear progression orders. Moreover, unlike most of the methods, the proposed method, called SPM, employs Cancer Cell Fraction (CCF) data instead of binary mutation profiles. The underlying rationale is that CCF can be used as a proxy for the time of mutation for heterozygous SNVs in diploid regions. This helps to identify the order of mutations.

The authors prove that the mentioned problem is NP-hard and they use an integer linear programming (ILP) approach to solve it. The objective function for the ILP is the difference between the predicted and observed data. The authors provide variables and constraints for clustering, mutual exclusivity, and subtype-specific linear progression. The experimental results with synthetic and real data indicate that the identified subtypes, pathways, and progression orders are consistent with the domain knowledge and the method outperforms PLPM [105].

2.3.11 Tofigh et al. [131]

This reference proposes Hidden-variable Oncogenetic Trees (HOTs), and extension to oncogenetic trees that uses hidden variables to represent the mutations/nodes on a tree structure. These hidden variables are then related to the observed variables. This allows for modeling the data noise, i.e. false positives and false negatives.

A HOT consists of a tree structure T and two parameters $\theta_Z(u)$ and $\theta_X(u)$ for each vertex u in T. The former parameter is the conditional parameter for the value of vertex/mutation u given its parents' values. The latter is the conditional probability of the observed mutations X given their true values Z. In the standard expectation maximization algorithm, only the hidden variables, i.e. Z, are optimized, leaving the structure T out. Accordingly, the authors use Edmond's optimal branching algorithm to optimize the tree structure. They produce a complete, directed and weighted graph by assigning weights to the edges based on the expected log-conditional likelihoods and use Edmond's algorithm to find the maximum arborescence (directed rooted tree) of that complete graph. This is conceived as the optimal tree structure maximizing the likelihood, and thus the optimal oncogenetic tree. Finally, they generalize their method to HOT-mixtures by allowing multiple HOTs and provide an expectation maximization method for learning HOT-mixtures. The authors compare the performance of their algorithm with Mtreemix [17] and show that HOT-mixtures significantly outperforms when the number of mutations and the amount of noise is large.

2.3.12 Loohuis et al. [82]

This work is one of the few that use Suppes' causation theory [125] to infer the causal relationships between the mutations. Suppes' probabilistic causation theory indicates that for any two events c and e occurring respectively at times t_c and t_e with probabilities $0 \leq P(c), P(e) \leq 1$, the event c is *prima facie* cause of the event e if it happens before that, i.e. $t_c < t_e$, and raises its probability, i.e. $P(e|c) > P(e|\bar{c})$. Because the time of mutation events are not available, the authors show that mutation frequency in binary data can be used as a proxy for time, with higher frequency indicating earlier occurrence.

The method proposed in this work, called CAPRESE, has two main characteristics: 1) it uses probabilistic causation mentioned above instead of correlation to infer progression structures, and 2) it uses a shrinkage-like estimator to measure causation among any pair of events. This estimator finds the optimal balance between probability raising and correlation depending on the amount of noise. It estimates the confidence in causation from event a to event b as:

$$m_{a\to b} = (1-\lambda)\frac{P(b|a) - P(b|\bar{a})}{P(b|a) + P(b|\bar{a})} + \lambda\frac{P(a,b) - P(a)P(b)}{P(a,b) + P(a)P(b)} \tag{2.1}$$

The first term is the probability raising term and the second term denotes the correlation. λ indicates the balance between these two terms. This estimator is is computed for all pairs of events. Then, the evolutionary tree is computed based on these weights by assigning to each mutation b a causal event a if $m_{a\to b} > m_{b\to a}$ and $\forall a', m_{a\to b} > m_{a'\to b}$. The authors also provide criteria for attaching an event to the root node (germline), meaning that the event is not an effect of any causal mutation. The method is applied to synthetic and real data. Comparisons with Conjunctive Bayesian Networks [19, 14] indicates higher accuracy of CAPRESE.

2.3.13 Ramazzotti et al. [103]

Unlike in [82], where Suppes' probabilistic causation theory is used for deriving evolutionary trees, this work uses a similar theoretical ground for inferring evolutionary DAGs. In most of the previous works on DAGs each node represents one event. Differently, in this work, a parent node can be a logical combination of the events. This relates this work to [114]. In this framework the problem reduces to the following tasks: for each input event e, assess a set of logical selectivity/parent patterns, filter the spurious ones, and combine the rest in a DAG, augmented with logical symbols.

The input to the proposed method, named CAPRI, is a binary matrix of cross-sectional data and an optional set of hypothesis about relationships between a logical combination of some of the events and their consequent/child event. If there is any hypothesis provided, the first step is to lift the input data by adding columns that correspond to the combinations in the hypothesis. Then, a DAG is constructed by adding potentially causal relationships based on Suppes' theorem between the columns of the lifted data (singleton events and logical combinations). Each causal relationship is represented by an edges such that the head edges can only be a singleton event. CAPRI employs a bootstrapping technique to compute p-values for the edges and remove more random edges that result in cycles. Similar to CBN, this DAG induces a distribution of observing a particular mutation profile if its parameters are assigned a value. Also similarly, the parameters of the model can be learned given the input data.

The DAG produced as above contains both genuine and spurious causal relationships. For filtering the spurious relationships out, the authors use the fact that spurious relationships reduce the data likelihood given a model. Accordingly, they provide an approach to optimize Bayesian Information Criterion (BIC), which penalizes the model complexity.

2.3.14 Discussion

Many of the existing methods for studying tumor evolution operate on tumor data from a single cancer patient. The earliest developed methods used bulk sequencing data from a single sample (*e.g.* rec-BTP [52], CTPsingle [36]) or multiple samples from the same individual (*e.g.* PhyloWGS [33], AncesTree [41], LICHeE [101], CITUP [86]). These were followed by the development of several methods that work on single-cell data (*e.g.* OncoNEM [108], SCITE [60], SiFit [145]). The most recently introduced methods, B-SCITE [87] and

PhISCS [88], simultaneously utilise the complementary strengths of both single-cell and bulk sequencing data. Indeed, most of the methods above have limitations when faced with the input consisting of a single sample low-to-medium coverage bulk sequencing dataset, which are predominant in existing databases (*e.g.* TCGA [1] and cBioPortal [25, 42]). Since this type of data contains numerous ambiguous cases (i.e. cases where the input data is consistent with more than one possible phylogenetic tree for the tumor), the existing algorithms for ITH detection based on a single tumor sample (*e.g.*, CTPsingle [36]) will yield several possible solutions for those cases [87].

In addition to ITH, inter-tumor heterogeneity is another phenomenon complicating the understanding and treatment of cancer. Inter-tumor heterogeneity is a direct consequence of the fact that individual tumors are genetically distinct. Despite the inter-tumor heterogeneity, one can still expect partially similar evolutionary trajectories among subsets of tumors [23, 99]. Leveraging the phylogenetic similarities among tumors from a cohort of patients in a collaborative fashion can guide the process of exploring the solution space and reduce the above-mentioned ambiguities in inferring tumor phylogenies, especially for cases when the input is low to medium coverage bulk sequencing data from a single tumor sample. Most of the existing methods that employ population level data are based on binary mutation data. Some of these methods (e.g. CAPRESE [82], CAPRI [103], and Beerenwinkel et al. [15]) exploit Suppes' probabilistic causation theory [125] to determine the pairwise order of mutations. Some other methods (e.g. Conjunctive Bayesian Networks [19, 43] and Bayesian Mutation Landscape [93]) model the phylogenetic relationships as a Bayesian network and propose approaches for learning the network structure. Although using more information by considering the whole population, these methods gain general knowledge about cancer progression and do not provide personalized evolutionary details.

Some of the existing methods [64, 16, 17, 131, 82] address the mentioned issue by computing subtype specific progression models. Although useful, one might still be interested in patient-specific evolutionary model for more confident design of personalized treatment strategies. Moreover, since most of the mentioned methods use binary mutation data, they do not fully utilize the potential of sequencing data by overlooking the intrinsic information about the timing of evolutionary events.

A recent method, REVOLVER [23], fills in these gaps by using non-binary sequencing data of the whole population to learn the personalized evolutionary models. It exploits the repeating evolutionary patterns for ITH detection in individual tumors by transferring information across all tumors. In this method, the assumption is that a particular mutation usually has the same predictor (preceding mutation) across different tumors in a particular cancer type. Accordingly, the authors consider the frequency of the direct ancestors of a mutation across different tumors and use that information when inferring the phylogeny for a specific tumor. This approach use in REVOLVER decreases the uncertainty of phylogenetic structures by incorporating the ancestry information. However, the underlying evolutionary assumption and the learning approach used in this method might result in incorrect predictions.

In chapter 5, we further describe REVOLVER and discuss its disadvantages. We propose HINTRA as an improved method with a Bayesian approach and benchmark its performance against REVOLVER.

Chapter 3

Unsupervised Patient Stratification

In empirical medicine, every patient of a particular disease initially receives almost the same treatment. However, although working for simpler diseases to a degree, this approach has not been successful for more complex diseases like cancer. Therefore, the paradigm in medicine is shifting from Empirical to so called Personalized Medicine, which is a patient derived approach with the goal of providing individual treatments for each patient according to his/her particular conditions and features. As an intermediate step currently being investigated, "Stratified Medicine is an approach by which groups of patients with the same disease are subdivided into different categories depending on the underlying mechanism of disease and their probable response to a therapeutic intervention [140]." According to the definition of stratified medicine, a cohort of patients is divided into subgroups, called subtypes, and the specific features of each subtype that constitute the disease mechanism for that subtype are identified and can then be used to design subtype-specific treatments.

The task of identifying the disease subtypes, which is central to stratified medicine, is also known as *patient stratification*. Although, as discussed in section 2.1, there have been several different patient stratification methods proposed in the literature, there are still significant open issues. First, it is still unclear if integrating different datatypes will help in detecting disease subtypes more accurately, and, if not, which datatype(s) are most useful for this task. Second, as most of the proposed stratification methods are deterministic, there is a need for investigating the potential benefits of applying probabilistic methods. Third, one possible approach to patient stratification is Biclustering, which although proven useful for this task [102] is not yet fully utilized in an integrative probabilistic framework. A comprehensive discussion of bi-clustering methods can be found in [98, 100]. Most of the existing patient stratification methods that employ biclustering are based on factor analysis. Alternative approaches (e.g. [91]) should be investigated.

In this chapter, we address these open issues by proposing a novel Probabilistic Graphical Model (PGM), which we call B2PS (Bayesian Biclustering for Patient Stratification), and appropriate evaluation metrics. To the best of our knowledge, the model provided here is the first integrative Bayesian biclustering model. While there are solutions for integrative biclustering [146] as well as Bayesian biclustering [91] in the literature, no work so far combines integrative, Bayesian, and biclustering concepts in one model.

Pontes et al. [100] provided a taxonomy of biclustering methods on expression data based on two aspects:

- Gene expression patterns: This is a classification based on the patterns that a bicluster's genes exhibit across a bicluster's samples. It includes biclusters with constant values (genes with similar expression values across the samples), constant values on rows or columns (genes with similar expression values across samples or vice versa, but possibly different values from gene to gene or sample to sample), coherent values on both rows and columns (correlated values or log-values of expression between each pair of genes or samples) and coherent evolution (up- or down-regulation of genes across the samples without any specific mathematical model for values inside a bicluster).

- Structure: This classification accounts for the way rows and columns of the input matrix are incorporated into biclusters. This divides the methods into row exhaustive (every row should be assigned to at least one bicluster), column exhaustive (similar but for columns), non exhaustive, row exclusive (each row can be assigned to at most one bicluster, i.e. no overlaps on rows), column exclusive (similar but for columns) and non exclusive.

Considering this taxonomy, B2PS belongs to the class of constant value exhaustive row and column exclusive class. We compare the performance of B2PS against NMF, a state-of-the-art deterministic method. Experimental results demonstrate the superiority of B2PS over NMF regarding both patient stratification and feature clustering in different experimental settings.

The main contributions of B2PS are as follows:

- The proposed model allows for incorporation of prior knowledge, which is useful for dealing with noisy data. Our experimental results show that this ability is useful for processing noisy biological data and improves the stratification performance.

- Given a prior upper bound number, the proposed method is able to detect the natural number of clusters for each dimension (i.e., row and column), identification of which requires an iterative trial process in deterministic methods. Measured evaluation metrics indicates that the natural sample clusters detected by our method form a better partitioning than the one detected by conventional NMF.

- Unlike conventional bi-clustering methods, the number of row and column clusters is not assumed to be the same in our model. This is a useful assumption that is more consistent with typical biological datasets and, according to our experimental results, provides a more informative clustering across both dimensions.

- The integrative method proposed here allows for examination of patient stratification results when using different combinations of diverse datatypes with no theoretical limitation on the number of data types. This makes it possible to identify the datatypes that are more useful for patient stratification. Experimental results with two TCGA datasets suggest that gene expression data is more informative than genomic data for patient stratification. This confirms the natural choice of this data type as the core input in many of the existing methods as discussed in section 2.1.

We believe that the outputs of the proposed method can be a useful basis for detecting the subtype-specific driver aberrations, which is one of the goals of stratified and personalized medicine. The R code of B2PS is available at https://github.com/sahandk/B2PS.

3.1 Problem Definition

We assume three input matrices for the problem of integrated biclustering. These are denoted by $S \in \{0,1\}^{r\times n^s}$ for point mutation, $E \in \{-1,0,+1\}^{r\times n^e}$ for gene expression and $V \in \{-2,-1,0,+1,+2\}^{r\times n^v}$ for copy number variation, where r is the number of patients and n^s, n^e and n^v are respectively the numbers of genes for point mutation, expression and copy number variation datasets. In S, 1 indicates existance of a mutation and 0 indicates otherwise. In E, -1, 0 and +1 respectively denote under-, neutral and over-expression. In V, the value indicates the amount of copy number variation for the corresponding patient-gene pair. We note that the value of copy number can be more than 2 in general. However, because we did not observe any larger values in our data, we do not consider any category for values larger than 2 here.

Given these inputs and a maximum number of patient clusters K^p and a maximum number of gene clusters for each dataset K^s, K^e and K^v, we are interested in producing patient clustering vector c^p and gene clustering vectors c^s, c^e and c^v, such that patients within a cluster or stratum are similar with respect to part of the gene clusters across different datatypes. Also, genes within a cluster should have similar values across part of the patient clusters. Emphasis is on finding biclusters that have a density skewed towards a specific value, e.g. biclusters with mostly +1's for expression data or biclusters with mostly 0's for point mutation data.

3.2 Methods

3.2.1 Model

The integrative probabilistic graphical model for B2PS is shown in Figure 3.1. Observed variables are shaded and hyper-parameters are in dotted circles. Table 3.1 includes a detailed description of the variables of the model.

Type	Name	Description	Distribution
Observed Variables	E_{il}	Expression status of gene j of patient i	$E_{il} \sim Categorical_3(\theta_{c_i^p, c_l^e})$
	S_{ij}	Mutation status of gene j of patient i	$S_{ij} \sim Bernoulli(\theta_{c_i^p, c_j^n})$
	V_{ik}	Copy number variation of gene k of patient i	$V_{ik} \sim Categorical_5(\theta_{c_i^p, c_k^v})$
Hyper-parameters	α^p and K^p	Parameter for prior Dirichlet distribution for patient clusters and the number of patient clusters and	$\alpha^p > 0$, $K^p \geq 1$
	α^x and K^x	Parameter for prior Dirichlet distribution for gene clusters and the number of gene clusters of data type x	$\alpha^x > 0$, $K^x \geq 1$
	λ^x and β^x	Parameters for prior distributions of θ^x.	$\lambda^e = \{\beta^e_{-1}, \beta^e_0, \beta^e_1\}$ $\lambda^s = \{\beta^s_0, \beta^s_1\}$ $\lambda^v = \{\beta^v_{-2}, \beta^v_{-1}, \beta^v_0, \beta^v_1, \beta^v_2\}$
Parameters	$\theta^X_{c^p, c^x}$	Parameters for distribution of the values inside bicluster (c^p, c^x)	$\theta^s_{c^p, c^s} \sim Beta(\lambda^s)$ $\theta^e_{c^p, c^e} \sim Dirichlet_3(\lambda^e)$ $\theta^v_{c^p, c^v} \sim Dirichlet_5(\lambda^v)$
	π^p	Parameter for distribution over patient clusters	$\pi^p \sim Dirichlet_{K^p}(\alpha^p)$
	π^x	Parameter for distribution over gene clusters	$\pi^x \sim Dirichlet_{K^x}(\alpha^x)$
Latent Variables	c_i^p	Cluster index for ith sample	$c_i^p \sim Categorical_{K^p}(\pi^p)$
	c_j^x	Cluster index for jth gene of data type x	$c_j^x \sim Categorical_{K^x}(\pi^x)$

Table 3.1: Variables and probabilistic relationships in the B2PS model. In this table, x indicates the data type and can be replaced with s (point mutation), e (gene expression), or v (copy number variation).

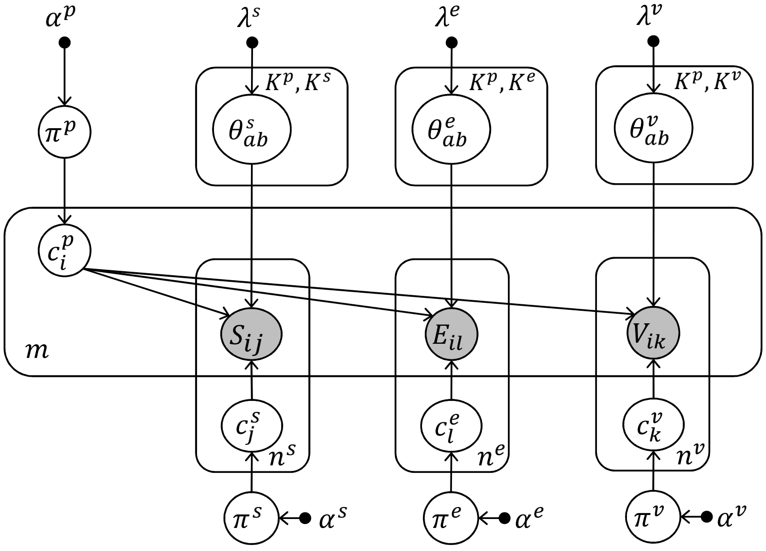


Figure 3.1: The probabilistic graphical model of B2PS.

Because the goal is to integrate different datatypes about the same set of patients, in our model, different datasets are assumed to have the same patients but they can have different genes. Accordingly, the patient clustering is shared across different datatypes, but each dataset has its particular gene clustering. However, gene clusterings of different datatypes are indirectly related to each other through the shared patient clustering. While, no direct dependency is assumed between patient clusters c^p_i and gene clusters c^e_l, c^s_j and c^v_k in this model, they are indirectly dependent given the observed data variables S, E and V. In terms of clustering structures discussed in [91], B2PS produces a single non-overlapping clustering, meaning that each patient or gene belongs to a single cluster that has no overlap with other clusters.

As for the generative process, for each data matrix, each element can be generated based on the parameter θ_{ab} where a is the patient cluster and b is the gene cluster corresponding to that element. Since we assume discrete values for the data matrices, Categorical distribution can be used for modeling data. Then, parameter θ can follow a conjugate Dirichlet prior. Similarly, the cluster vectors can be generated using a Dirichlet-Categorical distribution.

3.2.2 Parameter Learning

The Gibbs sampling method [24] is used for parameter learning and latent variable inference. After random initialization, the latent cluster vectors are iteratively sampled one by one based on marginal conditional probabilities. Model parameters π^p, π^s, π^e, π^v, θ^p, θ^s, θ^e and θ^v can be integrated out as they are continuous and difficult to be sampled.

For computing the conditional probabilities, we start by deriving the marginal joint probability by integrating over the model parameters:

$$
\begin{aligned}
&P(S,E,V,c^p,c^s,c^e,c^v|H) = \\
&\int_{\Theta} P(S,E,V,c^p,c^s,c^e,c^v,\Theta|H)\ d\Theta = \\
&\int_{\theta^s} P(S|\theta^s,c^p,c^s)P(\theta^s|\lambda^s)\ d\theta^s \int_{\theta^e} P(E|\theta^e,c^p,c^e)P(\theta^e|\lambda^e)\ d\theta^e \\
&\int_{\theta^v} P(V|\theta^v,c^p,c^v)P(\theta^v|\lambda^v)\ d\theta^e \int_{\pi^p} P(c^p|\pi^p)P(\pi^p|\alpha^p)\ d\pi^p \\
&\int_{\pi^s} P(c^s|\pi^s)P(\pi^s|\alpha^s)\ d\pi^s \int_{\pi^e} P(c^e|\pi^e)P(\pi^e|\alpha^e)\ d\pi^e \int_{\pi^v} P(c^v|\pi^v)P(\pi^v|\alpha^v)\ d\pi^v \qquad (3.1)
\end{aligned}
$$

In the above equation $H = (\alpha^p, \alpha^s, \alpha^e, \alpha^v, \lambda^s, \lambda^e, \lambda^v)$ is the set of hyper-parameters and $\Theta = (\theta^s, \theta^e, \theta^v, \pi^p, \pi^s, \pi^e, \pi^v)$ is the set of model parameters. In this equation, the integral is factorized based on the parameters included in each distribution. The three first integrals are marginal data likelihoods. These factors are computed similarly. We show the computation for the first factor in the following:

$$
\begin{aligned}
&\int_{\theta^s} P(S|\theta^s,c^p,c^s)P(\theta^s|\lambda^s)\ d\theta^s = \\
&\int_{\theta^s} \prod_{k_1=1}^{K^p} \prod_{k_2=1}^{K^e} \left[\prod_{i:c_i^p=k_1} \prod_{j:c_j^e=k_2} P(S_{ij}|\theta^s_{k_1k_2}) \right] P(\theta^s_{k_1k_2}|\lambda^s)\ d\theta^s = \\
&\prod_{k_1=1}^{K^p} \prod_{k_2=1}^{K^e} P(\theta^s_{k_1k_2}|\lambda^s) \left[\int_{\theta^s_{k_1k_2}} \prod_{i:c_i^p=k_1} \prod_{j:c_j^e=k_2} P(S_{ij}|\theta^s_{k_1k_2})\ d\theta^s_{k_1k_2} \right] = \\
&\prod_{k_1=1}^{K^p} \prod_{k_2=1}^{K^e} \frac{\Gamma(\beta_0^s+\beta_1^s)}{\Gamma(\beta_0^s)\Gamma(\beta_1^s)} {\theta^s_{k_1k_2}}^{\beta_1^s-1}(1-\theta^s_{k_1k_2})^{\beta_0^s-1} \times \\
&\int_{\theta^s_{k_1k_2}} \prod_{i:c_i^p=k_1} \prod_{j:c_j^e=k_2} {\theta^s_{k_1k_2}}^{S_{ij}}(1-\theta^s_{k_1k_2})^{(1-S_{ij})}\ d\theta^s_{k_1k_2} = \\
&\prod_{k_1=1}^{K^p} \prod_{k_2=1}^{K^e} \frac{\Gamma(\beta_0^s+\beta_1^s)}{\Gamma(\beta_0^s)\Gamma(\beta_1^s)} \int_{\theta^s_{k_1k_2}} {\theta^s_{k_1k_2}}^{(\bar{\mathbf{s}}_{k_1}^{k_2}(1)+\beta_1^s-1)}(1-\theta^s_{k_1k_2})^{(\bar{\mathbf{s}}_{k_1}^{k_2}(0)+\beta_0^s-1)}\ d\theta^s_{k_1k_2}, \qquad (3.2)
\end{aligned}
$$

where $\bar{\mathbf{s}}_{k_1}^{k_2}(z) = |\{S_{uw} : S_{uw} = z, c_u^p = k_1, c_w^s = k_2\}|$, i.e. the number of values in bicluster (k_1, k_2) of mutation data that are equal to z. Because

$$
\begin{aligned}
\int_{\theta^s_{k_1k_2}} &\frac{\Gamma(\bar{\mathbf{s}}_{k_1}^{k_2}(0)+\beta_0^s+\bar{\mathbf{s}}_{k_1}^{k_2}(1)+\beta_1^s)}{\Gamma(\bar{\mathbf{s}}_{k_1}^{k_2}(0)+\beta_0^s)\Gamma(\bar{\mathbf{s}}_{k_1}^{k_2}(1)+\beta_1^s)} \times \\
&{\theta^s_{k_1k_2}}^{(\bar{\mathbf{s}}_{k_1}^{k_2}(1)+\beta_1^s-1)}(1-\theta^s_{k_1k_2})^{(\bar{\mathbf{s}}_{k_1}^{k_2}(0)+\beta_0^s-1)}\ d\theta^s_{k_1k_2} = 1
\end{aligned}
$$

we can rewrite equation 3.2 as below:

$$\int_{\theta^s} P(S|\theta^s, c^p, c^s) P(\theta^s|\lambda^s)\ d\theta^s =$$

$$\prod_{k_1=1}^{K^p} \prod_{k_2=1}^{K^s} \frac{\Gamma(\beta_0^s + \beta_1^s)}{\Gamma(\beta_0^s)\Gamma(\beta_1^s)} \times \frac{\Gamma(\bar{\mathbf{s}}_{k_1}^{k_2}(0) + \beta_0^s)\Gamma(\bar{\mathbf{s}}_{k_1}^{k_2}(1) + \beta_1^s)}{\Gamma(\bar{\mathbf{s}}_{k_1}^{k_2}(0) + \beta_0^s + \bar{\mathbf{s}}_{k_1}^{k_2}(1) + \beta_1^s)}$$

$$\int_{\theta_{k_1 k_2}^s} \frac{\Gamma(\bar{\mathbf{s}}_{k_1}^{k_2}(0) + \beta_0^s + \bar{\mathbf{s}}_{k_1}^{k_2}(1) + \beta_1^s)}{\Gamma(\bar{\mathbf{s}}_{k_1}^{k_2}(0) + \beta_0^s)\Gamma(\bar{\mathbf{s}}_{k_1}^{k_2}(1) + \beta_1^s)} \times$$

$$\theta_{k_1 k_2}^s{}^{(\bar{\mathbf{s}}_{k_1}^{k_2}(1) + \beta_1^s - 1)} (1 - \theta_{k_1 k_2}^s)^{(\bar{\mathbf{s}}_{k_1}^{k_2}(0) + \beta_0^s - 1)}\ d\theta_{k_1 k_2}^s =$$

$$\prod_{k_1=1}^{K^p} \prod_{k_2=1}^{K^s} \frac{\Gamma(\beta_0^s + \beta_1^s)}{\Gamma(\beta_0^s)\Gamma(\beta_1^s)} \times \frac{\Gamma(\bar{\mathbf{s}}_{k_1}^{k_2}(0) + \beta_0^s)\Gamma(\bar{\mathbf{s}}_{k_1}^{k_2}(1) + \beta_1^s)}{\Gamma(\bar{\mathbf{s}}_{k_1}^{k_2}(0) + \beta_0^s + \bar{\mathbf{s}}_{k_1}^{k_2}(1) + \beta_1^s)} \tag{3.3}$$

The approach for the latent clusters is similar, but simpler. For example, the fourth factor in equation 3.1, which is the marginal patient clustering probability, is computed as follows:

$$\int_{\pi^p} P(c^p|\pi^p) P(\pi^p|\alpha^p)\ d\pi^p =$$

$$\int_{\pi^p} P(\pi^p|\alpha^p) \prod_{k=1}^{K^p} \prod_{i:c_i^p=k} P(c_i^p|\pi_k^p)\ d\pi^p =$$

$$\int_{\pi^p} \frac{\Gamma(K^p \alpha^p)}{\Gamma(\alpha^p)^{K^P}} \prod_{k=1}^{K^p} (\pi_k^p)^{\alpha^p - 1} (\pi_k^p)^{|\{u:c_u^p=k\}|}\ d\pi^p =$$

$$\frac{\Gamma(K^p \alpha^p)}{\Gamma(\alpha^p)^{K^P}} \times \frac{\prod_{k=1}^{K^p} \Gamma(|\{u : c_u^p = k\}| + \alpha^p)}{\Gamma(m + K^p \alpha^p)} \times$$

$$\int_{\pi^p} \frac{\Gamma(m + K^p \alpha^p)}{\prod_{k=1}^{K^p} \Gamma(|\{u : c_u^p = k\}| + \alpha^p)} \prod_{k=1}^{K^p} (\pi_k^p)^{|\{u:c_u^p=k\}| + \alpha^p - 1}\ d\pi^p =$$

$$\frac{\Gamma(K^p \alpha^p)}{\Gamma(\alpha^p)^{K^P}} \times \frac{\prod_{k=1}^{K^p} \Gamma(|\{u : c_u^p = k\}| + \alpha^p)}{\Gamma(m + K^p \alpha^p)} \tag{3.4}$$

Based on the marginal joint probability in equation 3.1 and computed as above (see equations 3.2 and 3.4), one can derive conditional probability for sample clusters:

$$P(c_i^p = q | c_{-i}^p, c^s, c^e, c^v, S, E, V, H) \propto P(c_i^p = q, c_{-i}^p, c^s, c^e, c^v, S, E, V, H)$$

Using the property that $\Gamma(y + 1) = y\Gamma(y)$ and keeping only the terms of marginal joint probability that depend on q, we have:

$$\begin{aligned}
&P(c_i^p = q|c_{-i}^p, c^s, c^e, c^v, S, E, V, H) \propto \\
&(|\{l : c_l^p = q, l \neq i\}| + \alpha^p)\times \\
&\prod_{t=1}^{K^s} \frac{\prod_{z\in\{0,1\}} \frac{\Gamma(\bar{\mathbf{s}}_q^t(z)+\beta_z^s+[c_i^p\neq q]\,\hat{\mathbf{s}}_i^t(z))}{\Gamma(\bar{\mathbf{s}}_q^t(z)+\beta_z^s-[c_i^p=q]\,\hat{\mathbf{s}}_i^t(z))}}{\frac{\Gamma(\sum_{z\in\{0,1\}}\bar{\mathbf{s}}_q^t(z)+\beta_z^s+[c_i^p\neq q]\,\hat{\mathbf{s}}_i^t(z))}{\Gamma(\sum_{z\in\{0,1\}}\bar{\mathbf{s}}_q^t(z)+\beta_z^s-[c_i^p=q]\,\hat{\mathbf{s}}_i^t(z))}}\times \\
&\prod_{t=1}^{K^e} \frac{\prod_{z\in\{-1,0,1\}} \frac{\Gamma(\bar{\mathbf{e}}_q^t(z)+\beta_z^e+[c_i^p\neq q]\,\hat{\mathbf{e}}_i^t(z))}{\Gamma(\bar{\mathbf{e}}_q^t(z)+\beta_z^e-[c_i^p=q]\,\hat{\mathbf{e}}_i^t(z))}}{\frac{\Gamma(\sum_{z\in\{-1,0,1\}}\bar{\mathbf{e}}_q^t(z)+\beta_z^e+[c_i^p\neq q]\,\hat{\mathbf{e}}_i^t(z))}{\Gamma(\sum_{z\in\{-1,0,1\}}\bar{\mathbf{e}}_q^t(z)+\beta_z^e-[c_i^p=q]\,\hat{\mathbf{e}}_i^t(z))}}\times \\
&\prod_{t=1}^{K^v} \frac{\prod_{z\in\{-2,-1,0,1,2\}} \frac{\Gamma(\bar{\mathbf{v}}_q^t(z)+\beta_z^v+[c_i^p\neq q]\,\hat{\mathbf{v}}_i^t(z))}{\Gamma(\bar{\mathbf{v}}_q^t(z)+\beta_z^v-[c_i^p=q]\,\hat{\mathbf{v}}_i^t(z))}}{\frac{\Gamma(\sum_{z\in\{-2,-1,0,1,2\}}\bar{\mathbf{v}}_q^t(z)+\beta_z^v+[c_i^p\neq q]\,\hat{\mathbf{v}}_i^t(z))}{\Gamma(\sum_{z\in\{-2,-1,0,1,2\}}\bar{\mathbf{v}}_q^t(z)+\beta_z^v-[c_i^p=q]\,\hat{\mathbf{v}}_i^t(z))}},
\end{aligned}\tag{3.5}$$

where $\hat{\mathbf{x}}_i^t(z)$ indicates the number of features of patient i that are in feature cluster t of data type $x \in \{s, e, v\}$ and have a value equal to z, and [.] is the Iverson bracket, which is equal to 1 if the expression within the bracket is true and 0 otherwise.

The first term of the right hand side of equation 3.5 accounts for the sizes of clusters (i.e., larger clusters are assigned greater probability). The three other terms correspond to the likelihood of the observed data for patient i given the parameters of the biclusters that correspond to patient cluster q.

Due to computational overhead of the Γ function, we use a faster method and estimate equation 3.5 based on the expected value of the model parameters conditional on all data excluding the data related to patient i. This is equivalent to using the posterior predictive distribution of the variables related to patient i. Then, equation 3.5 can be rewritten as:

$$\begin{aligned}
&P(c_i^p = q|c_{-i}^p, c^s, c^e, c^v, S, E, V, H) \propto \\
&(|\{l : c_l^p = q, l \neq i\}| + \alpha^p)\times \\
&\prod_{t=1}^{K^s} \frac{\prod_{z\in\{0,1\}}(\bar{\mathbf{s}}_q^t(z) + \beta_z^s - [c_i^p = q]\,\hat{\mathbf{s}}_i^t(z))^{\hat{\mathbf{s}}_i^t(z)}}{\left(\sum_{z\in\{0,1\}}(\bar{\mathbf{s}}_q^t(z) + \beta_z^s - [c_i^p = q]\,\hat{\mathbf{s}}_i^t(z))\right)^{\sum_{z\in\{0,1\}}\hat{\mathbf{s}}_i^t(z)}}\times \\
&\prod_{t=1}^{K^e} \frac{\prod_{z\in\{-1,0,1\}}(\bar{\mathbf{e}}_q^t(z) + \beta_z^e - [c_i^p = q]\,\hat{\mathbf{e}}_i^t(z))^{\hat{\mathbf{e}}_i^t(z)}}{\left(\sum_{z\in\{-1,0,1\}}(\bar{\mathbf{e}}_q^t(z) + \beta_z^e - [c_i^p = q]\,\hat{\mathbf{e}}_i^t(z))\right)^{\sum_{z\in\{-1,0,1\}}\hat{\mathbf{e}}_i^t(z)}}\times \\
&\prod_{t=1}^{K^v} \frac{\prod_{z\in\{-2,-1,0,1,2\}}(\bar{\mathbf{v}}_q^t(z) + \beta_z^v - [c_i^p = q]\,\hat{\mathbf{v}}_i^t(z))^{\hat{\mathbf{v}}_i^t(z)}}{\left(\sum_{z\in\{-2,-1,0,1,2\}}(\bar{\mathbf{v}}_q^t(z) + \beta_z^v - [c_i^p = q]\,\hat{\mathbf{v}}_i^t(z))\right)^{\sum_{z\in\{-2,-1,0,1,2\}}\hat{\mathbf{v}}_i^t(z)}}
\end{aligned}\tag{3.6}$$

The exponents in this equation become cheap multiplications in log space.

Gene clusters for different data types are sampled similarly. As an example, equation 3.7 is the conditional probability of feature clusters according to gene expression data.

$$
\begin{aligned}
&P(c^e_j = q|c^p, c^e_{-j}, c^s, c^v, S, E, V, H) \propto \\
&(|\{l : c^e_l = q, l \neq j\}| + \alpha^e)\times \\
&\prod_{t=1}^{K^p} \frac{\prod_{z\in\{-1,0,1\}}(\bar{\mathbf{e}}^q_t(z) + \beta^e_z - [c^e_j = q]\,\tilde{\mathbf{e}}^t_j(z))^{\tilde{\mathbf{e}}^t_j(z)}}{\left(\sum_{z\in\{-1,0,1\}}(\bar{\mathbf{e}}^t_q(z) + \beta^e_z - [c^e_j = q]\,\tilde{\mathbf{e}}^t_j(z))\right)^{\sum_{z\in\{-1,0,1\}}\tilde{\mathbf{e}}^t_j(z)}},
\end{aligned}
\tag{3.7}
$$

where $\tilde{\mathbf{e}}^t_j(z)$ is the number of patients that belong to patient cluster t and have value z for feature j in gene expression dataset. It is possible to efficiently keep the track of counts $\bar{\mathbf{x}}$, $\hat{\mathbf{x}}$ and $\tilde{\mathbf{x}}$ by updating them only when a cluster assignment changes.

The random initialization of cluster latent variables c^p, c^s, c^e and c^v produces a uniform distribution of entities to the clusters. However, according to the terms included in above conditional probabilities, sampling tends to minimize the number of clusters such that the members of a cluster are highly similar. So, as the biclustering converges throughout the iterations, some clusters become empty with no entities assigned to them, if the values for K^s, K^s, K^e and K^v are set large enough. Accordingly, after each execution of learning algorithm the natural number of clusters can be determined as the number of occupied clusters. In terms of the computational complexity, each iteration of sampling, which respectively samples all c^s, c^e, c^v and c^p, is equal to $O(\max\{r\sum_x n^x, \sum_x(r + n^x)K^pK^x\})$ where $x \in \{s, e, v\}$.

3.2.3 Consensus Clustering and Parameter Estimation

Due to the stochastic nature of Gibbs sampling, the results of two distinct executions can be different. Therefore, as in [57] and [75], a consensus method based on repeated execution of the learning algorithm is used to yield a more robust clustering. This method is based on a similarity matrix, where the similarity is measured as the number of times (out of several executions) that two entities (samples or genes) belong to the same cluster at the end of an execution. Then, the consensus matrices (one for each dimension) are used to perform UPGMA hierarchical clustering to identify the final sample and gene clusters. The number of clusters used for hierarchical clustering is the average of the number of clusters occupied at the end of different executions. After finding the final clustering structures, the model parameters can be computed using maximum *a posteriori* estimation.

3.2.4 Comparison Partner

To compare the performance of the proposed probabilistic model with deterministic methods, we use a popular method for patient stratification based on Non-negative Matrix Factorization (NMF). We used the multiplicative NMF algorithm of Lee and Seung [75]. We downloaded the MATLAB implementation by Zhang et al. [146], who modified and used the algorithm for biclustering genomic and transcriptomic data. We amended the code to produce consensus matrices for further post-processing described in section 3.2.6.

3.2.5 Evaluation

Between two main categories of internal and external measures used to evaluate clustering results, we used external measures, which are more suitable for assessing the performance of patient or gene clustering algorithms [102]. According to the goal of patient stratification, different patient groups are expected to exhibit distinctive responses to treatments. Therefore, for evaluating the patient clustering results, we use clinical data and perform survival analysis. We use the log-rank test [89] implemented in R 'survival' package [129]. The smaller the log-rank *p-value*, the more distinct the survival behavior of different patient clusters. We note that, despite the popularity of this metric, a small *p-value* does not always indicate that all clusters are mutually distinct, but we observe small *p-value* for cases when there are only two distinct behaviors among more than two clusters (i.e. clusters can be further grouped based on survival). However, this metric is still useful for comparison purposes.

Since the main goal of this study is sample stratification, we also measure the stability and robustness of sample clustering outputs in terms of the Cophenetic Correlation Coefficient using the method described by Brunet et al. [22]. This is a measure between 0 and 1 and approaches 1 as results of a method are more reproducible and robust. Since almost all of the features of the datasets used in our experiments are genes, the Gene Ontology Term Overlap (GOTO) [94] criterion is used for evaluating the feature clustering. This metric measures the average within-cluster consistency of biological functions of the member genes. First, for each gene cluster, the overlap between GO terms associated with all pairs of genes is computed and averaged. The the average of all of these cluster-specific values is reported as GOTO. Larger values of this metric imply more meaningful clustering in terms of biological relationships between cluster members.

3.2.6 Hyper-parameter Tuning

One hyper-parameter for NMF is the number of clusters. To determine the best number of clusters for NMF, the method proposed by Brunet et al. [22] is used, which is based on the Cophenetic Correlation Coefficient briefly described in section 3.2.5. Similar to method described in section 3.2.3 for B2PS, a consensus matrix is computed throughout execution of NMF for the same number of times as for B2PS. This experiment is repeated with different numbers of clusters and the Cophenetic Correlation Coefficient is recorded for each experiment. Finally, a chart showing the trend of the Cophenetic Correlation Coefficient versus the increasing number of clusters is drawn and the number after which the coefficient value decreases considerably is chosen as the optimal number of clusters.

The hyper-parameters of B2PS include patient clustering hyper-parameter α^p, gene clustering hyper-parameters α^s, α^e and α^v and hyper-parameters for bicluster data distributions λ^s, λ^e and λ^v. α^p is common among all datatypes, however, gene clustering and bicluster value distribution priors are distinct for different datatypes. Bicluster data prior distribution hyper-parameters are set according to their real distribution in the corresponding datatype. Their magnitude is controlled using an scaling coefficient which is tuned.

All hyper-parameters are tuned through a grid search with the evaluation metrics discussed in section 3.2.5 and with a higher weight assigned to log-rank *p-value*. First, tuning is performed for each datatype independently. For integrated analysis of mutiple datatypes, the prior settings of individual data types are used. For common hyper-parameter α^p, the value used for the datatype producing the best stratification (i.e. gene expression) is used.

3.3 Experimental Results

3.3.1 Data

Data for this research are obtained from The Cancer Genome Atlas (TCGA) online dataset [1]. This includes genomic data, namely somatic point mutation and genome-wide copy number variation, and transcriptomic gene expression data. These information are collected from Glioblastoma Multiform (GBM) and Breast Invasive Carcinoma (BRCA) patients. For each disease, datasets for a subset of patients/samples having records for all three datatypes mentioned above are downloaded. To be analyzable with our method, these dataset are preprocessed into three matrices where rows refer to samples and columns refer to features (i.e., genes or miRNAs). According to different properties of the three datatypes, different preprocessing methods are used. Final values are 0 (for genes not containing any non-silent mutation) and 1 (otherwise) for point mutation data, -2, -1, 0, 1, 2 (the change in the normal number of copies of a gene or miRNA computed by GISTIC2.0 [92]) for CNV, and -1 (under-expression), 0 (no change), and +1 (over-expression) for gene expression data capturing expression changes more than two fold compared to normal tissue. Number of features of preprocessed final datasets for somatic point mutation, CNV, and expression data were respectively 4117, 23082, 11874 for 102 GBM samples and 13776, 23082, and 17814 for 501 BRCA samples. Clinical data were also available for the patients and contained information required for survival analysis, i.e. overall survival. We retrieved gene ontology data for GOTO analysis using the 'biomaRt' R package [39].

Because NMF only accepts non-negative values, for experiments with NMF these data are further preprocessed using the method described in [69]. In this process, first the number of columns of the dataset is doubled with two of the columns corresponding to one gene. One of the two columns stores the originally positive values and the other is dedicated for the absolute value of the originally negative values. We note that this transformation is not expected to have a negative effect on NMF's performance. In the contrary, it increases that chance of separating patient strata associated with positive values across a set of genes from other strata with negative values.

3.3.2 Results

The experiments are designed with three goals in mind: 1) to show the benefit of the ability to incorporate prior hyper-parameters enabled by the Bayesian approach, 2) to identify the best combination of datatypes for patient stratification, and 3) to compare the proposed method with a state-of-the-art method. In all experiments, the learning algorithm is executed 50 times for both B2PS and NMF and the consensus results are computed as described. Also, the initial number of clusters is set 20 for patients and about 100 for genes in each data type.

Effects of Prior Hyper-parameters

To investigate the effects of priors on performance of B2PS, different combinations of large and small values for different hyper-parameters are examined. As an example, the results of a subset of different possible settings for GBM expression dataset are shown in Table 3.2. Since the main goal of this research was sample stratification, final selected priors (shown in bold in table) favor better sample clustering over better gene clustering. According to these and similar results for the BRCA dataset (not reported here), large hyper-parameters for bicluster data distribution increase the performance regarding the sample clustering. This can be explained by the fact that strong priors cancel part of the noise of gene expression data, which generally, is expected to increase the sizes of sample and gene clusters. For sample clusters, this effect is somewhat attenuated according to strong patterns in expression profiles of each cluster and the number of clusters remain the same as when a small data hyper-parameter is used. However for gene clusters, this effect merges more similar gene clusters resulting in fewer clusters.

Large hyper-parameters for clustering have a reverse effect on clustering structure. As the clustering hyper-parameters increase, we should expect smaller and more precise clusters and, consequently, larger number of clusters. Because the number of genes is much more than the number of patients, the conditional probability for patient clustering is much more influenced by data rather than the clustering hyper-parameter. Therefore, hyper-parameter of gene clustering has more effect than sample cluster hyper-parameter. Generally the results

Priors			Sample	Feature	Log-rank	GOTO
Data	Sample Clustering	Gene Clustering	Clusters	Clusters	*p-value*	
small	small	small	8	66	0.018	3.44
large	small	small	8	25	**0.004**	3.41
large	large	small	9	21	0.017	3.40
large	small	large	8	73	0.019	3.42
large	large	large	8	70	0.008	3.42

Table 3.2: Different *a priori* hyper-parameter settings for experiments with GBM gene expression dataset

endorse the usefulness of ability to include prior knowledge about data noise in patient stratification.

Informative Datatypes for Patient Stratification

To identify the most informative datatypes for patient stratification we examined different combinations of three datatypes: somatic point mutation, CNV and gene expression. Results are summarized in Table 3.3 for GBM and BRCA datasets. Here, no results are reported for point mutation data, because, due to high heterogeneity of these data, independent experiments with point mutation dataset did not converge to any stable results. Moreover, point mutation data did not have any effects on the output when integrated with other datatypes.

According to the results, when used as the only input, gene expression data produces the best result with respect to sample clustering (log-rank *p-value*). This can be related to the fact that gene expression profiles are closer to the final phenotypes and reflect the cumulative effects of molecular aberrations including point mutations and CNVs which occur in earlier stages of the central dogma of biology.

In addition, gene clusters based on gene expression are associated with the highest GOTO score. This is consistent with the fact that genes with similar expression patterns across different samples are more likely to share the same functions in cell compared to the genes with similar CNV or point mutation. This is due to the dependency of CNV to the location of gene on the genome which results in non-deleterious CNVs. For point mutations, one expects very rare co-occurrence of the functionally related genes within the same pathway due to mutual exclusivity [142].

Moreover, according to the results, combination of expression and CNV data types introduces noise and decreases the robustness (the Cophenetic Correlation Coefficient) of the results and, deteriorates performance of sample and gene clustering compared to when gene expression is used alone. This is related to the inconsistency between different data types and the fact that different genotypes can be transcribed and translated into similar phenotypes.

Dataset	Data	Sample	Feature Clusters		Log-rank	Cophenetic	GOTO	
	Types	Clusters	Exp.	CNV	*p-value*	Corr. Coef.	Exp.	CNV
GBM	Exp.	8	25	NA	0.004	0.96	3.41	NA
GBM	CNV	19	NA	86	0.410	0.98	NA	1.82
GBM	Both	7	22	68	0.290	0.80	3.40	1.80
BRCA	Exp.	8	69	NA	0.140	0.94	2.60	NA
BRCA	CNV	20	NA	63	0.350	0.91	NA	1.85
BRCA	Both	11	69	68	0.540	0.90	2.58	1.86

Table 3.3: Results of integrative and single input experiments for GBM and BRCA

Dataset	Method	Sample	Feature	Log-rank	Cophenetic	GOTO
		Clusters	Clusters	*p-value*	Corr. Coef.	
GBM	B2PS	8	25	0.004	0.96	3.41
GBM	NMF	3	3	0.460	0.97	2.54
GBM	B2PS	3	29	0.047	0.97	3.41
GBM	B2PS	3	6	0.220	1.00	3.39
BRCA	B2PS	8	69	0.140	0.94	2.60
BRCA	NMF	3	3	0.230	0.99	2.54
BRCA	B2PS	3	101	0.120	1.00	2.60
BRCA	B2PS	3	6	0.490	0.98	2.55

Table 3.4: Comparison between B2PS and NMF

B2PS versus NMF

Comparison between the proposed method and NMF is conducted using gene expression data, which is detected here as the most informative datatype for patient stratification. To identify the number of clusters of NMF, the method described in section 3.2.6 is used. The results of NMF with the selected number of clusters and B2PS with the detected number of clusters are included in Table 3.4 for GBM and BRCA datasets. According to the results, B2PS produces more meaningful stratification and feature clusters in general, and specifically for GBM dataset.

In another experiment to evaluate the number of patient clusters detected by B2PS, B2PS is forced to produce the same number of subtypes as detected by NMF. Results shown in Table 3.4 indicates that the original number of clusters detected by B2PS results in a better stratification and, interestingly, when the number of patient clusters of B2PS is restricted, the number of detected gene clusters increases while maintaining the same GOTO score. To examine if this flexibility in the number of clusters across two different dimensions is an advantage that is effective in superior performance of B2PS, the results are compared with the case when this flexibility is discarded by restricting both patient and gene clustering. For this, the numbers of patient and gene clusters are set to equivalent values for both methods. Since, unlike NMF, B2PS inputs consists of both negative and positive values, then equivalent setting for B2PS is when the number of gene clusters is twice the number of patient clusters. This is due to the preprocessing of data for NMF as described in section 3.3.1. As a result of this process, the final NMF gene clusters might contain both over- and under-expressed genes for a particular patient cluster as opposed to B2PS that will only include one of the two possibilities. The results of these restricted experiments are also included in Table 3.4. As it can be seen, this additional restriction distorts the performance in both aspects of sample and feature clustering considerably. Accordingly, these results imply that flexibility in the number of clusters improves the performance of B2PS.

Chapter 4

Supervised Patient Stratification

Patient stratification methods are key to the vision of precision medicine. As discussed earlier, patient strata are expected to be different with respect to their phenotypes. Therefore, we consider incorporating a phenotype data to segment the patient population into subsets relevant to the given phenotype. As discussed in section 2.2, most of the existing methods for supervised patient stratification are not generally applicable due to specific assumption about the phenotype, i.e. specific models for single phenotypes or considering multiple phenotypes. Moreover, the focus is more on the prediction performance than patient stratification in some of the methods.

In the previous chapter, we observed that transcriptional data provide better stratification. Also, we found that this type of omics data is very popular for patient stratification. Therefore, we consider using transcriptional data in this chapter. However, the proposed model can be applied to any data that is or can be converted to binary.

Most of the many thousands of measured transcripts will not be related to the desired phenotype directly but rather fulfill other biological functions. As the number of samples is generally small compared to the number of transcript, it is difficult to distinguish irrelevant measurements from relevant ones. Consequently, a key task is to reliably identify and weight transcriptional features based on their relevance to the target phenotype and use these weights for patient stratification in a predictive setting.

Considering these issues, we introduce a Bayesian method called SUBSTRA that uses regularized non-parametric biclustering to identify patient subtypes and interpretable subtype-specific transcript clusters. Whereas most existing patient stratification methods focus either on predictive performance or interpretable features, we developed a method striking a balance between these two important goals. The method iteratively re-weights feature importance to optimize phenotype prediction performance by producing more phenotype-relevant patient subtypes. To the best of our knowledge, SUBSTRA is the first method that provides all of the following features:

- *Producing phenotype-relevant subtypes*: SUBSTRA includes phenotype data in the patient stratification process to identify subtypes with distinct phenotype-relevant mechanisms.

- *Producing phenotype-relevant transcript weights and clusters*: The transcript weights are learned using a Gradient Descent (GD) approach minimizing the phenotype prediction error. The transcript clusters are dependent to the phenotype-relevant subtypes and, consequently, to the phenotypes.

- *Noise handling*: The probabilistic Bayesian approach captures data uncertainty by estimating local distribution parameters.

- *Providing good interpretability-accuracy trade-off for phenotype prediction*: SUBSTRA learns a biclustering model and feature weights that simultaneously optimize two objectives: (1) the posterior probability of biclustering variables given the data and the transcript weights, and (2) the prediction error given the data and the biclustering variables. The former objective corresponds to interpretability and the latter to accuracy.

We investigate the performance of SUBSTRA in finding relevant features using simulated data and successfully benchmark it against state-of-the-art unsupervised stratification methods and supervised alternatives. Moreover, SUBSTRA achieves predictive performance competitive with the supervised benchmark methods and provides interpretable transcriptional features in diverse biological settings, such as drug response prediction, cancer diagnosis, or kidney transplant rejection. The R code of SUBSTRA is available at https://github.com/sahandk/SUBSTRA.

4.1 Problem Definition

Given a transcriptomic matrix $E \in \{0,1\}^{m \times n}$ for m patients and n transcripts as well as a phenotype $f \in Y^m$, where Y is the set of possible categorical values for the phenotype, we are interested in computing patient subtypes, transcript clusters and transcript weights such that:

1. Patients of each subtype have similar phenotypes (phenotype mislabeling is handled through a penalty). This assumption leads to phenotype-relevant subtypes and transcript weights.

2. Each subtype is associated with a local expression pattern across a subset of transcripts.

3. These patterns are unique for each subtype but might be noisy and based on only a few transcripts.

4. The relevant transcripts corresponding to the local patterns are weighted more than others to boost the signal for the biclustering and enable the identification of the correct subtype structure.

4.2 Methods

SUBSTRA performs two tasks in an iterative way: biclustering and feature weighting. At each iteration, biclustering produces patient strata as well as transcript clusters. The feature weighting task leverages the phenotype data to weight the transcripts according to their relevance to the phenotype. The relevance is identified as the contribution of the feature to prediction accuracy. The weights are then used for biclustering in the next iteration. The two tasks are elaborated in the following sections.

4.2.1 Biclustering

Our method extends the biclustering approach of Khakabimamaghani and Ester [63], called B2PS (Bayesian Biclustering for Patient Stratification). Similar to that method, we assume that (1) there is a cluster variable per patient $1 \leq i \leq m$ indicated by c_i^p, (2) there is a cluster variable per transcript $1 \leq j \leq n$ indicated by c_j^t, (3) the numbers of patient and transcript clusters are not necessarily equal, (4) the clustering is exhaustive and exclusive (*i.e.*, each patient/transcript belongs to exactly one cluster), and (5) variance of the values inside a bicluster is minimal (*i.e.*, biclusters with constant values).

To introduce supervision to patient stratification, we extend the B2PS model by two random variables: phenotype data and transcript weights. All model variables are connected to and exert influence on each other in the resulting model shown in Figure 4.1. These variables and their dependencies are elaborated in the next section. In addition, unlike B2PS which needed an upper bound for the number of clusters as input, we use a non-parametric Bayesian solution based on Chinese Restaurant Process (CPR) for inferring the natural number of patient and transcript clusters automatically.

The probabilistic graphical model of SUBSTRA is shown in Figure 4.1. All of the distributions and variables of this model are described in detail in Table 4.1. The central assumption is that the expression level of transcript j of patient i, which is indicated by E_{ij}, follows a probability distribution with parameter $\theta_{(c_i^p, c_j^t)}$ associated to bicluster (c_i^p, c_j^t). Depending on whether continuous or discrete expression data is considered, the probability distribution of variable E_{ij} can be Gaussian or categorical. We choose to use categorical expression data for two reasons: (1) using categorical data, modeled through a multinomial distribution, instead of continuous data, modeled through a Gaussian distribution, reduces the computational costs considerably due to simpler functional forms and parameters, and (2) discrete expression data have been shown to improve the prediction accuracy and generality of the trained model (e.g., applicability to different array platforms) [54, 62].

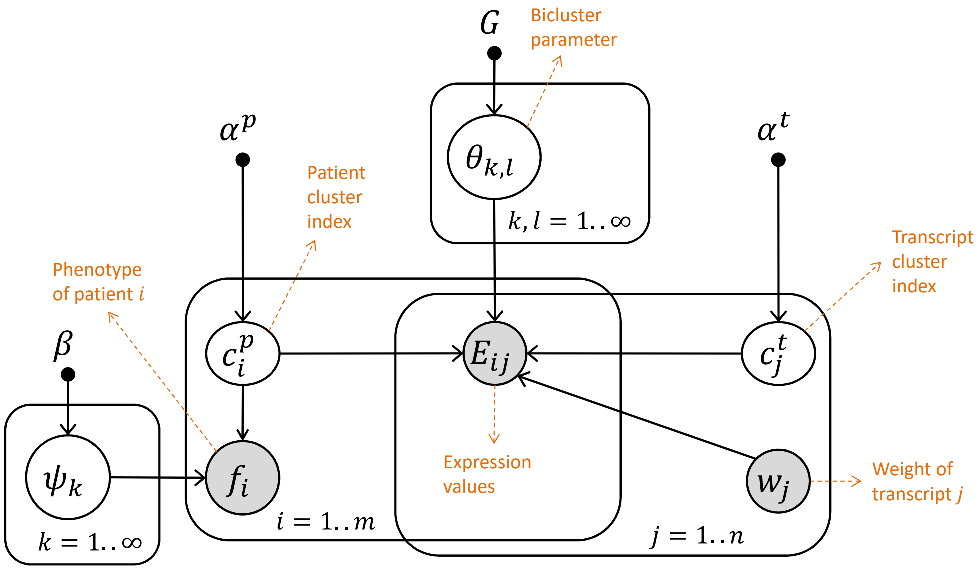


Figure 4.1: The probabilistic graphical model of SUBSTRA. The observed variables are shown with shaded circles and hyper-parameters are indicated by solid small circles. Other variables and parameters are shown with white circles. Please refer to the text for detailed explanation.

We assume binary expression values where 0 indicates low and 1 indicates high expression levels. So, E_{ij} follows a Bernoulli distribution in SUBSTRA.

4.2.2 Feature Weighting

In addition to the transcriptomic data, SUBSTRA incorporates the following information:

- *Phenotype information:* Phenotype of patient i shown by f_i. This information can be drug response, treatment effect, disease status, survival time, genetic risk score, etc.

- *Transcript weights:* A vector $w = [w_j]$ $(1 \leq j \leq n)$ of real values assigned to transcripts 1 to j. To compensate for the low influence of a single phenotype compared to the high dimensionality of the transcriptomic data, SUBSTRA propagates the effect of phenotype using phenotype-relevant transcript weights. Each weight is interpreted as the number of times that the corresponding transcript is considered during the biclustering. Thus, the higher the weight of a transcript, the stronger its effect on the biclustering. This variable is considered observed (shaded) in Figure 4.1, because, unlike the model latent variables that are inferred based on the joint probability of the model, we learn the transcript weights using a different objective function (*i.e.*, prediction error) based on a Gradient Descent approach. More details are provided in section 4.2.3.

Type	**Name**	**Description**	**Distribution**
Observed Variables	E_{ij}	Expression status of transcript j of patient i	$E_{ij} \sim Bern.(\theta_{c_i^p, c_j^t})$
	f_i	Phenotype of patient i	$f_i \sim Cat.(\psi_{c_i^p})$
	w_j	Weight of transcript j	Initiated to μ
Hyper-parameters	α^p	Parameter of prior CRP for patient clusters	$\alpha^p = 1$
	α^t	Parameter of prior CRP for transcript clusters	$\alpha^t = 1$
	G	Parameter for prior Beta base distribution of θ	$G = 1$
	β	Parameter for the prior Beta base distributions of ψ	Described in section 4.2.3
	μ	Gradient descent learning rate and initial transcript weights	Described in section 4.2.3
Parameters	$\theta_{k,l}$	Probability distribution of the values inside bicluster (k, l)	$\theta_{k,l} \sim Beta(G)$
	ψ_k	Probability distribution of the values of the phenotypes of patient cluster k	$\psi_k \sim Dirichlet(\beta)$
Latent Variables	c_i^p	Cluster index for ith patient	$c_i^p \sim CRP(\alpha^p)$
	c_j^t	Cluster index for jth transcript	$c_j^t \sim CRP(\alpha^t)$

Table 4.1: Variables and probabilistic relationships in the SUBSTRA model.

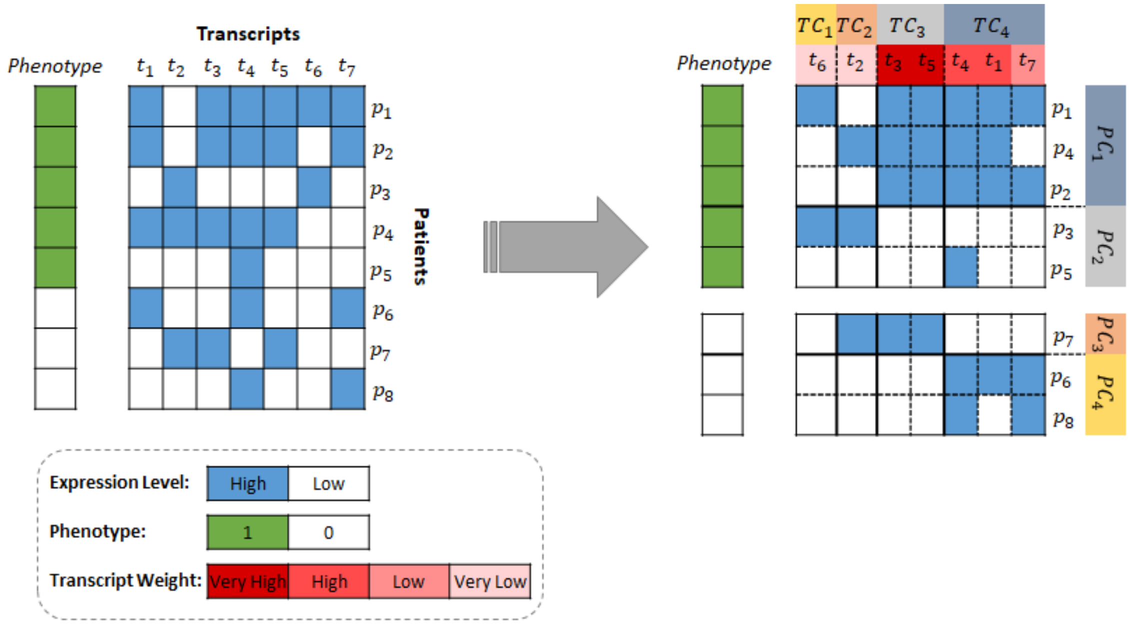


Figure 4.2: Input and output of SUBSTRA. The matrix on the left is reordered into the matrix on the right by SUBSTRA. The patients and transcripts are assigned to appropriate clusters and the transcript weights indicate the significance of features with regard to the phenotype. The patient and transcript clusters are formed in a way that the values inside biclusters are as consistent as possible, especially for those biclusters that are related to transcripts with higher weights. High-weight transcripts are those that form a biclustering more consistent with the phenotypes. For example, using the combination of transcripts in TC_3 and TC_4, one can produce the four patient clusters with homogeneous phenotypes (*i.e.* PC_1 to PC_4) as shown in the figure. So, the TC_3 and TC_4 transcripts are assigned high weights. On the other hand, t_2 and t_6 cannot form a consistent patient clustering when used alone or in combination with other transcripts and get low weights. Although in this sample the number of patient clusters is equal to the number of gene clusters, this is not a constraint in our algorithm.

As shown in Figure 4.1 and Table 4.1, phenotype of patient i indicated by f_i, follows a subtype-specific distribution with parameter $\psi_{c_i^p}$. Furthermore, the transcript weights w_j influence the biclustering variables through expression variable E_{ij}. The information flow between the transcript weights and phenotypes are through E_{ij} and c_i^p variables (this is possible because E_{ij} is observed and c_i^p is latent). We use this information flow to adjust the transcript weights as described in section 4.2.3. Without loss of generality, we assume that phenotype is a binary variable following a Bernoulli distribution in the current work. In practice, any distribution could be used based on the type of phenotype. A sample input for SUBSTRA and the expected output is shown in Figure 4.2.

4.2.3 Parameter Learning and Inference

Parameter learning and inference are performed via Gibbs sampling.

The sampler infers the latent variables and learns the transcript weights simultaneously. The algorithm consists of three below phases.

Phase 0 (Initialization)

The latent variables of SUBSTRA (*i.e.* patient clusters c_i^p and transcript clusters c_j^t) are initialized randomly, such that two patients with different phenotypes are not assigned to the same cluster. This constraint satisfies the assumption 3 stated in section 4.1 during the initialization. However, the strictness of the constraint during the sampling can be controlled by the hyper-parameter β. If the mislabeling rate is low in the observed phenotypes, we should set the hyper-parameter β to a small value to make this constraint stricter. Otherwise, a larger β is used. The transcript weights are all initialized equal to μ, which is an input and indicates the magnitude of weights. If μ is large, the algorithm will be more sensitive to the values of transcript expressions and will fit faster to the data, increasing the probability of over-fitting or local optima. This works for the cases with strong relevant signals. On the other hand, when there are strong irrelevant signals in the data, a smaller μ is preferred as it provides more flexibility and increases the exploration space. We use cross-validation to tune this value. The value that produces more accurate phenotype prediction is selected, because higher accuracy implies more relevant biclustering and weighting.

Phase I

In this step, only the latent variables are sampled and the transcript weights are fixed. This is required since the initial random values of parameters can be misleading if used for adjusting the weights. In this phase, the Gibbs sampler uses the conditional probabilities of the latent variables. The conditional probabilities are computed based on the joint probability, which factorizes as below:

$$\begin{aligned}
&P(E, w, f, c^p, c^t, \theta, \psi | \alpha^p, \alpha^t, \beta, G) \\
&= P(c^p|\alpha^p) \times P(c^t|\alpha^t) \\
&\quad \times P(E|\theta, c^p, c^t, w) P(\theta|G) \times P(f|\psi, c^p) P(\psi|\beta)
\end{aligned}$$

Considering this dependency structure and the distributions given in Table 4.1, the conditional probabilities of latent variables are computed as below:

$$\begin{aligned}
&P(c_i^p = q | c_{-i}^p, E_{i.}, w, f_i, c^t, \theta, \psi, \alpha^p) \\
&\propto P(c_i^p = q | c_{-i}^p, \alpha^p) P(f_i | \psi, c_i^p = q) P(E_{i.} | \theta, c_i^p = q, c^t, w) \\
&= \pi(q | c_{-i}^p, \alpha^p) \times \psi_q[f_i] \times \prod_{j=1}^{n} (\theta_{c_i^p, c_j^t}[E_{ij}])^{w_j} \qquad (4.1)
\end{aligned}$$

$$\begin{aligned}
&P(c_j^t = r | c_{-j}^t, E_{.j}, w, c^p, \theta, \alpha^t) \\
&\propto P(c_j^t = r | c_{-j}^t, \alpha^t) P(E_{.j} | \theta, c_j^t = r, c^p, w) \\
&= \pi(r | c_{-j}^t, \alpha^t) \times \prod_{i=1}^{m} \theta_{c_i^p, c_j^t}[E_{ij}], \qquad (4.2)
\end{aligned}$$

where $\pi(q|c_{-i}^p, \alpha^p)$ is the CRP probability and is defined as below:

$$\pi(q|c_{-i}^p, \alpha^p) = \begin{cases} \frac{\alpha^p}{m-1+\alpha^p} & \text{if } x \text{ is an empty cluster} \\ \frac{|\{d | c_d^p = q \ \wedge \ d \neq i\}|}{m-1+\alpha^p} & \text{otherwise} \end{cases}$$

We assume that ψ_q and $\theta_{c_i^p, c_j^t}$ are simplex vectors with probabilities corresponding to every possible value of phenotype and data elements respectively. For example, $\theta_{c_i^p, c_j^t}[0] + \theta_{c_i^p, c_j^t}[1] = 1$.

We use the predictive posterior distribution parameters to estimate the model parameters θ and ψ of equations 4.1 and 4.2 as follows:

$$\begin{aligned}
\theta_{q,r}[x] &= \frac{\text{no. of } x\text{'s in bicluster } (q,r) + G/2}{\text{no. of data points in bicluster } (q,r) + G} \\
\psi_q[x] &= \frac{\text{no. of patients in cluster } q \text{ with phenotype } x + \beta/2}{\text{no. of patients in cluster } q + \beta} \\
\pi_q &= \pi(q | c^p, \alpha^p) \qquad (4.3)
\end{aligned}$$

During *Phase I*, we repeat the following for each c_i^p:

1. Estimate the parameters using equation 4.3 based on the current values of the model variables excluding $E_{i.}$, f_i, and c_i^p

2. Use equation 4.1 to sample c_i^p

Similarly for each c_j^t, we:

1. Estimate the parameters based on the current value of the model variables excluding $E_{.j}$ and c_j^t

2. Use equation 4.2 to sample c_j^t

At each Gibbs sampling round we sample all latent variables as described above. As we use CRP, we consider the possibility of belonging to an empty cluster when sampling each latent variables for patients and transcripts. The sampling round is repeated until convergence or for a predefined number of iterations. The convergence is measured based on the Rand index similarity between the biclustering in two consecutive iterations, which is achieved when Rand index > 0.95 for patient and transcript clustering. Then we move to *Phase II*.

Phase II

In this phase, we adjust the transcript weights and simultaneously modify the biclustering structure. Since the weights should indicate the relevance of a transcript to the phenotype, we use the phenotype prediction error, which is a function of the weights, as the objective function for weight adjustment. The input to this phase is the latent variable values at the end of the previous phase. In addition to the steps in *Phase I*, we adjust transcript weights before sampling each c_i^p in this phase following the below steps:

1. Estimate the parameters based on the current value of the model variables except $E_{i.}$, f_i, and c_i^p
2. Adjust the weights to reduce the phenotype prediction error for patient i
3. Use equation 4.1 to sample c_i^p

The weights are adjusted such that the objective function defined as the squared prediction error $[1 - p(f_i = x_i|...)]^2$ (x_i is the true value of f_i) is minimized. Using a Gradient Descent approach, we use the slope of this function to adjust the weights. So, the weights are updated as follows:

$$w_j = w_j + \nu \times 2 \frac{\partial p(f_i = x_i|...)}{\partial w_j} [1 - p(f_i = x_i|...)], \tag{4.4}$$

where ν is the learning rate and we set $\nu = \mu$, the magnitude of weights, to maintain the magnitude of weights.

Because the cluster assignment of patient i is unknown at this stage (*i.e.* we are about to sample it in step 3) and according to the information flow in the model (Figure 4.1), we have:

$$\begin{aligned}
&p(f_i = x_i|...)\\
&= p(f_i = x_i|\psi, E_{i.}, \pi, c^t, w, \theta)\\
&= \sum_{q \in O} p(f_i = x_i, c_i^p = q|\psi, E_{i.}, \pi, c^t, w, \theta)\\
&= \sum_{q \in O} p(f_i = x_i|c_i^p = q, \psi)p(c_i^p = q|E_{i.}, \pi, c^t, w, \theta)\\
&\propto \sum_{q \in O} p(f_i = x_i|c_i^p = q, \psi)p(c_i^p = q, E_{i.}|\pi, c^t, w, \theta),
\end{aligned}$$

where O is the set of occupied patient clusters. The second term in the last summation can be factorized based on the model (very similar to equation 4.1). Let us define:

$$p_y = \sum_{q \in O} p(f_i = y|c_i^p = q, \psi)p(c_i^p = q, E_{i.}|\pi, c^t, w, \theta),$$

where $y \in Y$ indicates one of the values that the patient phenotype can take. Then we have:

$$p(f_i = x_i|...) = \frac{p_{x_i}}{\sum_y p_y}$$

Then, the derivative term in equation 4.4 is computed as below:

$$\frac{\partial p(f_i = y|...)}{\partial w_j} = \frac{\partial \frac{p_y}{\sum_z p_z}}{\partial w_j} = \frac{(\sum_z p_z)\frac{\partial p_y}{\partial w_j} - p_y \sum_z \frac{\partial p_z}{\partial w_j}}{(\sum_z p_z)^2} \tag{4.5}$$

So, we need to compute $\frac{\partial p_y}{\partial w_j}$ for every y. We have:

$$\begin{aligned}
\frac{\partial p_y}{\partial w_j} &= \frac{\partial \sum_{q \in O} p(f_i = y|c_i^p = q, \psi)p(c_i^p = q, E_{i.}|\pi, c^t, w, \theta)}{\partial w_j}\\
&= \sum_{q \in O} p(f_i = y|c_i^p = q, \psi) \times \partial \frac{p(c_i^p = q, E_{i.}|\pi, c^t, w, \theta)}{\partial w_j}\\
&= \sum_{q \in O} \psi_q[y] \times \pi_q \times \partial \frac{\prod_{l=1}^{n} (\theta_{q, c_l^t}[E_{il}])^{w_l}}{\partial w_j}\\
&= \sum_{q \in O} \psi_q[y] \times \pi_q \times \log(\theta_{q, c_j^t}[E_{ij}]) \prod_{l=1}^{n} (\theta_{q, c_l^t}[E_{il}])^{w_l} \qquad (4.6)
\end{aligned}$$

The next step is to compute the left-hand-side of equation 4.5 based on the equation 4.6 and then use it in equation 4.4 for computing the new weights:

$$
\begin{aligned}
& w_j = w_j + \nu \times \\
& 2\frac{\sum_{q,q' \in O, q' > q} \pi_q \pi_{q'} \prod_{l=1}^{n} (\theta_{q,c_l^t}[E_{il}] \theta_{q',c_l^t}[E_{il}])^{w_l} (\psi_{q'}[x_i] - \psi_q[x_i])[\log(\theta_{q',c_j^t}[E_{ij}]) - \log(\theta_{q,c_j^t}[E_{ij}])]}{\sum_{q,q' \in O} \pi_q \pi_{q'} \prod_{l=1}^{n} (\theta_{q,c_l^t}[E_{il}] \theta_{q',c_l^t}[E_{il}])^{w_l}} \\
& \times (1 - \frac{\sum_{q \in O} \psi_q[x_i] \times \pi_q \times \prod_{l=1}^{n} (\theta_{q,c_l^t}[E_{il}])^{w_l}}{\sum_{q \in O} \pi_q \times \prod_{l=1}^{n} (\theta_{q,c_l^t}[E_{il}])^{w_l}})
\end{aligned}
\tag{4.7}
$$

In practice, this can be done more efficiently by computing the summations like equation 4.6 separately and then multiplying the results.

We note that the squared error objective function is not convex. However, since it is Lipschitz continuous, i.e. the function is bounded and differentiable for every $w \geq 0$ with a bounded slope (limit of the slope in equation 4.7 as $w \to \infty$ is 0 and the denominator never becomes 0 for other values), gradient descent can be used to find the local optima. Moreover, to guarantee continuous improvement, after each update the new weights are accepted only if they reduce the squared error. Otherwise, the algorithm keeps the previous weights and continues to the next patient.

In this phase, a certain number of iterations is executed and the model performance in terms of the Area Under the Receiver Operating Characteristic Curve (AUC) over the training set is monitored. Finally, the model that corresponds to the iteration with the highest AUC is selected. Ties are broken with respect to the Mean Squared Error (MSE) of the predicted probabilities. Although the training set AUC and MSE are used for model selection, over-fitting is avoided because the data corresponding to patient i is not included when updating the weights based on that patient.

4.3 Experimental Results

In this section we describe the experiments performed for testing the accuracy of SUBSTRA. The method produces two types of outputs: predictive outputs (predicted phenotypes) and descriptive outputs (*i.e.*, patient strata, transcript clusters, and transcript weights). We benchmark against other methods with respect to these outputs.

4.3.1 Predictive Performance Evaluation

To investigate the predictive ability of SUBSTRA, it is benchmarked against the following methods:

- Support Vector Machine (SVM): A well-known state-of-the-art prediction method with high accuracy. The implementation of SVM in R package ’e1071’ is used.

- Regularized Logistic Regression (LR): A popular prediction method that assigns model-based (not ad-hoc) weights to the predictor features. We used the Elastic Net Generalized Linear Models implementation in R package 'caret'.

- Predictive Chain (PCH): This method is evaluated as a simple baseline method that performs biclustering and prediction in two separate steps, rather than in one integrated step as SUBSTRA does. It first applies NMF[74] (a popular biclustering method) for deriving a low-rank representation of the patients and then trains the LR model on that representation. We investigate whether using NMF output will have positive or negative effects on the prediction accuracy of LR.

The Area Under the Receiver Operating Characterisitics (AUC) metric is computed to measure the prediction accuracy of all three methods through nested CV with inner 3-fold cross validation for hyper-parameter tuning and outer 5-fold cross validation for evaluation. For SVM, the *radial basis function* kernel is used and the model is tuned through grid search over the kernel parameter $\gamma \in \{10^i | -8 \leq i \leq -1\}$ and soft margin parameter $C \in \{1..5\}$. For SUBSTRA, the weight magnitude variable μ is tuned over values $\{0.0001, 0.001, 0.01, 0.1, 1, 10\}$.

4.3.2 Descriptive Performance Evaluation

We benchmarked the biclustering accuracy of SUBSTRA against similar biclustering methods that do not consider phenotype data (i.e., unsupervised patient stratification). SUBSTRA performs exhaustive and exclusive biclustering with constant values inside the biclusters. Based on a review over 47 biclustering algorithms for gene expression data provided by Pontes et al. [100], we found HARP [143] to be the most consistent method with these features. Two other comparable methods not listed in [100], include B2PS [63], which is an exhaustive, exclusive, and constant value biclustering method, and NMF.

As stated in the beginning of this chapter, many existing supervised stratification methods either leverage several phenotypes or make specific assumptions for compound phentypes, e.g. assume survival data. This makes it hard to compare SUBSTRA with those methods. Therefore, we define an additional simple baseline method that first identifies feature weights using LR. Then, the feature weights are given to weighted NMF (wNMF) [138] for biclustering. We call this method Descriptive Chain (DCH). This is to investigate the influence of the provided weights on the biclustering accuracy, as well as comparison against SUBSTRA's biclustering.

We compare SUBSTRA against HARP, B2PS, NMF, and DCH in terms of the following metrics:

- Patient Strata: Whenever the ground-truth patient clusters are available, we use Rand index to measure the patient clustering accuracy.

- Transcript Clustering: Transcripts fall into two categories of relevant (signal) and irrelevant (noise) to the phenotype. We only focus on the clustering results for the relevant transcripts. Two metrics, cluster purity and class purity are used for evaluation. Clusters refer to the outputs of the methods and classes refer to the ground-truth transcript clusters. Class purity (CSP) measures how well the true signal clusters are separated from each other by the method. Cluster purity (CLP) indicates how much of the signal transcripts are captured in the method clusters. Together, these two metrics reflect how well the method has been able to capture the true signal clusters. More details are provided in supplementary section D. For HARP, we note that it is only exclusive with regard to patient clustering and might produce overlapping transcript clusters. Thus, only CLP can be reported for this method.

- Transcript Weights: Pearson correlation coefficient between the ground-truth weights and method weights are reported when the ground-truth information is available. When unavailable, GO term enrichment analysis of the top ranked genes is used as described later.

To accommodate for random initialization, the descriptive experiments are repeated 5 times for each dataset and the weights are averaged and the clusters are identified through consensus clustering [95]. For HARP, the user should provide a lower bound for the number of clusters, for which we used the true number of clusters 4 for the AND, OR and XOR datasets and 6 for the UNCLES dataset. To idendify the number of components k for NMF, we used a method based on Mean Squared Error (MSE) of NMF-based missing value imputations. This method is provided in R package NNLM [80]. First, 20% of the matrix entries are set to missing values. Then, using the information from the remaining elements and for different values of $2 \leq k \leq 22$, missing values are imputed based on the latent factors learned by NMF and the MSE is measured. The k resulting in the smallest MSE is selected. For the synthetic AND, OR, and XOR datasets, this approach was not successful and returned $k = 2$ which was meaningless according to the structure of the datasets. Therefore, we used the true value $k = 4$ for these experiments.

4.3.3 Experiments with Synthetic Data

We used synthetic data to have access to the ground-truth information to benchmark SUBSTRA for detecting the true patient and transcript clusters, true feature weights and accurate prediction. Different synthetic datasets were generated considering the assumptions mentioned in section 4.1. In separate simulations, we tested different types of relations between the transcript clusters and the phenotype: AND, OR, and XOR. For this purpose, we assumed that the expression values of two transcript clusters A and B are correlated with the phenotype through the mentioned relations. As an example, for an XOR relationship,

PC#	TC# (Size)			Phenotype
(Size)	A (10)	B (10)	Noise (1980, 380, 180)	
1 (30)	0.7	0.7	0.5	1
2 (30)	0.7	0.23	0.5	0
3 (20)	0.1	0.8	0.5	0
4 (20)	0.3	0.3	0.5	0

Table 4.2: Parameters and cluster sizes for AND data.

PC#	TC# (Size)			Phenotype
(Size)	A (10)	B (10)	Noise (1980, 380, 180)	
1 (30)	0.17	0.7	0.5	1
2 (30)	0.76	0.17	0.5	1
3 (20)	0.7	0.8	0.5	1
4 (20)	0.3	0.3	0.5	0

Table 4.3: Parameters and cluster sizes for OR data.

the value of phenotype will be 1 if and only if the transcripts of only one of the clusters A or B are expressed.

Each dataset consists of 200 patients constituting 4 patient clusters with four different possible combinations of parameters for signals A and B (*i.e.*, A high-B high, A high-B low, A low-B high, and A low-B low). Each of these two clusters includes 10 transcripts. Bicluster parameters larger than 0.5 indicate high expression and vice versa. A third transcript cluster is included as the noise, with parameter equal to 0.5 across different patient clusters (*i.e.*, biclusters with Bernoulli distribution with parameter 0.5). The values of parameters for different settings are provided in tables 4.2, 4.3 and 4.4. The same parameters are used for simulating different noise levels. The performance of the three methods are compared for different datasets with 90%, 95%, or 99% of transcripts belonging to the noise cluster. These datasets will respectively contain 200, 400, and 2000 transcripts 20 of which are relevant signals and the rest are noise.

To avoid biases towards our own assumptions, we include another synthetic microarray dataset introduced in [2]. This dataset, to which we refer as UNCLES (the title of the paper), consists of two patient classes (positive and negative) and three gene clusters. The gene cluster $C1$ (75 genes) includes genes consistently co-expressed for all patients, and the

PC#	TC# (Size)			Phenotype
(Size)	A (10)	B (10)	Noise (1980, 380, 180)	
1 (40)	0.7	0.25	0.5	1
2 (20)	0.1	1.0	0.5	1
3 (30)	0.33	0.35	0.5	0
4 (10)	1.0	0.95	0.5	0

Table 4.4: Parameters and cluster sizes for XOR data.

gene cluster $C2$ (85 genes) includes genes consistently co-expressed only in the positive class while being poorly co-expressed in the negative class. Among the two clusters, $C1$ is more correlated with the patient classes as it has in general higher expression in the positive class and lower expression in the negative class. Accordingly, although we evaluate the methods for detecting the two clusters, we only consider $C1$ when evaluating the capabilites of the methods in up-weighting the phenotype-relevant genes. The rest of the genes (1040 genes) are poorly co-expressed everywhere and are considered noise. The dataset contains 42 positive and 40 negative patients. The UNCLES dataset contains continuous data. We use the original continuous as well as the discretized data. To monitor the sensitivity to different discretization methods, three different approaches, namely Equal-Frequency Binning (EFB), Equal-Width Binning (EWB), and k-means (KM), are used for discretization as described in [62].

Table 4.5 shows the predictive and descriptive results for different simulation settings. Among the methods, HARP and NMF has the lowest performance for most of the datasets with respect to patient stratification. Adding supervision to NMF as in DCH improves the results in high noise datasets (i.e., AND and OR 99%), however, it does not have significant effects on the other cases. B2PS and SUBSTRA perform relatively better than other methods both in our simulations and UNCLES dataset. SUBSTRA outperforms B2PS considerably (difference larger than 0.05) in high noise datasets as well as XOR relationship, which is more complex than AND and OR.

With respect to transcript clustering, HARP and NMF has similarly lower CLP in most of the cases. The reason is that both methods detect uniformly large and impure clusters. On the other hand, NMF has superior ability in separating the signal clusters from each other compared to DCH. Although, adding supervision in DCH improves cluster purity (CLP) for some low-noise datasets compared to solo NMF, it increases the chance of mixing the true signal clusters in a single transcript cluster (lower CSP). Top methods with respect to transcript clustering are B2PS and SUBSTRA, with SUBSTRA being superior in certain cases (high noise AND and EFB UNCLES). This indicates that supervision as in SUBSTRA improves the clustering quality.

Table 4.5 also shows the transcript weighting results for SUBSTRA and DCH. The values indicate the correlation between the method and the ground-truth weights. The ground-truth weights are produced by assigning weight 1 to the signal transcripts (members of A, B, and $C1$ clusters) and 0 to the other transcripts. Based on the results, SUBSTRA produces consistently more correlated weights for the synthetic data than DCH, which uses LR for weighting. This can be associated to the probabilistic nature of the method and its ability to capture more complex relationships like XOR, which are not detectable by linear methods such as LR (note the low correlation values of DCH for XOR and UNCLES). Transcript clustering in SUBSTRA can increase the weight consistency inside the transcript clusters beside improving the accuracy of the weights due to inter-cluster discrepancies. The descriptive results are visualized in supplementary section A.

Regarding the AUC measures in table 4.5, SUBSTRA also outperforms the other predictive benchmark methods in most of the experiments and is more robust to the noise levels and the task complexity. On the other hand, PCH and LR are sensitive to noise and the type of discretization and SVM is sensitive to noise but robust to the discretization method. Binary data, compared to continuous data, is associated with better performance except for the predictive accuracy of LR.

As an example visualization, figure 4.3 shows the heatmaps of SUBSTRA and B2PS results corresponding to the quantitative results shown in table 4.5 for the AND relationship. As it can be seen, B2PS does not discover the two important signal clusters A and B for a high level of noise (note the SR values) and the identified clusters are mixed (note the PR values). It also does not assign a weight to the transcript clusters. On the other hand, SUBSTRA has successfully detected the signals with the highest weights assigned to them. The purity of subtypes with respect to class label is also consistent for SUBSTRA. Visual results for the other relationships are provided in Appendix A.

4.3.4 Experiments with Real Data

We also tested SUBSTRA with real data. These datasets are listed in Table 4.6. The Kidney 1 and 2 datasets are taken from studies Khatri et al. [67] and Einecke et al. [40]. They include baseline gene expression profiles for patients before kidney transplantation and whether the patient rejected the transplantation (phenotype). We also used a dataset from the Cancer Cell Line Encyclopedia (CCLE) [10], which provides a collection of genomic information (including baseline transcriptomic data) and pharmacological profiles (including response to various drugs for several cell lines derived from different tissues). A subset of cell lines which had information about their response to AZD6244 (a drug that targets MEK, a gene mediating cellular response to growth signals) was selected from this dataset. Response to the drug was recorded in terms of IC50. We used a cut-off value of 7 to discretize IC50 values to 0 (not responding) and 1 (responding). Two datasets, namely Lung Cancer from Gordon et al. [49] and Multiple Myeloma from Tian et al. [130], were also used from the R package "datamicroarray" [104]. The package is a collection of microarray datasets with phenotypes. They are from different studies and can be used for machine learning.

All datasets are pre-processed. For each dataset, the first 5000 features with the highest coefficient of variation are selected. Then, the three mentioned discretization methods are used to binarize the continuous expression data into 0 (low) and 1 (high). These methods are non-parametric and do not depend on any threshold. Continuous data is also considered where applicable.

Since no ground-truth data are available about patient strata and transcript clusters, we only benchmarked the predictive performance and transcript weights of SUBSTRA against

Metric	Type	Method	AND			OR			XOR			UNCLES			
			90%	95%	99%	90%	95%	99%	90%	95%	99%	EFB	EWB	KM	NO
PC Rand,#	Descriptive	HARP	0.66,4	0.63,4	0.61,4	0.66,4	0.63,4	0.63,4	0.66,4	0.66,4	0.61,4	0.78,6	0.72,6	0.74,6	0.68,6
		B2PS	0.91,8	0.91,10	0.41,8	0.87,8	0.89,10	0.89,9	0.84,8	0.87,9	0.85,10	0.86,19	0.86,19	0.86,18	NA
		NMF	0.68,4	0.67,4	0.62,4	0.70,4	0.69,4	0.62,4	0.71,4	0.70,4	0.62,4	0.82,8	0.77,6	0.80,7	0.79,9
		DCH	0.70,4	0.69,4	0.70,4	0.71,4	0.71,4	0.75,4	0.69,4	0.62,4	0.61,4	0.80,8	0.75,6	0.77,7	0.74,9
		SUBSTRA	**0.96**,10	0.93,8	**0.89**,12	0.87,12	0.92,10	**0.95**,11	**0.95**,10	**0.96**,9	**0.94**,11	0.85,15	0.82,9	0.82,10	NA
CSP%,CLP%	Descriptive	HARP	NA,25	NA,07	NA,01	NA,24	NA,09	NA,02	NA,27	NA,14	NA,03	NA,15	NA,23	NA,35	NA,13
		B2PS	100,100	95,100	65,02	100,100	100,100	100,100	100,100	100,100	100,100	83,95	81,96	93,97	NA
		NMF	100,30	100,11	75,02	100,24	100,12	95,02	100,27	100,14	90,02	97,21	94,26	81,28	66,32
		DCH	90,60	85,89	85,01	80,55	50,05	65,15	80,21	70,12	95,01	81,10	78,29	97,26	94,26
		SUBSTRA	100,100	**100,100**	**100,100**	100,100	100,100	100,100	100,100	100,100	100,100	**98,97**	79,97	89,97	NA
WPC	Descr.	DCH(LR)	0.58*	0.68*	0.59*	0.30*	0.33*	0.31*	0.10	0.04	-0.01	-0.01	0.08	0.03	-0.04
		SUBSTRA	0.65*	0.72*	**0.70***	**0.59***	**0.54***	**0.47***	**0.49***	**0.44***	**0.28***	**0.21***	-0.05	0.14*	NA
AUC	Predictive	LR	0.88	0.84	0.85	0.72	0.73	0.58	0.46	0.52	0.60	0.74	0.51	0.56	0.99
		SVM	0.87	0.82	0.70	0.65	0.61	0.55	0.62	0.62	0.34	0.98	0.94	0.94	0.88
		PCH	0.78	0.71	0.53	0.78	0.68	0.63	0.44	0.44	0.40	0.94	0.51	0.55	0.81
		SUBSTRA	**0.97**	**0.97**	**0.97**	**0.91**	**87**	**0.93**	**0.91**	**0.89**	**0.88**	1.00	0.96	0.97	NA

Table 4.5: Results for the experiments with the synthetic data. Abbreviations used include PC Rand,# – Rand index for patient clustering (comparison with the ground truth) and the number of patient clusters, CSP%,CLP% – class purity and cluster purity (described in section 4.3.2) as percentage, WPC – Pearson correlation coefficient between the true weights and the method's weights with statistically significant results (after Bonferroni correction) marked by *, and NO – no discretization. Best performance in each dataset is shown in bold if the gap with the second best performance is larger than or equal to 0.05 for Rand index, purity, and AUC and larger than or equal to 0.1 for WPC when at least one of the correlations is significant.

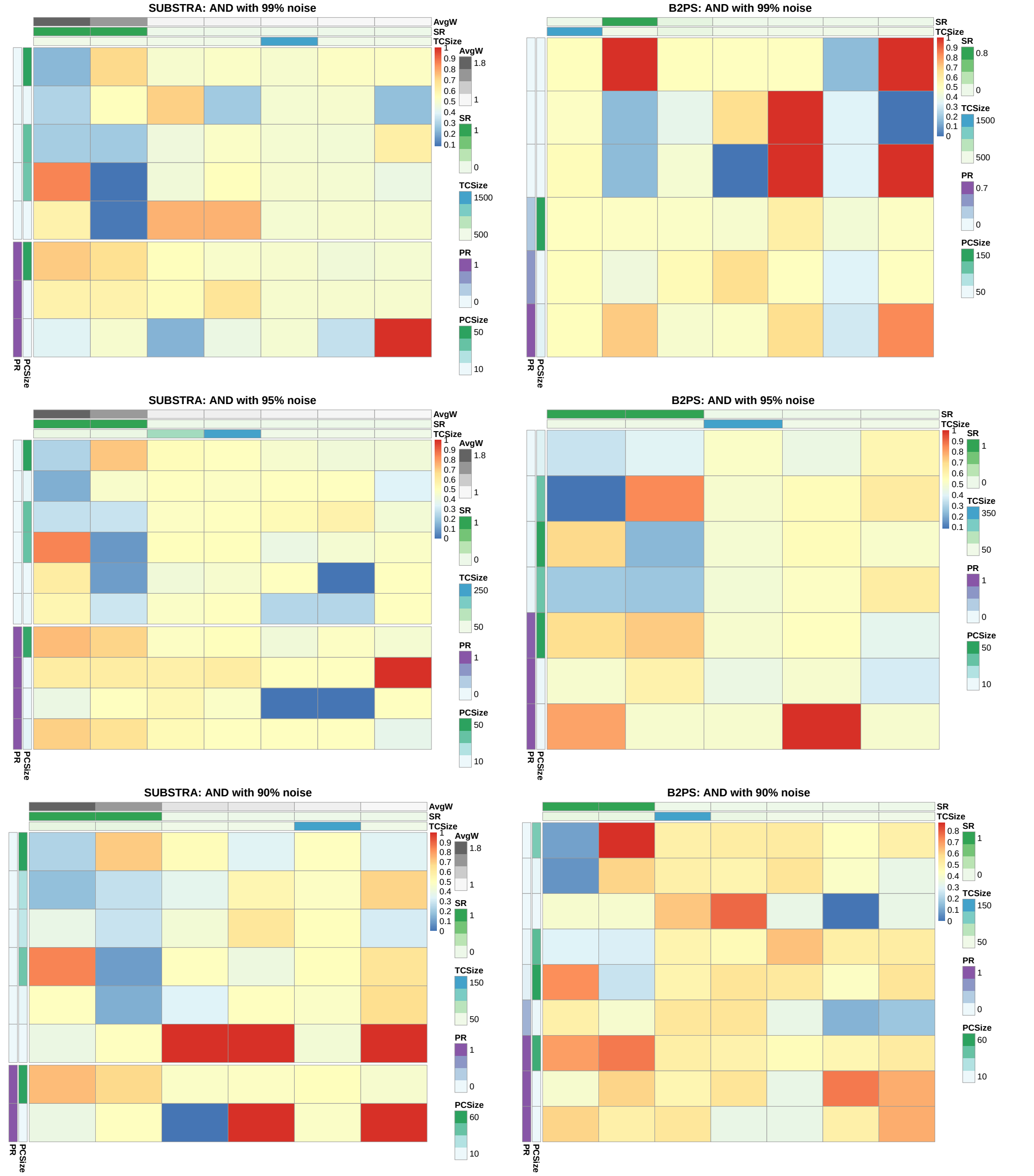


Figure 4.3: Resulted heatmaps of SUBSTRA and B2PS for synthetic data for the AND relationship. The heatmaps depict the behavior of the transcript clusters (columns) across different patient groups (rows).The value inside each cell/bicluster indicates the average expression (a value between 0 and 1), with red being high expression and blue being low expression. $TCSize$ is size of transcript cluster, $AvgW$ is average weight of the gene/transcript cluster, SR is the Signal Ratio (the proportion of signals within each transcript cluster), $PCSize$ is patient cluster size, and PR is the Phenotype Ratio (the proportion of '1' phenotypes within the patient cluster). For SUBSTRA, the transcript clusters are sorted based on the $AvgW$ from left to right in descending order.

Dataset	**#Patients**	**#Features**	**Phenotype**	**Neg.-Pos.**
Kidney 1 [40]	282	18,089	Kidney transplant response	63%-37%
Kidney 2 [67]	101	18,988	Kidney transplant response	57%-43%
Drug Response [10]	490	42,869	Response to AZD6244	26%-74%
Multiple Myeloma [130]	173	12,625	Existence of focal bone lesions	21%-79%
Lung Cancer [49]	181	12,533	MPM or ADCA	17%-83%

Table 4.6: Datasets used in the predictive and descriptive experiments

the comparison partners. All methods were executed on the same cross-validation folds and experiments were repeated and averaged to accommodate for the random initialization effects. More details are provided in supplementary section C.

Figure 4.4 shows the predictive results for the above datasets. According to these results, all methods have in general similar predictive performance when considering the best performing configuration (i.e., discretization). Looking closer LR has a slightly better performance than the others in three out of five experiments. SUBSTRA and SVM are performing similar taking all experiments into account. SUBSTRA produces more stable results than the other methods as reflected in the error bars. Considering similar discretizations, SUBSTRA performs better than the predictive alternative PCH. Using continuous data, which is not yet implemented in SUBSTRA, PCH approaches SUBSTRA, especially in 'Multiple Myeloma' and 'Drug Response' datasets. These results match those of simulation experiments and indicate that simple chaining of the existing methods does not reproduce the quality of SUBSTRA. As a multi-purpose method, SUBSTRA, provides reasonable predictive performance while producing more relevant descriptive outputs (as described later), thus maintaining a good trade-off between accuracy and interpretability that is lacking in most of the existing methods.

Discretization has positive effect for some datasets and methods and negative effects for the others. However, there is a general indifference with respect to the discretization techniques. The exception here is 'Multiple Myeloma', for which EFB resulted in better performance than the other techniques, matching the findings in [62].

To evaluate the plausibility of the weights assigned to the transcripts, we compared SUBSTRA with DCH using the following analysis. We ran both methods using the preprocessed data corresponding to the best predictive performance (among EFB, EWB, and KM) in figure 4.4. Experimental settings are described in supplementary section C. Then, the transcripts were sorted in descending order with respect to the weights obtained by each method. Top 100 transcripts were selected for each dataset and each method. We mapped transcripts to genes, and conducted Gene Ontology (GO) enrichment analysis for the top 100 genes for each dataset. The only exception was the 'Kidney 1' dataset for which we

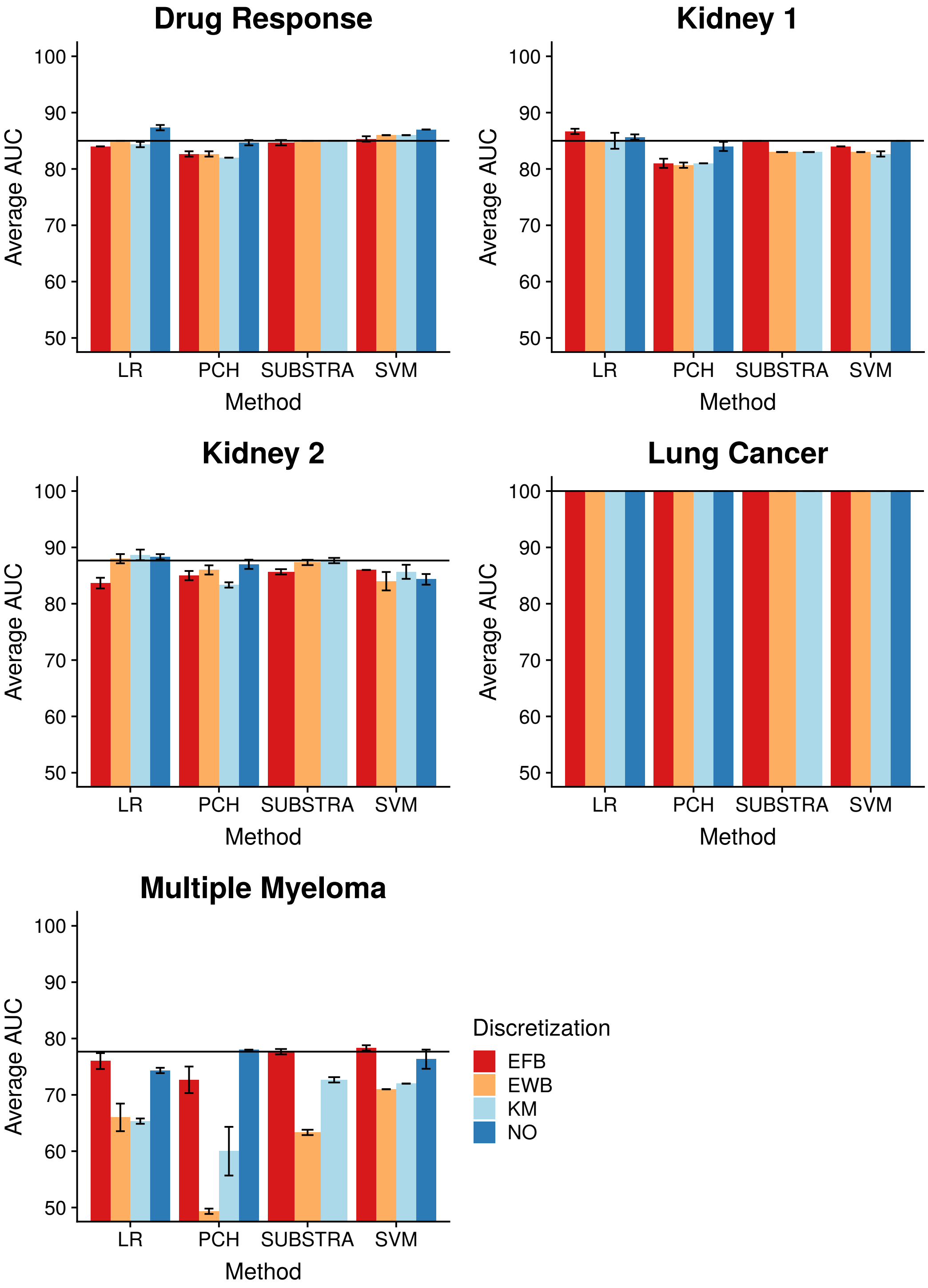


Figure 4.4: Predictive results for the real data. The horizontal line indicates the best performance of SUBSTRA. The error bars are based on standard deviation. NO – no discretization.

selected top 200 to obtain enriched GO terms for at least one of the methods (top 100 genes were not significantly associated to any GO term). To compare consistency of the top genes across the two methods, we computed the GO terms that are significantly enriched with top genes (*q-value* < 0.05 after false discovery rate correction using Benjamini-Hochberg method) for both methods (i.e., common enriched GO terms). Then, we compared the *q-values* associated to these GO terms by the two methods to see which method produces more significant enrichment for the common terms. We used paired Wilcoxon signed-rank test on the logarithms of the q-values. Next, we performed a similar analysis for the top transcript cluster of each method according to the average weight. For 'Drug Response', since the top 3 clusters of none of the methods were significantly associated with any GO terms, we looked at the 4th clusters.

The statistical significance of the difference between the enrichment of the top genes and clusters selected by the two methods are shown in Table 4.7. According to these results, top genes of SUBSTRA for 'Lung Cancer' and 'Kidney 2' result in significantly stronger enrichment. For 'Kidney 1', DCH top genes were not associated with any GO terms while SUBSTRA top genes were related to 15 significantly enriched GO terms indicative of higher consistency among them. For 'Multiple Myeloma' and 'Drug Response', there was no statistically significant difference the two methods. Overall, SUBSTRA detected significantly more relevant genes in 2 out of 5 experiments and was equally well in the others, which indicated its descriptive abilities compared to existing methods.

For the top transcript clusters, the results were more different among the two methods. In 4 out of 5 cases, no enrichment was detected for DCH while SUBSTRA could detect significantly enriched clusters. The reason might be the relatively small clusters that wNMF detected. For 'Kidney 1', both methods produced large top clusters, but SUBSTRA's cluster was very significantly more enriched. This indicates the meaningfulness of the transcript clusters detected by SUBSTRA. In the next section, we look at the relevance of these clusters to the phenotypes.

4.3.5 SUBSTRA Finds Relevant Transcript Clusters

SUBSTRA detects transcript clusters that define patient subtypes. Sorting clusters by the average of the transcript weights gives an indication of their relevance to the phenotype under consideration. We further analyzed the top 5 transcript clusters that SUBSTRA identified for each real dataset through Gene Ontology (GO) and Pathway (PW) enrichment analysis. The results indicate the uniform relevance of the identified transcript clusters and match the existing literature beside detecting novel signals requiring further investigation. After mapping the transcripts to the corresponding genes, Gene Ontology (GO) and Pathway (PW) enrichment analysis based on the "tmod" R package [139] was performed for these transcript sets. Biological Process (BP) GO terms and KEGG pathways from the Molecular Signatures Database (MSigDB) [79, 122] are used as the candidate gene modules,

	Metric	**Kidney 1**	**Kidney 2**	**Drug Response**	**Multiple Myeloma**	**Lung Cancer**
Genes	WSRT(Com.)	NA(0)	0.03(6)	0.65(9)	0.87(67)	0.01(36)
	SUBSTRA	**NA**	**-20.96**	-4.20	-6.34	-4.86
	DCH	NA	-5.33	-4.41	-6.03	-6.04
Cluster	WSRT(Com.)	0.00(69)	NA(0)	NA(0)	NA(0)	NA(0)
	SUBSTRA	**-31.53**	**NA**	**NA**	**NA**	**NA**
	DCH	-4.34	NA	NA	NA	NA

Table 4.7: Comparison between the weights assigned by SUBSTRA and DCH to the transcripts. Abbreviations used include WSRT(Com.) – Wilcoxon Signed-Rank Test (WSRT) *p-value* and the number of common GO terms in the parentheses. The top and bottom halves of the table correspond respectively to the evaluation of the top weighted genes and cluster. In each of the two parts, the second and third rows show the mean of the logarithm of the *q-values* of the enrichment tests for SUBSTRA and DCH, respectively. NAs indicate the situations when there have been no common enriched GO term between the two methods. In all NA cases, this was due to one of the methods (DCH) having empty enriched GO term set. The best performances are shown in bold.

and all MSigDB genes are used as the background gene set. Modules with *q-value* < 0.05 are selected as significantly enriched. As an example, figures 4.5, 4.6 and 4.7 show the heatmap and gene enrichment results for the 'Kidney 2' dataset. These will be explained later in the corresponding paragraph.

In the following paragraphs we provide the highlights of the descriptive results based on the gene clusters identified in SUBSTRA's outputs. The 'Kidney 1' dataset was obtained from biopsies extracted more than a year after the kidney transplants [40]. The authors of this study developed a classifier for transplant failure versus acceptance, and identified 886 genes whose expression was significantly associated with graft failure. Of the 30 top genes most frequently used by the classifier, five (HAVCR1, ITGB3, LTF, PLK2 and SERPINA3) were clustered in the second top cluster (C2) identified by SUBSTRA. SUBSTRA clusters suggests that inflammatory processes (cluster C1) can be implicated separately from pathways associated with cellular death and differentiation, extra-cellular matrix organization and circulatory system development (cluster C2), in allograft rejection. In fact Einecke et al. [40] implicate inflammatory processes in early graft rejection, and pathways enriched in SUBSTRA cluster C2 in later graft loss, suggesting that SUBSTRA correctly captures and distinguished among different mechanisms responsible for rejection (see figures A.4 and A.5). Although genes in C3, a cluster enriched in transmembrane transport, and C5, a cluster enriched in organ morphogenesis and tissue development, are present among the 886 classifying genes in the original publication, SUBSTRA makes a novel prediction that these additional mechanisms play distinct and central roles in graft rejection.

In the study associated with the 'Kidney 2' dataset, Khatri et al. [67] identified a 'common rejection module' consisting of 11 genes that were differentially expressed in rejection of

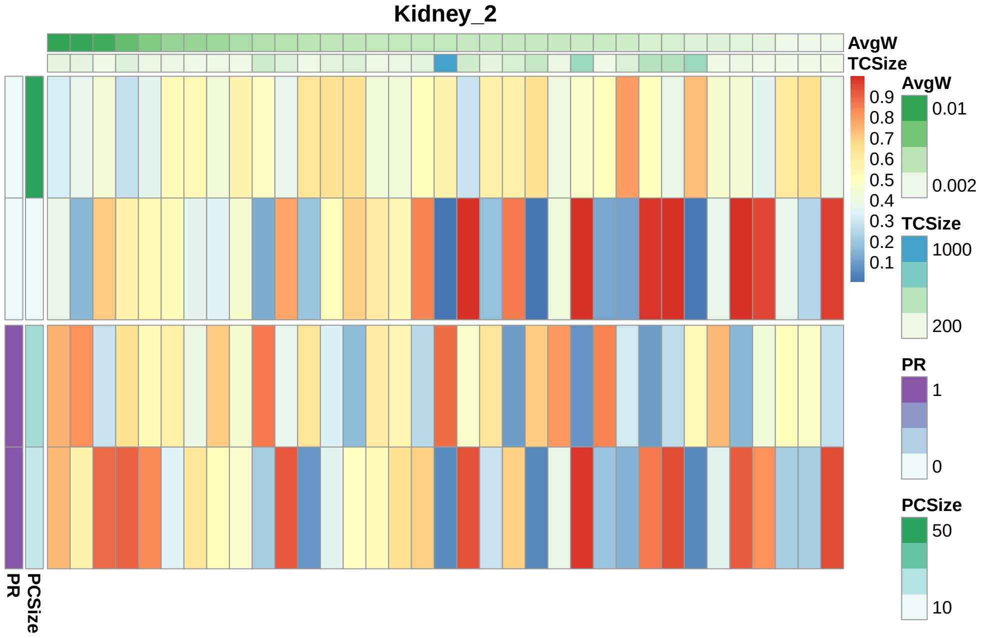


Figure 4.5: Heatmap for Kidney 2 Dataset.

transplanted organs : BASP1, CD6, CD7, CXCL9, CXCL10, INPP5D, ISG20, LCK, NKG7, PSMB9, RUNX3 and TAP1. SUBSTRA placed six of these genes – CXCL9, CXCL10, LCK, NKG7, PSMB9 and RUNX3, in the fourth gene cluster, supporting the conclusions of Khatri *et al.*, that these genes form a distinct module that differentiates graft rejection from non-rejection. The second top cluster shows enrichment of 'graft versus host disease', allograft rejection, immune signaling pathways, as well as related pathways such as cell, leukocyte, and lymphocyte activation (see figures 4.6 and 4.7). The first two pathways are active in almost half of the rejection cases (see figure 4.5). The rest of the rejection cases are associated with C3 to C5, which exhibit related but slightly different enrichment of immune response pathways.

'Drug Response' dataset [10] contains gene expression information from cancer cell lines treated with AZD6244, known as selumetinib. Selumetinib's target, MEK, is implicated in the epithelial-mesenchymal transition (EMT), which is an important step in the initiation of metastasis [11]. Among many other physiological changes, EMT involves the loss of cell-cell junctions such as tight junctions that are characteristic of epithelial cells. Our method identifies a transcript cluster related to EMT involved in cell-substrate adhesion as key pathways that respond to selumetinib (see figure A.7).

In 'Multiple Myeloma', Tian et al. [130] identified DKK1 as an important gene involved in the formation of focal bone lesions. As an inhibitor of the Wnt signaling pathway, DKK1's exact role in modulating this phenotype can be related to any of the pathway's many downstream effects, such as cell fate determination, cell motility, body axis formation, cell proliferation and stem cell renewal [70]. SUBSTRA recapitulated the original analysis

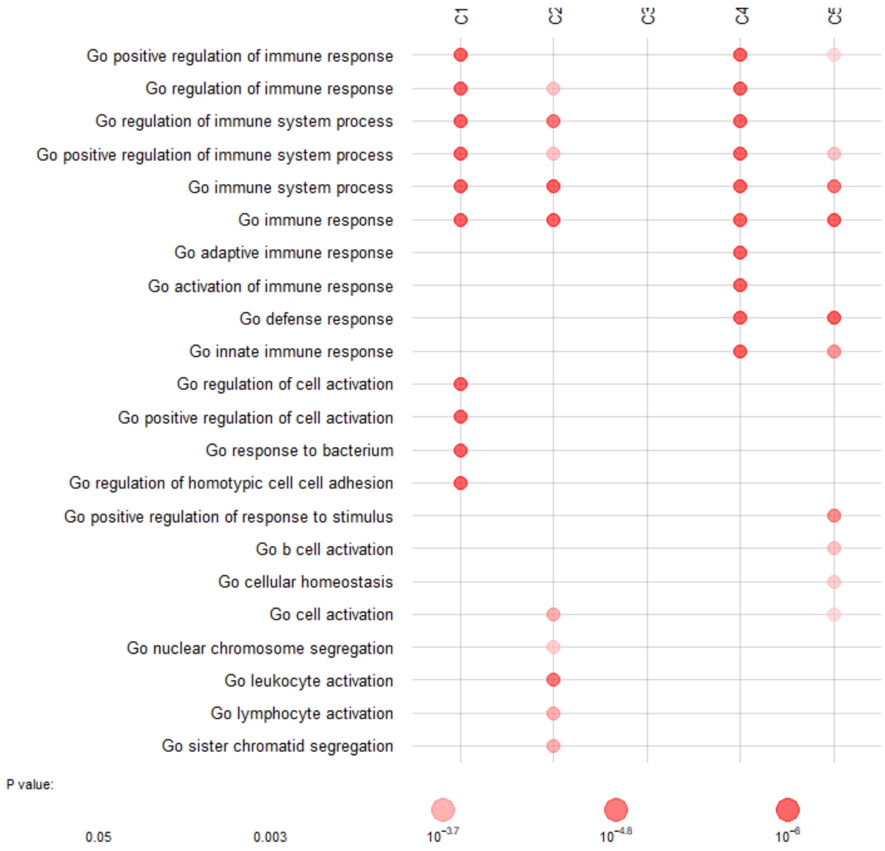


Figure 4.6: GO Enrichment for Kidney 2 Dataset.

by assigning the greatest weight to DKK1 within the third relevant cluster C3. Interestingly, this cluster also harbors some of the most significantly enriched pathways. Gene set enrichment analysis identified the cell cycle and MAPK, signaling as pathways enriched in genes of this cluster (C3 in figures A.9 and A.10). This result suggests that DKK1 might be modulating cell proliferation as opposed to other cellular processes associated with the Wnt signaling pathway. Furthermore, previous work has shown an interplay between the Wnt and MAPK signaling pathways in skeletal development [147]. MAPK ,signaling may be playing an important role in the formation of osteolytic lesions, a potential discovery that is not described in the original study. This shows that SUBSTRA biclustering and weight assignment can complement other methods such as differential gene expression analysis to provide additional biological context.

For the 'Lung Cancer' dataset, Gordon et al. [49] originally identified eight genes differentially expressed between adenocarcinoma of the lung (ADCA) and malignant pleural mesothelioma (MPM): CALB2, ANXA8, EPCAM, CLDN7, NKX2-1, CD200, PTGIS, and COBLL1. SUBSTRA reported all but one gene (CLDN7) in the top 3 transcript clusters, although other claudin genes, namely CLDN3 and CLDN4, were included in the top cluster.

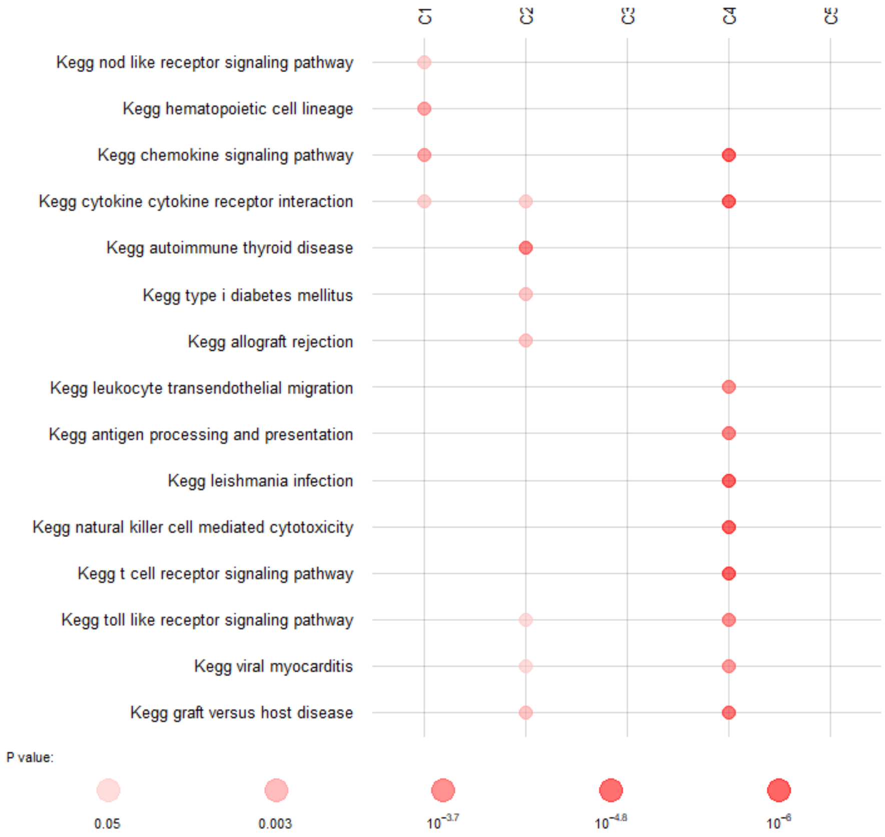


Figure 4.7: Pathway Enrichment for Kidney 2 Dataset.

Consistent with the eight genes, cell and focal adhesion are among the enriched GO terms and KEGG pathways in the top 5 transcript clusters (see figures A.12 and A.13). Moreover, SUBSTRA suggests several additional pathways, including extracellular receptor interaction, MAPK signaling, and cytokine receptor interactions, that may biologically distinguish ADCA and MPM.

4.3.6 Runtime of SUBSTRA

In a series of experiments on synthetic data, the influence of the input size factors on the runtime of SUBSTRA are identified. The studied factors include the number of patients m, the number of transcripts n, the number of patient strata, and the number of transcript clusters. When examining the effect of each of the four factors, the other three factors were kept constant. For each setting of the factors, first a corresponding synthetic dataset was generated. Next, 20 iterations of SUBSTRA consisting of 10 Phase I and 10 Phase II iterations were performed. The procedure was executed 10 times for each setting and the runtimes were averaged.

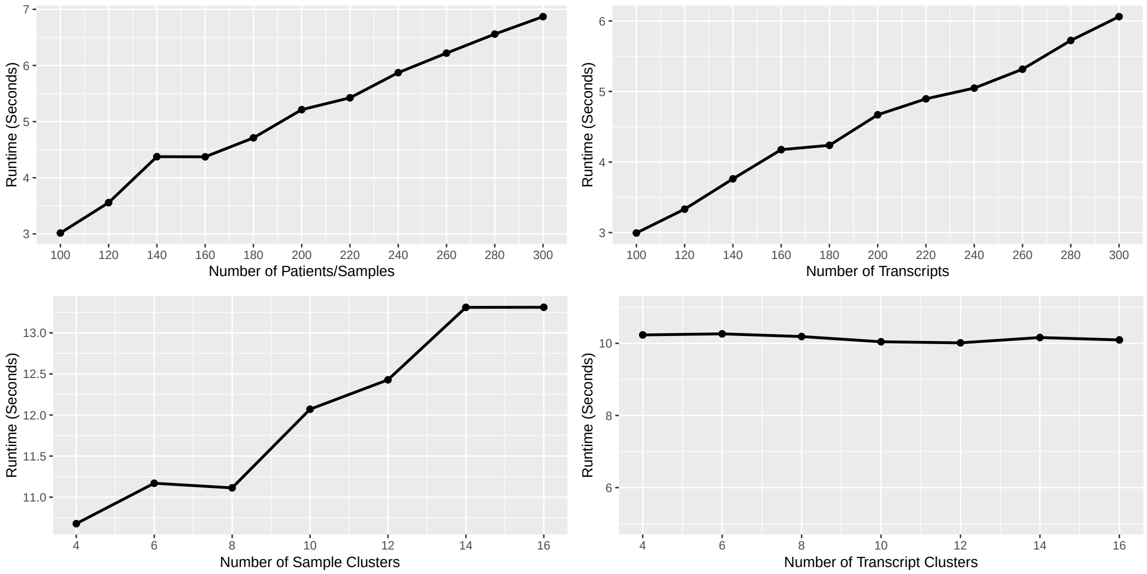


Figure 4.8: Results of the runtime analysis experiments. For the top left curve, the numbers of sample and transcript clusters were set to 4, the number of transcripts was fixed at 100, and the number of samples varied from 100 to 300. For the top right curve, the numbers of sample and transcript clusters were set to 4, the number of samples was fixed at 100, and the number of transcripts varied from 100 to 300. For the bottom left curve, the numbers of samples and transcripts were fixed at 240, the number of transcript clusters were set to 4, and the number of sample clusters varied between 4 and 16. For the bottom right curve, the numbers of samples and transcripts were fixed at 240, the number of sample clusters were set to 4, and the number of transcript clusters varied between 4 and 16.

Figure 4.8 shows the results. Based on these results, the runtime scales linearly with respect to the number of patients, the number of transcripts, and the number of patient clusters. However, the number of transcript clusters did not have any effects on the runtime in our experiments. This can be explained by the fact that the most expensive computations of SUBSTRA are the steps for learning the feature weights, which include only the other three factors.

Chapter 5

Collaborative Intra-Tumor Heterogeneity Detection

Despite the remarkable advances in sequencing and computational techniques, noise in the data and complexity of the underlying biological mechanisms render deconvolution of the phylogenetic relationships between cancer mutations difficult. As discussed in section 2.3, many of the existing methods for studying tumor evolution operate on tumor data from a single cancer patient. These methods have limited applicability for the majority of the existing data, which is in the form of a single sample low-to-medium coverage bulk sequencing dataset. Since this type of data contains numerous ambiguous cases implying more than one possible phylogenetic tree for the tumor, the existing algorithms for ITH detection based on a single tumor sample will yield several possible solutions for those cases [87].

Inter-tumor heterogeneity is another phenomenon increasing the complexity of understanding and treatment of cancer. However, despite that, tumors still might share evolutionary patterns among the same set of mutations [23, 99]. Therefore, these evolutionary similarities can be leveraged to create a collaboration between the information to guide the inference and reduce the ambiguities, especially for cases when the input is low to medium coverage bulk sequencing data from a single tumor sample.

Most of the existing methods that look at the above-mentioned similarities have two limitations: 1) they are based on binary mutation data and do not fully utilize the potential of sequencing data by overlooking the intrinsic information about the timing of evolutionary events, and 2) they infer phylogeny at the population or sub-population levels resulting in general instead of personalized evolutionary knowledge.

The above issues are partially addressed by a recent method, REVOLVER [23], which uses non-binary sequencing data and exploits the repeating evolutionary patterns for ITH detection in individual tumors instead of general evolution inference. REVOLVER assumes that a particular mutation usually has the same predictor (preceding mutation) across different tumors in a particular cancer type. Accordingly, the authors consider the frequency of the direct ancestors of a mutation across different tumors and use that information when inferring the phylogeny for a specific tumor. REVOLVER uses an Expectation-Maximization (EM) approach for finding the optimum phylogenetic trees. In the first step, an existing method (e.g., ClonEvol [31]) is used for deriving a set of high-scoring candidate phylogenetic trees for each tumor, the best of which is chosen as the current tree for each tumor. Then, the frequencies of the direct ancestors of each mutation are learned from the currently selected trees for all tumors. This information, which constitutes the parameters of the distribution over tree topologies, is then used for reevaluating the tree set for each tumor and selecting the ones with the highest new scores. These two steps of updating the parameters/frequencies (E-step) and updating the current trees based on the new parameters (M-step) continues until convergence or until termination criterion is met.

This approach decreases the uncertainty of phylogenetic structures by incorporating the ancestry information. However, the underlying evolutionary assumption in REVOLVER, which is the dependency of a mutation only on its direct ancestor (the preceding mutation), is a limitation because earlier mutations inherited by a subclone might also be decisive in the selection of the next mutation during the cancer evolution. Therefore, considering only the direct ancestor as the predictor of a mutation might result in a loss of information. Another issue, which is discussed further in section 5.2.2, is that the tree topology distribution used in REVOLVER is biased towards more branching structures. If not controlled, this bias may produce unrealistic results with too much branching.

In this chapter, we discuss the consequences of the above key issues and introduce a collaborative ITH detection method to address them. Our method, HINTRA, integrates sequencing data for a cohort of tumors and infers tumor phylogeny for each individual based on the evolutionary information shared between different tumors. Through a Bayesian iterative process, HINTRA learns the repeating evolutionary patterns and uses this information for resolving the phylogenetic ambiguities of individual tumors.

Our contributions can be summarized as follows:

- We introduce a Probabilistic Graphical Model (PGM) called HINTRA for collaborative ITH detection, as well as a corresponding parameter learning method. The proposed PGM is based on read count data, instead of summary values such as Cancer Cell Fraction (CCF) or Variant Allele Frequency (VAF), to account for the uncertainty of the measurements. To reduce the bias of existing methods, we propose a Bayesian EM method that leverages the topology uncertainty when learning the parameters, using a distribution over possible phylogenetic tree topologies instead of a point estimate.

- HINTRA includes a novel factorization approach for phylogenetic tree topologies. Addressing the information loss issue mentioned earlier, HINTRA considers all the mutations preceding a particular mutation in the phylogenetic tree, instead of only the most recent one. Moreover, the proposed factorization allows for the prediction of the next mutation that might happen in a subclone given its current mutational landscape. This capability, which is lacking in the existing methods, can be used for prognostic clinical applications.

Using both synthetic and real data, we evaluate performance of HINTRA and compare it to the state-of-the-art methods including REVOLVER (as a collaborative ITH detection method) and ClonEvol [31] (as a standalone ITH detection method). Our results for synthetic data based on different scenarios indicate that HINTRA outperforms the existing methods. Our results for real data were biologically consistent and provided new information of potential clinical interest. The C++ source code for HINTRA is available at https://github.com/sahandk/HINTRA.

5.1 Problem Definition

We now formally define the collaborative ITH detection problem. We assume that the input consists of read count data across m tumors. For each tumor, we consider read counts for a given set G of n known driver genes. The input data is organized into two matrices, one for the reference read counts denoted by $R = [r_{ij}] \in \mathbb{N}_0^{m \times n}$ and the other for the variant read counts denoted by $V = [v_{ij}] \in \mathbb{N}_0^{m \times n}$, where $\mathbb{N}_0$ denotes the set of non-negative integers. More precisely, r_{ij} and v_{ij} respectively denote the number of reference and variant reads supporting driver gene j in tumor i.

The output is a set of phylogenetic trees $\{T_i\}_{1 \leq i \leq m}$, where T_i is the phylogenetic tree of tumor i indicating the phylogenetic order of mutations in that tumor. A phylogenetic tree is a representation of the evolutionary events that are observed in a tumor. The root of the tree corresponds to the germline (GL) cell and the other nodes indicate the subclones of the tumor. Each edge stands for a mutation that occurs in a cell of the subclone corresponding to the edge's tail (the parent) and triggers the growth of the subclone corresponding to the edge's head (the child). In this work, we assume that the mutations satisfy the infinite sites assumption (ISA). This assumption means that each mutation appears exactly once in the phylogenetic tree at the node in which the mutation appears for the first time, and is present (conserved) in all the descendants of the subclone in which it first occurs. See Figure 5.3 for an example of a phylogenetic tree. Our goal is to infer the tree for each tumor by considering the evolutionary patterns of similar mutations in the other tumors' trees.

As a byproduct, we also learn model parameters that can be used to compute the probability that a particular mutation occurs in a cell having a specific set of mutations. For example, the parameters can contain information on the most frequent mutation occurring in (i.e. providing competitive advantage to) a breast cancer cell already containing the mutations *TP53* and *PIK3CA*. This parameter provides predictive information with prognostic applications.

Although it is theoretically possible to consider the exact position where each of the input mutations occurs within its gene, we chose to analyze the data at the gene level to increase the frequency of each mutation and gain statistical power. For genes affected by multiple mutations in the same tumor, we use the read count data for the most prevalent of such mutations, i.e. the mutation with the largest Cancer Cell Fraction (CCF). The CCF represents the fraction of cells of a tumor that harbor a particular mutation and most of the existing methods preprocess the read counts from sequencing data into CCFs before using them for phylogeny inference. This allows the use of the existing CCF computation tools (e.g. PyClone [110]) that can handle complicated cases such as mutations involving copy number variations (CNV). However, it ignores the uncertainty in the computed CCFs, which may lead to incorrect results by assigning high weights to uncertain inputs or vice versa. Incorporating read counts directly into the inference provides a more accurate representation and can help prioritize informative inputs over uncertain ones. Moreover, in cases with CNV, the computed CCFs can be simply translated into read counts based on an appropriate approximation of the locus coverage (e.g. mean sequencing coverage). Therefore, we choose read counts as the format of our input.

5.2 Methods

For the sake of simplicity, the methods in sections 5.1 to 5.2.3 are presented assuming that a single sample is available for each tumor. Later, in section 5.2.4, we generalize our model to allow multiple samples per tumor.

5.2.1 Probabilistic graphical model

The proposed PGM for HINTRA is shown in Figure 5.1. In this model, each tumor i, for $1 \leq i \leq m$, is associated with a phylogenetic tree T_i, whose structure depends on a parameter β. The tree structure constrains the possible values of the read count data. This is done through a latent variable $\theta_{i.}$, which is a vector of size n of Cancer Cell Fractions (CCFs) of driver mutations in tumor i. The dot in the index $i.$ denotes a vector. The CCF of a mutation indicates the proportion of cells in the tumor sample that harbour that mutation. A larger CCF is, in general, evidence of earlier occurrence of the mutation during tumor evolution. Accordingly, $\theta_{i.}$ depends on the tree structure T_i corresponding to tumor i and influences the noisy observed reference and variant read counts for tumor i.

According to the PGM, the joint probability of the model variables is factored as:

$$P(V, R, \theta, T, \beta) = P(V|R, \theta)P(\theta|T)P(T|\beta) \tag{5.1}$$

The first term on the right hand side of equation 5.1 is the likelihood term and is defined as below:

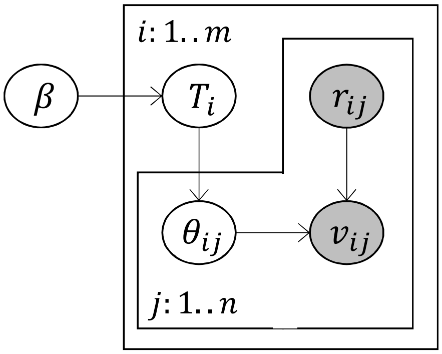


Figure 5.1: Probabilistic Graphical Model of HINTRA. Latent and observed variables are indicated by white and shaded circles, respectively.

$$
\begin{aligned}
P(V|R,\theta) &= \prod_{i=1}^{m}\prod_{j=1}^{n} P(v_{ij}|r_{ij},\theta_{ij}) \\
v_{ij}|r_{ij},\theta_{ij} &\sim \text{Binomial}(v_{ij}+r_{ij},\theta_{ij}/2)
\end{aligned} \tag{5.2}
$$

The Binomial distribution parameter is equal to $\theta_{ij}/2$ because CCF is computed as $\theta_{ij} = \frac{2v_{ij}}{v_{ij}+r_{ij}}$ for driver mutation j of tumor i (note the multiplication by 2 in the nominator). The second factor in RHS of equation 5.1 is defined as below:

$$
\begin{aligned}
P(\theta|T) &= \prod_{i=1}^{m} P(\theta_{i.}|T_i) \\
\theta_{i.}|T_i &\sim \text{Uniform(possible values)}
\end{aligned} \tag{5.3}
$$

The possible values for vector $\theta_{i.}$ are restricted by: 1) the *sum rule* indicating that the CCF for a mutation should not be smaller than the sum of the CCFs of all of its children in the phylogenetic tree T_i [61], and 2) $0 \leq \theta_{ij} \leq 1$ for $1 \leq j \leq m$.

The third factor and its computation is discussed in section 5.2.2.

5.2.2 Prior probability of phylogenetic trees

The underlying assumption of collaborative ITH detection is that some of the evolutionary patterns (i.e., phylogenetic relationships between the evolutionary events in a tumor) are common among different tumors. Accordingly, the goal is to define the entire phylogenetic tree in terms of its substructures representing the evolutionary patterns. One can then investigate the frequency of the patterns to find the more frequent patterns and use them as a reference whenever there is ambiguity for a tumor with respect to the phylogenetic relationships between the events involved in those frequent patterns. Here, ambiguous case refers to the case where multiple phylogenetic trees are consistent with the observed bulk data read counts. For a simple example of an ambiguous case we can consider a tumor with CCF values $[0.2, 0.3, 1.0]$. In this case, relying solely on CCF values, one can easily observe that both the chain and the branching topology are possible explanations of the observed data. Several more complicated examples for this were recently provided in the analysis of acute lymphoblastic leukemia patients in [87]. For an example of a non-ambiguous case we can consider a tumor with mutations having CCFs $[0.5, 0.8, 1.0]$. In this case, only the chain topology is consistent with the observed CCFs. Namely, for the branching topology, the frequencies of the two child nodes would add up to 1.3, which is larger than the CCF of their parent, thus violating the sum rule.

To the best of our knowledge, the most recent ITH detection method that is based on the assumption of common evolutionary patterns is REVOLVER [23]. REVOLVER assumes independence of the edges and defines the probability of the phylogenetic tree of tumor i as the product of the probabilities of the observed edges (i.e., the probability of attaching a given child node to a particular parent node) as follows:

$$P(T_i|\beta) \propto \prod_{p \to c \, \in \, E_i} P(p|c, \beta), \tag{5.4}$$

where p and c are the parent and child nodes of a given edge $p \to c$ of tree T_i and E_i is the set of all edges of the tree for tumor i. The parameter β governs the edge probabilities and is shared across all tumors.

In the above approach, each node is assumed to be dependent only on its direct ancestor. However, the selection of the next mutation that brings competitive advantage to a cell does not only depend on the last mutation, but it depends on the entire current mutational burden of the cell. Figure 5.2A shows a scenario in which the above assumption is violated, leading to a poor performance for REVOLVER (see section 5.3.1). In this scenario, the two trees have different truncal mutations (**a** for topology 1 and **e** for topology 2). Because of this difference, mutation d is attached to different parents in the two trees. However, considering only pair-wise relationships, because **d** happens after **b** in 70% of the tumors, the factors of REVOLVER will assign it under **b** even when inferring trees for a tumor having true topology 2.

Another drawback of the above factorization is that it cannot be translated into a prognostic application of predicting the next driver mutation based on the current mutational landscape of a subclone. The reason is that the conditional probability of the next mutation given all current mutations is not computable as a function of the parameters learned based on the above assumptions. In other words, based on the above assumptions, the next mutation depends only on the most recent ancestor.

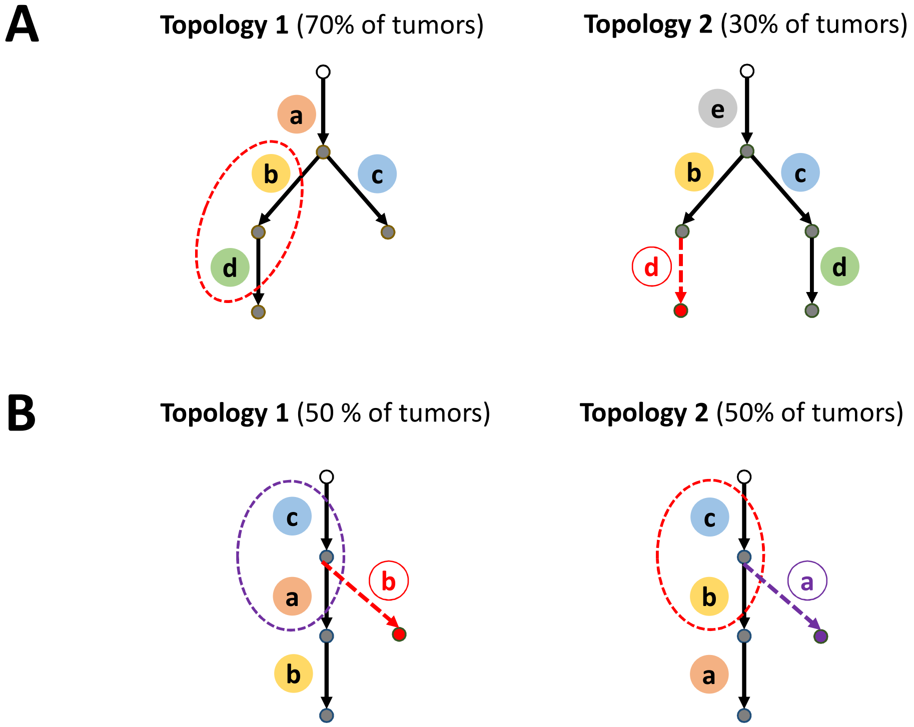
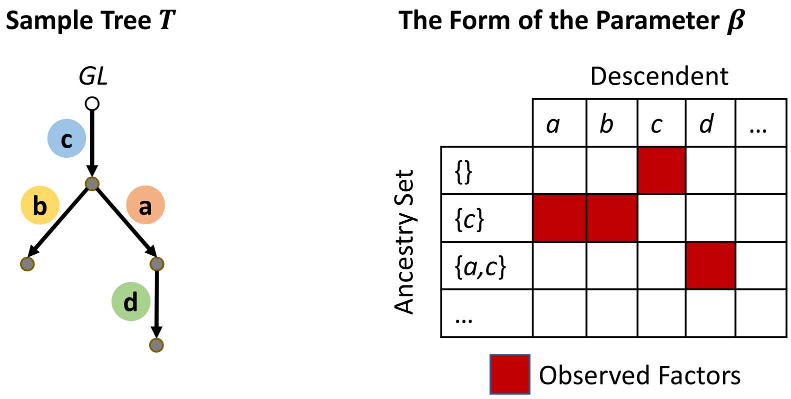

A
Topology 1 (70% of tumors)
a
b
c
d
Topology 2 (30% of tumors)
e
b
c
d
d
B
Topology 1 (50 % of tumors)
c
a
b
b
Topology 2 (50% of tumors)
c
b
a
a

Figure 5.2: Two sample scenarios in which tree factorization and parameter learning as in [23] results in undesired inference. The small circles denote the tumor subclones and the empty circle is the germline cell. The edges are labeled with the mutations, denoted by letters within larger circles. The true tree topologies are shown with solid edges. Each ambiguous situation is shown in a different color, with dashed ovals indicating the conflicting evidence (source of ambiguity) and the dashed edge indicating the possible mistake due to that evidence.



	a	b	c	d	…
{}			🟥		
{c}	🟥	🟥			
{a,c}				🟥	
…					

🟥 Observed Factors

Factorization of Tree T

$f(T) = \{\{\} \rightarrow c, \{c\} \rightarrow b, \{c\} \rightarrow a, \{a,c\} \rightarrow d\}$

$P(T|\beta) \propto \beta_{\{\},c}\, \beta_{\{c\},b}\, \beta_{\{c\},a} \beta_{\{a,c\},d}$

Figure 5.3: A sample phylogenetic tree and its factorization

To overcome the above limitations, we extend the tree factorization to capture the effect of all existing driver mutations (ancestors) on the next driver mutation (descendant). The occurrence order of the ancestors is not taken into account because the selection of the next mutation depends only on the set of current mutations, but not the order of these mutations. We define the prior tree probability as below:

$$P(T_i|\beta) \propto \prod_{\mathcal{P}\rightarrow c \;\in\; f(T_i)} \beta_{\mathcal{P},c}, \tag{5.5}$$

where $\mathcal{P}$ is the set of possible ancestor mutations of c in tree T_i, which we call the *ancestry set* hereafter. An ancestry set consists of all the mutations on the path in T_i from the root to any internal (i.e., non-leaf) node/subclone and captures the mutational landscape of that node/subclone. The function $f(T_i)$ returns the set of edges of T_i consisting of the ancestry sets and their children. The parameter β is a matrix with rows corresponding to all possible ancestry sets for all tumors in the dataset. The columns correspond to the set G of all mutations. The entry β_{kj} of the β matrix indicates the amount of evidence for an edge labelled with the j-th mutation whose tail is a node with the k-th ancestry set. Figure 5.3 illustrates these concepts.

5.2.3 Parameter learning

Although the proposed prior probability in section 5.2.2 resolves the information loss issue, it inherits the bias towards branching structures. Scenario B in Figure 5.2 shows a sample situation where this bias can lead to unexpected phylogeny detection. In topology 1 in scenario B, mutation **b** occurs after **a**, which is not consistent with topology 2, in which **b** occurs after **c**. A similar inconsistency exists between the ancestors of mutation **a** in the two topologies. Accordingly, based on both the REVOLVER and HINTRA factorization approaches, any ambiguous case that suggests a branching topology in which **a** and **b** can occur in parallel has supporting evidence due to the conflicting orders of **a** and **b** in the two topologies, even if it is originally associated with one of the two topologies. However, in case of a slight ambiguity (e.g. a 5% ambiguous cases for each of the two topologies), the evidence for the branching topology is very small and the chain topology should be favored (which has support from, for example, 95% of the samples). So, if the inherent bias towards branching topologies in the factorization approaches is not controlled, the methods infer the incorrect branching structure (shown with the dashed edges) for the ambiguous cases. We control this bias by employing a Bayesian EM parameter learning method described next. This method accounts for uncertainty of each of the possible topologies when learning the parameters and, in this scenario, only accepts a branching topology in cases with high certainty (i.e., when the subclones corresponding to **a** and **b** are very small).

We propose a Bayesian EM approach to learn the parameters of the PGM of HINTRA. The goal is to optimize the value of β, the topology distribution parameter, by maximizing the marginal likelihood $P(V|R,\beta)$ and utilizing the data's uncertainty. This is performed using an iterative approach with the following steps at each iteration:

1) Compute $\beta^{'}$ using $P(T|\beta,V,R)$ (see equation 5.6).

2) If $P(V|R,\beta) \leq P(V|R,\beta^{'})$ (see equation 5.10), then set $\beta = \beta^{'}$ and continue; otherwise output $\beta^{'}$ and terminate.

Initially, the tree priors are assumed to be uniform. Then, in the first step, $\beta_{\mathcal{P},c}$ is updated for each ancestry set $\mathcal{P}$ and descendant mutation c using the following equation:

$$\beta^{'}_{\mathcal{P},c} = \sum_{i=1}^{m} \sum_{T_i} \mathbf{1}_{f(T_i)}(\mathcal{P} \to c) \times P(T_i|\beta, v_{i.}, r_{i.}) + \epsilon, \tag{5.6}$$

where $\mathbf{1}_{A(x)}$ is the indicator function for $x \in A$ and the value ϵ is the pseudo-count for avoiding zero probabilities. Equation (5.6) is the sum of evidence for factor $\mathcal{P} \to c$ over all tumors, where the evidence is weighted by the posterior likelihood of every possible tree topology that contains the factor $\mathcal{P} \to c$. Accordingly, $\beta^{'}_{\mathcal{P},c}$ indicates the updated evidence for the factor $\mathcal{P} \to c$. The posterior likelihood for tree topology is computed as:

$$P(T_i|\beta, v_{i.}, r_{i.}) = \frac{P(v_{i.}|r_{i.}, T_i)P(T_i|\beta)}{\sum_X P(v_{i.}|r_{i.}, X)P(X|\beta)} \tag{5.7}$$

In the above equation, the marginal data likelihood is computed as below:

$$P(v_{i.}|r_{i.}, T_i) = \int_{\theta_{i.}} \prod_{j=1}^{n} P(v_{ij}|r_{ij}, \theta_{ij}) P(\theta_{ij}|T_i) \, \mathrm{d}\theta_{i.} \tag{5.8}$$

Because the term containing the integral over the vector $\theta_{i.}$ is not in closed form, we approximate that term using discrete values as below:

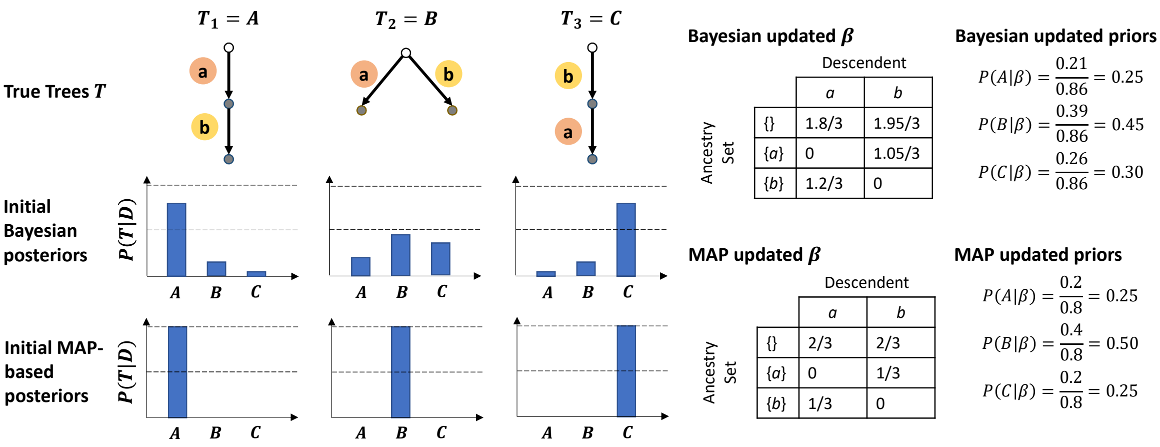
$$P(v_{i.}|r_{i.}, T_i) \approx \sum_{\theta_{i.} \in \delta_{\Delta}(T_i)} \prod_{j=1}^{n} P(v_{ij}|r_{ij}, \theta_{ij}) P(\theta_{ij}|T_i), \tag{5.9}$$

where $\delta_{\Delta}(T_i)$ is a function that enumerates all discrete values of the vector $\theta_{i.}$ with step-size Δ considering the constraints imposed by topology T_i. In our experiments (section 5.3), we use $\Delta = 0.05$.

In the second step, the marginal probability conditional on β (i.e., the maximization objective) is computed as:

$$P(V|R, \beta) = \prod_{i=1}^{m} \sum_{T_i} P(v_{i.}|r_{i.}, T_i) P(T_i|\beta) \tag{5.10}$$

Figure 5.4 illustrates, with an example, the Bayesian EM approach described above as well as the EM approach used in REVOLVER, which uses MAP point estimate. It explains how using a Bayesian approach that employs the entire spectrum of possible topologies (i.e. data uncertainty) instead of the point estimates as used in REVOLVER, can reduce the bias inherent in both of the tree prior probability definitions used in REVOLVER and HINTRA. The figure shows the first step of different EM approaches on a dataset with three tumors all having two driver mutations, a and b. Topologies A, B, and C constitute all possible trees with the two mutations. Each of the three tumors has a different topology, as shown. The bar charts show the initial posterior probabilities of the topologies (i.e., $P(T|D) \propto P(D|T)P(T|\beta)$) for the tumors computed assuming a uniform initial topology prior $P(T|\beta)$ and hypothetical data likelihoods. Two types of posteriors are computed, one based on the Bayesian estimation (the top row) and one based on the MAP estimation (the bottom row). The evidence β updated based on the two types of posteriors is shown in the middle. At the right, the updated priors based on the updated evidences are presented. Despite the fact that each of the three topologies is observed only once, the bias in the prior definition makes the most branching topology B more likely. However, the Bayesian approach considers the entire spectrum of possible topologies (i.e. based on the data likelihood given each possible topology), which reduces the bias. As shown in this figure, ambiguous cases like B often have a more uniform distribution over topologies than the other cases, resulting in reduced support for branching. This mitigates the effects of the prior bias during the learning process. Besides this, since we optimize the marginal data likelihood (equation 5.10) instead of the



Bayesian updated β

Ancestry Set	Descendent a	Descendent b
{}	1.8/3	1.95/3
{a}	0	1.05/3
{b}	1.2/3	0

Bayesian updated priors

$$P(A|\beta) = \frac{0.21}{0.86} = 0.25$$
$$P(B|\beta) = \frac{0.39}{0.86} = 0.45$$
$$P(C|\beta) = \frac{0.26}{0.86} = 0.30$$

MAP updated β

Ancestry Set	Descendent a	Descendent b
{}	2/3	2/3
{a}	0	1/3
{b}	1/3	0

MAP updated priors

$$P(A|\beta) = \frac{0.2}{0.8} = 0.25$$
$$P(B|\beta) = \frac{0.4}{0.8} = 0.50$$
$$P(C|\beta) = \frac{0.2}{0.8} = 0.25$$

Figure 5.4: Bias in the topology prior probabilities and how the Bayesian approach mitigates this bias

maximum likelihood, we adapt the parameter β to the whole information in the observed data as opposed to fitting it to the point estimates. This results in further reduction in bias.

To reconstruct the phylogenetic tree structures, we use MAP estimation after the parameter β has been computed using the above Bayesian EM approach. For each tumor i, we have:

$$T_i = \arg\max_{T_i} \{ \ P(v_{i.}|r_{i.}, T_i) P(T_i|\beta) \ \} \tag{5.11}$$

5.2.4 Generalization to multiple samples per tumor

In the more general case where multiple samples (obtained, for example, by sequencing multiple regions of the tumor) are available for a given tumor, we define the likelihood of tumor data (previously shown in equation 5.9 for the single-sample case) as the product of the likelihoods of the individual samples:

$$\begin{aligned} P(v_{i.}|r_{i.}, T_i) &= \prod_{q=1}^{s_i} P(v_{i.}^q | r_{i.}^q, T_i) \\ &\approx \prod_{q=1}^{s_i} \sum_{\theta_{i.}^q \in \delta_\Delta(T_i)} \prod_{j=1}^{n} P(v_{ij}^q | r_{ij}^q, \theta_{ij}^q) P(\theta_{ij}^q | T_i), \end{aligned} \tag{5.12}$$

where s_i is the number of samples for tumor i and $v_{i.}^q$ and $r_{i.}^q$ are the read count data for sample q of tumor i and $\theta_{i.}^q$ is the corresponding parameter vector.

5.2.5 Extracting prognostic information

The likelihood that each mutation c follows an ancestry set $\mathcal{P}$ is computed as:

$$P(c|\mathcal{P}) = \frac{\beta_{\mathcal{P},c}}{\gamma_{\mathcal{P}}}, \tag{5.13}$$

where $\gamma_{\mathcal{P}}$ is the evidence for $\mathcal{P}$ computed as:

$$\gamma_{\mathcal{P}} = \sum_{i=1}^{m} \sum_{T_i} \mathbf{1}_{g(T_i)}(\mathcal{P}) \times P(T_i|\beta, v_{i.}, r_{i.}) + n\epsilon, \tag{5.14}$$

where the function $g(T_i)$ returns all the ancestry sets in tree T_i, ϵ is the pseudo-count (a small value) and n is the number of mutations in the input dataset.

Because $P(c|\mathcal{P})$ is a proportion estimate, the minimum value for the sample size $\gamma_{\mathcal{P}}$ to have a 95% confidence interval of width W can be computed as $4/W^2$ (e.g. $\gamma_{\mathcal{P}} \geq 100$ for $W = 0.2$).

5.3 Experimental Results

5.3.1 Experiments with synthetic data

We evaluated the performance of HINTRA using synthetic data to have access to the ground-truth phylogenetic trees. The comparison partners included REVOLVER [23], as the only method that explores a similar idea of collaborative ITH detection, and ClonEvol [31], as the state-of-the-art method for standalone ITH detection. We used the same evaluation metric as in [23], namely true positive ratio, which is the proportion of predicted edges that exist in the ground-truth tree.

For comparison with REVOLVER, we conducted three different experiments. The first experiment evaluated the information transfer and de-noising capabilities of HINTRA. In this experiment, we followed exactly the same simulation procedure used for evaluating REVOLVER. The second experiment showcased one of our main contributions, the ability of HINTRA to capture more complete evolutionary patterns. This experiment was based on scenario A in Figure 5.2. The third experiment examined the ability of HINTRA to control the topology distribution bias and it was based on scenario B shown in Figure 5.2.

As in [23], the sensitivity to CCF noise levels are monitored in the three experiments, where noise follows a Gaussian distribution and was controlled through tweaking the standard deviation (e.g. 0 or 0.05). Moreover, ambiguity was introduced into the ground-truth models as the percentage of tumors with CCFs that had different possible phylogenetic structures (i.e. ambiguous cases). These experiments were conducted assuming a single sample per tumor. To evaluate the effect of the number of samples on the methods' performance, we conducted an additional set of experiments where 2 or 4 samples were generated per tumor, and considering the most difficult simulation configuration, i.e. higher noise and ambiguity.

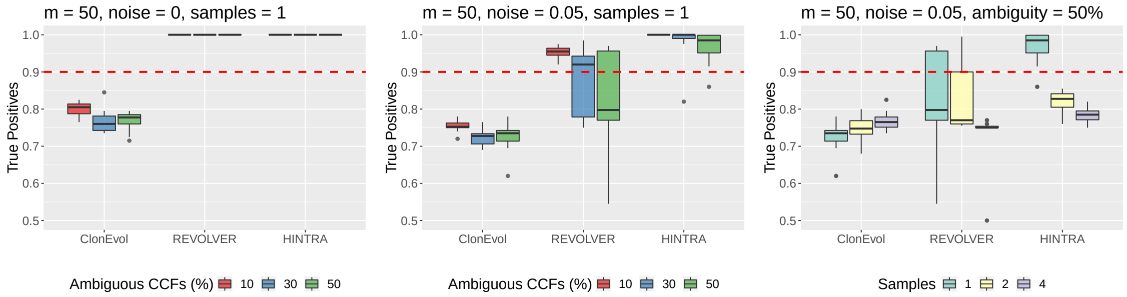


Figure 5.5: Results for the synthetic datasets from [23]

All samples of a tumor were assumed to be ambiguous for ambiguous cases and they were all non-ambiguous otherwise. For each configuration of the parameters, the experiment was repeated 10 times with *de novo* generation of the synthetic data at each repetition. For each single repetition, the average true positive ratio over all tumors was computed and plotted as a point. We simulate a sequencing coverage of 100x in all experiments.

The synthetic data in [23] was produced by assuming the same evolutionary tree (a chain structure) consisting of four mutations across all tumors. We repeated the same experiments for a cohort of 50 tumors. Two CCF noise levels of 0 and 0.05 and three different percentages of ambiguous cases (10%, 30%, and 50%) were simulated as in the original paper. The results are shown in Figure 5.5. According to Figure 5.5, unlike ClonEvol, both HINTRA and REVOLVER were able to detect the correct phylogenetic trees for all tumors and for all levels of ambiguity when there was no noise in the CCFs. However, after introducing noise with standard deviation 0.05 to the true CCFs, the level of ambiguity had a stronger effect on performance. HINTRA outperformed REVOLVER in all the datasets with noise, and the gap between the two methods increased with increasing level of ambiguity. This indicates the higher robustness of HINTRA to noise and ambiguity. Interestingly, by increasing the number of samples, the performance of the stand-alone algorithm improved but the collaborative methods exhibited decreasing accuracy. These results are consistent with the original study [23]. The most likely reason for this is the high level of ambiguity (50%). For ambiguous cases, multiple noisy samples create conflicting phylogenies which leads to a higher probability for branching structures. Thus, transferring information from the ambiguous cases decreases the overall accuracy. The stand-alone method is less sensitive because each tumor is analyzed separately. We note that it is very unlikely that in real data all samples of a tumor with a ground-truth chain structure are ambiguous as we assumed here. HINTRA, in general, performs slightly better than the two other methods for larger numbers of samples.

In the second experiment, we simulated scenario A shown in Figure 5.2 for a cohort of 50 tumors (35 with tree 1 and 15 with tree 2). Because the ground-truth topologies have branches and so are associated with ambiguous cases, we only investigated the noise level and

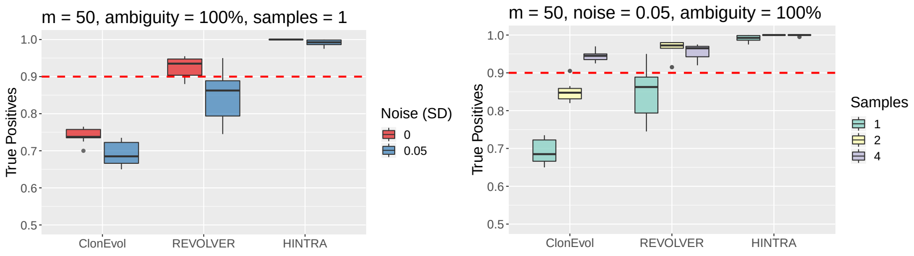


Figure 5.6: Results for the synthetic datasets based on scenario A from Figure 5.2

the number of samples as the variable factor. Two noise levels of 0 and 0.05 were simulated. The results are shown in Figure 5.6. We observe that there are considerable gaps between the performance of the three methods. As explained earlier, the gap between REVOLVER and HINTRA is due to the differences in the definitions of the tree topology factors, where HINTRA looks further back into the evolutionary history of a subclone and provides a more accurate assignment of the mutations based on that richer information. Unlike the previous scenario, having more samples improves the performance of the methods in this scenario. This is due to the fact that the trees are consistently branching topologies in this scenario. These topologies are associated with ambiguous cases and, unlike for chain topologies, having multiple ambiguous noisy samples is not misleading for these cases. Overall, HINTRA performs slightly better than the two other methods with a larger number of samples.

For the third experiment, we simulated scenario B as shown in Figure 5.2. Two levels of noise (0 and 0.05) were introduced to the CCFs and 6%, 10%, and 14% were used as the frequency of ambiguous cases. The goal of this experiment was to investigate the capability of the methods to control the bias towards branching structures. Accordingly, we set small ambiguity levels to leave enough evidence for the true structures and examined whether the methods could still infer branching structures in absence of direct evidence. The branching structure was still supported indirectly due to the two conflicting structures of tree 1 and tree 2, but there were not enough ambiguous cases to support that structure. Figure 5.7 shows the results. Because of the low levels of ambiguity, the true positive rates were in general high for all methods. However, CloneEvol and HINTRA performed better than REVOLVER in this experiment. CloneEvol performed standalone phylogeny inference and chose one of the two possible topologies (chain and branching) at random. However, REVOLVER had a bias towards branching structures and preferred that structure whenever it was possible. HINTRA had less bias in the topology distribution due to the Bayesian EM approach and opted for the branching structure only whenever it had high probability. Therefore, HINTRA controlled the bias effectively for all levels of ambiguity and noise in our experiments. When more samples were used per tumor with a noise of 0.05 and an ambiguity of 14%, all the methods showed improvement. Due to the small ambiguity, more samples improved the evidence for the correct topology and strengthened the information

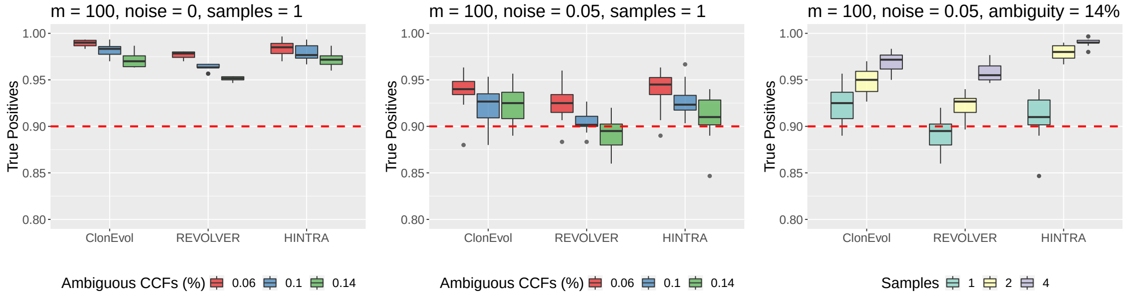


Figure 5.7: Results for the synthetic datasets based on scenario B from Figure 5.2

transferred from those cases, which resulted in a better resolution for the ambiguous cases. Overall, HINTRA once again outperformed the other methods when multiple samples were used.

5.3.2 Experiments with real data

In the absence of ground-truth phylogenetic trees for tumor mutations, performing an objective comparison of the accuracy of HINTRA and any other method is difficult. Instead, we evaluated HINTRA's performance based on the consistency of the learned parameters with existing biological domain knowledge. We chose Breast Cancer as the subject of study, since it is one of the most studied cancer types for which a rich body of domain knowledge is available. We used a public data set from [107], which includes 1756 advanced breast cancer patients. This dataset is the most recently published genomic dataset for Breast Cancer with clinical data.

In the available clinical data, these patients were stratified according to whether they do or do not express the genes for the receptors for the hormones estrogen and progesterone (HR) and HER2, resulting in the HR+/HER2-, HR+/HER2+, HR-/HER2+, and TN (Triple Negative) subtypes. We used this information to separate the patients into four corresponding groups and ran HINTRA for each group independently to infer tumour progression and phylogenetic trees. We only included tumors having Single Nucleotide Variations (SNVs) and a normal copy number in the considered loci. We considered mutations in breast cancer genes from COSMIC Cancer Gene Census dataset (cancer.sanger.ac.uk, [53]) and augmented the list by the genes mentioned in the original study [107]. After limiting to the selected genes and filtering out synonymous mutations, loss of heterozygosity, and weak signals (i.e. small read counts and mutations with less than 1% frequency in each subtype), the number of patients with at least one mutation was reduced to 1348. We also limited the number of mutated genes per tumor to 5 and removed the 47 tumors (3.5%) that did not satisfy this constraint.

Finally, the read count data was modified based on the available sample purity data. Sample purity is the proportion of cells within a biopsy sample that comes from tumor as opposed to normal cells captured in the sample. Modifying the data with respect to purity reduces the ambiguity by increasing the corresponding CCF values. For this correction, we reduce the number of reference read counts assuming that part of them are associated with the normal cells. The modified number of reference read counts can be computed as $r' = r - (r + v)(1 - z)$, where z is the tumor purity.

A large subset of the cohort were HR+/HER2- cases. In a majority of cases in this subtype, we observed clonal mutation acquisition in signaling cascades (TP53, PIK3CA, AKT, GATA3, PTEN, etc.) as discussed in the original findings [107], suggesting that HINTRA is able to reliably detect these early mutational associations. HINTRA also found that the most likely descendant of TP53 and PIK3CA combinatory events in HR+/HER2- subtype is PTEN, which occurs with probability 0.2. We consulted the literature and found some inconsistencies between studies with regarding the relationship of PTEN to PIK3CA. For example, some studies argue that PIK3CA is mutually exclusive to PTEN [121], while others state that PIK3CA could be characterized together with PTEN deletions for HR+ subtypes [96]. In the data set we used, the mutual exclusivity of PIK3CA and PTEN mutations was observed across the cohort. Our results suggest that TP53 may have some additive effects on PTEN and its association to PIK3CA. This may be a potentially interesting topic since TP53 and PTEN are both tumour suppressors and could provide a tumorigenic advantage to these aggressive subtypes. Consistent with the existing knowledge, HINTRA detected TP53 as the most important initiator preceding GATA3, CDH1, and FOXA1, which are commonly associated with invasive lobular carcinoma, a subtype within HR+/HER2-. A high proportion of HR+/HER2- cases acquire a CDH1 mutation, which is a hallmark of lobular carcinoma. Furthermore, it is known that CDH1 loss and PIK3CA gain of function are highly correlated with these outcomes; however, their order is not accounted for in the literature, and when these mutations are mentioned, they are characterized as a group [5]. Interestingly, we found that CDH1 is almost three times as likely to be the initiator of this association with PIK3CA, which may provide some insights on the development of lobular carcinoma.

The limited number of samples in the other three subtypes (HR-/HER2+, HR+/HER2+, and TN) resulted in weaker signals. Among the stronger patterns derived from the parameters learned by HINTRA, we observed that TP53 is almost twice as likely to be an initiator driver mutation when associated with PIK3CA in HR-/HER2+ and TN tumors. This adds to the results of PCAWG studies such as Gerstung et al. [45] demonstrating that driver mutations in PIK3CA and TP53 are more likely to be clonal.

5.3.3 Computational resources analysis

The size of the input for collaborative intra-tumor heterogeneity detection can be defined in terms of the CCF discretization hyper-parameter Δ, number of tumors m, the number of mutations per tumor and the number of unique mutation profiles referred to as "com-

Factor	Values
Samples (m)	20, $\underline{40}$, 60
Mutations per sample	3, $\underline{4}$, 5
Combinations	1, $\underline{5}$, 10
Δ	0.025, $\underline{0.050}$, 0.100
CPU Cores	2, $\underline{4}$, 6

Table 5.1: Problem size factors considered in the running time analysis. Default values are underlined.

binations". Here, "profile" stands for the set of observed mutations for the corresponding tumor. In addition to these factors, the number of utilized CPU cores can affect the time and memory resources consumed by the different methods.

To evaluate the effects of these factors, a set of experiments with synthetic data was conducted. The range of values tested and the default values for each of the factors are provided in Table 5.1. When studying the effect of each factor by changing its value, all other factors were set to their default values. The datasets were generated following the approach in [23]. Different combinations of mutations were generated based on the assumption that half of the mutations of a new combination should already exist in the previous combinations (each mutation can belong to a different existing combination) and half of them should be new mutations not existing in any of the previous combinations. This mimics the real data distribution in the way that it assigns higher probability of mutation to a few genes and promotes heterogeneity.

Both HINTRA and REVOLVER consist of two phases: preprocessing and EM. During phase I (preprocessing), REVOLVER uses ClonEvol to construct and score the trees for each tumor and selects the top trees as candidates. HINTRA computes the marginal likelihood using equation 5.9. The results of this phase could be stored in both algorithms to avoid recomputation costs. During phase II (EM), both algorithms learn the parameters. The maximum number of EM iterations was set to 100 for both HINTRA and REVOLVER in these experiments. We measured the running times for the two phases separately for better interpretation. The results are shown in Figure 5.8. According to these results, REVOLVER performed the first phase more efficiently and was less sensitive to the problem size. This is due to the efficient strategy used in ClonEvol for searching the tree topology space, whereby the search space is pruned based on the consistency of the subtrees with the CCFs, resulting in a considerably smaller search space. On the other hand, the current implementation of HINTRA enumerates all possible topologies, whose number is combinatorially related to the number of mutations. Furthermore, unlike HINTRA, ClonEvol requires the clonal mutation to be identified in the input and it builds the tree of the rest of the mutations under that clonal mutation. This has the significant effect of reducing the search space by fixing one node.

Another important factor affecting the running time of HINTRA is Δ. The running time of HINTRA contains a term proportional to $\binom{\Delta^{-1}+x}{x}$, where x is the number of mutations in a sample. Our experiments (results not shown) indicate that there was no noticeable difference between the accuracy of HINTRA when $\Delta = 0.05$ or $\Delta = 0.1$ and the latter can be used to improve the speed without sacrificing the accuracy. Yet, using $\Delta = 0.05$, our experiment with breast cancer HR+/HER2- subtype (see section 5.3.2), consisting of 1019 samples with up to 5 mutations, took about 50 minutes. The running time of both methods scales linearly with the number of tumors. In contrast to REVOLVER, which does not allow parallel processing in phase I, using more CPU cores improves the running time performance of HINTRA (see Figure 5.8).

In phase II, HINTRA was in general more efficient than REVOLVER. However, it was more sensitive to the number of mutations. This can be explained by the fact that HINTRA integrates over all tree topologies while REVOLVER focuses only on a set of top trees selected by ClonEvol, the size of which is bounded independently of the number of mutations.

The running time of HINTRA in both phase I and II can be improved by using alternative approaches. For example, one can use ClonEvol and then integrate only over the selected trees in the probabilistic framework of HINTRA. Alternatively, Monte Carlo Markov Chain approaches as in [108] can be used to sample high likelihood trees in constant time. These approaches are expected to result in a small loss of accuracy as the density over tree topologies would be concentrated in a small area of the search space. Another way of improving the running time is limiting the summation in equation 5.9 to θ values close to the observed CCFs. Because these values are associated with larger likelihoods, we expect this approximation to be close to the true value.

The memory consumption of HINTRA is also shown in Figure 5.8. Based on these results, the number of mutations per sample is the only important factor for the amount of memory used. This affects the number of possible ancestry sets as well as the total number of mutations n. These two values determine the size of the β parameter. Moreover, the number of mutations per sample determines the number of possible topologies, which indicates the number of marginal likelihoods that need to be computed and stored. While HINTRA consumes up to about 12 MB, REVOLVER uses about 4 GB of memory during its execution (not shown in the figure due to the large magnitude). This may be due to the implementation of REVOLVER in R, which is very inefficient compared to C++, which we used to implement HINTRA.

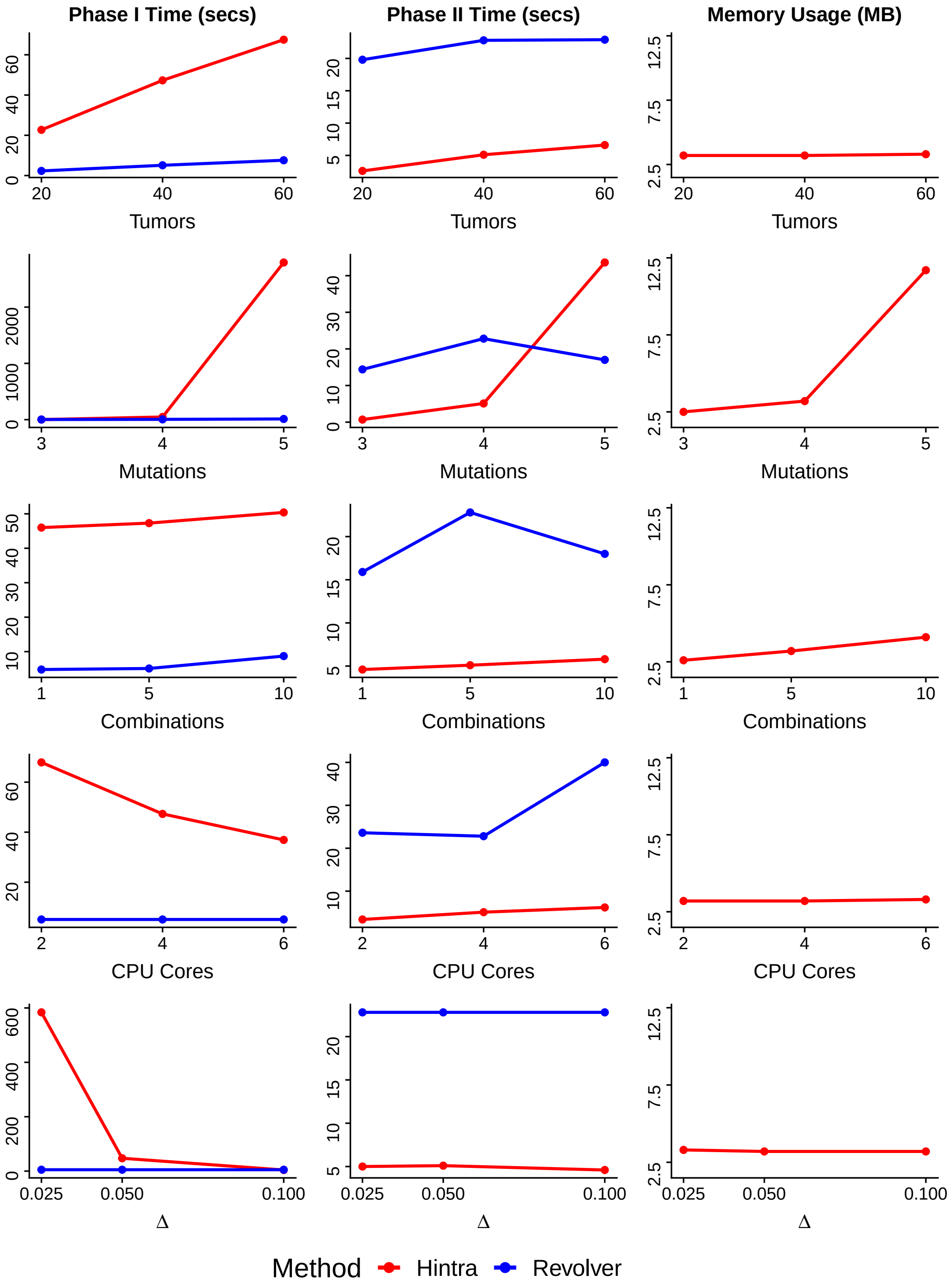


Figure 5.8: Results of running time and memory analysis.

Chapter 6

Conclusions

Understanding the heterogeneity of complex diseases is critical for discovering the underlying mechanisms and designing appropriate treatments. Omics data provide opportunities for studying this heterogeneity. In this thesis, we provided methods for modeling the heterogeneity at two levels: among samples and among cells within a tumor, i.e. intra-tumor heterogeneity. In the case of cancer, we note that these two types of heterogeniety are correlated and complementary and provide a higher resolution picture of disease heterogeneity for a cohort of tumor samples. Therefore, in one of the proposed methods, i.e. HINTRA , we simultaneously account for both levels of heterogeneity. In the next sections, we summarize the thesis and the experimental limitations and discuss future work.

6.1 Summary

In chapter 1, we briefly discussed probabilistic graphical models and motivations behind using them for modeling biological data. Different available omics data types were defined and the importance of understanding the heterogeneity of disease mechanisms captured in these data types was explained. In chapter 2, existing methods for unsupervised and supervised patient stratification and intra-tumor heterogeneity detection was reviewed, discussing their strengths and weaknesses. We described the existing gaps motivating the propositions in the next chapters.

In chapter 3, we propose a novel probabilistic graphical model, called B2PS, for integrative Bayesian biclustering of omic data for patient stratification. The method uses somatic point mutation, copy number variation and gene expression data to identify patient strata and gene clusters. Our experimental results demonstrate the effectiveness of the Bayesian approach for inclusion of prior knowledge and detection of a natural number of clusters. Based on experiments with different combinations of three different data types, we found that gene expression produces the best survival stratification for our datasets when used alone. Integrating with other datasets reduces the performance of B2PS. This is in accordance with the natural choice of gene expression in other stratification studies. It is consistent with the fact that gene expression is closer to the survival phenotype than other genomic data, whose collective effects are already reflected in gene expression data. Our experiments also show that B2PS is more effective in patient stratification than NMF using gene expression data, most likely due to the probabilistic nature of B2PS and its flexibility in the number of clusters across two dimensions. This work is published at [63].

In this chapter 4, an integrative Bayesian probabilistic model for simultaneous analysis of transcriptomic and phenotype data is presented. The model, called SUBSTRA, learns patient strata relevant to a phenotype and detects corresponding transcript clusters. The method also assigns weights to the transcripts based on their relevance to the phenotype and allows for interpretable prediction. SUBSTRA achieves both good interpretability (i.e., produces meaningful patient clusters, transcript clusters, and transcript weights) and accurate phenotype prediction, which is lacking in the state-of-the-art methods for phenotype prediction [134] such as SVM. Based on the simulation results, the combination of transcriptomic and phenotype data improves patient stratification results and helps detect relevant linear and non-linear signals in situations with high noise levels. The biclustering also improves the prediction accuracy in certain simulation experiments. We carried out gene set enrichment analysis of the transcripts identified as important by SUBSTRA in relevant biological scenarios, such as kidney rejection and drug response. We found that SUBSTRA selects more consistent genes with better enrichment values compared to regularized logistic regression models in most of the experiments. Also, analyzing the transcript clusters detected by SUBSTRA indicates that they capture key biological mechanisms that drive the differential fates of these samples and shed light on factors driving predictive performance. These clusters are shown to be more consistent than the alternative methods discussed in this chapter and the prediction accuracy of SUBSTRA is shown to be comparable with the common single-purpose predictive methods, such as LR and SVM. This work is published at [65].

In chapter 5, we presented HINTRA as a new method for collaborative intra-tumor heterogeneity detection. HINTRA is a probabilistic graphical model with a novel tree prior probability that considers all the mutations preceding a particular mutation in the phylogenetic tree, instead of only the most recent one. It uses a Bayesian approach to learn its parameters, which mitigates the bias towards branching topologies found in other tools. We compared HINTRA's performance using synthetic and real data against both a stand-alone and a collaborative method. In our experiments on synthetic datasets, we demonstrated the effectiveness of both proposed tree prior probability and Bayesian learning method using different scenarios. Similarly, for synthetic data from the literature, HINTRA inferred the true phylogenetic trees with more accuracy compared to the state-of-the-art. In our experiments on breast cancer data, HINTRA's findings were consistent with the existing domain knowledge. Moreover, based on the prognostic parameters learned, HINTRA provided new insights of potential interest. This work is published at [66].

6.2 Limitations

Although CNVs are crucial in cancer, we only included SNVs in the experiments with HINTRA on breast cancer data. CNVs were excluded due to the difficulties that they cause in inferring correct CCFs, which would be later transformed into read counts for phylogeny detection by HINTRA. Therefore, the implications in chapter 5 for real data are restricted to SNVs. Although there are tools for inferring the CCF values for CNVs (e.g. PyClone [110]), their accuracy is limited for low-coverage cross-sectional data. Including CNV data should be considered in future work as both the sequencing technologies and CCF inference tools improve.

6.3 Future Work

The subtypes and gene clusters produced by B2PS can serve as a starting point to find subtype-specific gene expression profiles and consequently subtype specific pathways or subnetworks. This information together with the mutation profiles can then be employed to find the driver genetic variations for each subtype, which is the hallmark of stratified medicine. Designing methods that can extract important biclusters based on B2PS's outputs is a potential direction for future work. In cases where gene expression data is collectible (e.g., cancer), this type of data turns out to be more informative than other genomic data for patient stratification at least for the datasets used in this study. For cases where gene expression data cannot be gathered from the relevant tissue, methods like the one proposed in [57], which preprocess the genomic data to reduce their heterogeneity, can be useful. In that respect, future research for B2PS may explore the integration of preprocessed genomic data as well as other data types (e.g., methylation, miRNA expression, and other structural variations like gene fusion).

For simplification and efficiency, we assumed binary expression data in B2PS and SUBSTRA. Examining alternative prior distributions, e.g. Gaussian distribution for continuous gene expression data, is needed. This generalization together with using alternative faster learning algorithms, e.g. variational inference and parallelization, are some technical directions for future research.

As another future work for SUBSTRA, one might extend the method to incorporate more patient and transcript information, such as pathways and interaction data. This might further improve the performance. In scenarios with temporary data access contracts, only the model learned from data is available, but not the dataset itself. For such scenarios, one might leverage the Bayesian properties of SUBSTRA for Lifelong Machine Learning/Continual Learning. The Bayesian nature of this method allows for incorporation of prior knowledge extracted from previously available datasets when training a new model, which might compensate for the lack of access to those data. Moreover, for learning feature weights, using non-convex optimization techniques instead of gradient descent might provide further improvements.

In the current implementation of HINTRA a limited number of mutations can be considered for each patient. This number depends on the available computational resources. Although in some datasets (e.g. the breast cancer dataset used in section 5.3.2) this limitation does not result in a considerable information loss (3.5% of the samples with more than 5 mutations), in general it can limit the findings to only well-known driver genes and a subset of the patients. Part of this problem is resolved by enabling parallel computing. Further improvements in running time can be achieved by using the ideas discussed in section 5.3.3 for future implementation. Although the presented probabilistic framework of HINTRA considers a model for read count data, generalization to other increasingly available data types (e.g. binary data from single cell sequencing) using appropriate distributions (e.g. Bernoulli) is also possible. Investigating the possibility of an intrinsically unbiased prior probability for phylogenetic structures, e.g. conjunctive Bayesian networks, and their applicability in a collaborative framework are other directions for future work.

Bibliography

[1] The Cancer Genome Atlas. http://cancergenome.nih.gov/.

[2] Basel Abu-Jamous, Rui Fa, David J. Roberts, and Asoke K. Nandi. Uncles: method for the identification of genes differentially consistently co-expressed in a specific subset of datasets. *BMC Bioinformatics*, 16(1):184, Jun 2015. ISSN 1471-2105. doi: 10.1186/s12859-015-0614-0. URL https://doi.org/10.1186/s12859-015-0614-0.

[3] Ashar Ahmad and Holger Fröhlich. Towards clinically more relevant dissection of patient heterogeneity via survival-based bayesian clustering. *Bioinformatics*, 33(22): 3558–3566, 2017. doi: 10.1093/bioinformatics/btx464. URL +http://dx.doi.org/10.1093/bioinformatics/btx464.

[4] Muhammad Ammaduddin, Suleiman A. Khan, Disha Malani, Astrid MurumÃďgi, Olli Kallioniemi, Tero Aittokallio, and Samuel Kaski. Drug response prediction by inferring pathway-response associations with kernelized bayesian matrix factorization. *Bioinformatics*, 32:i455–i463, 2016.

[5] Yeji An, Jessica R. Adams, Daniel P. Hollern, Anthony Zhao, Stephen G. Chang, Miki S. Gams, Philip E.D. Chung, Xiaping He, Rhea Jangra, Juhi S. Shah, Joanna Yang, Lauren A. Beck, Nandini Raghuram, Katelyn J. Kozma, Amanda J. Loch, Wei Wang, Cheng Fan, Susan J. Done, Eldad Zacksenhaus, Cynthia J. Guidos, Charles M. Perou, and Sean E. Egan. Cdh1 and pik3ca mutations cooperate to induce immune-related invasive lobular carcinoma of the breast. *Cell Reports*, 25(3):702 – 714.e6, 2018. ISSN 2211-1247. doi: https://doi.org/10.1016/j.celrep.2018.09.056. URL http://www.sciencedirect.com/science/article/pii/S2211124718315018.

[6] Elliott Antman, Scott Weiss, and Joseph Loscalzo. Systems pharmacology, pharmacogenetics, and clinical trial design in network medicine. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 4(4):367–383, 2012. doi: 10.1002/wsbm.1173. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/wsbm.1173.

[7] Samuel J. Aronson and Heidi L. Rehm. Building the foundation for genomics in precision medicine. *Nature*, 526:336, Oct 2015. URL https://doi.org/10.1038/nature15816.

[8] Camille Stephan-Otto Attolini, Yu-Kang Cheng, Rameen Beroukhim, Gad Getz, Omar Abdel-Wahab, Ross L. Levine, Ingo K. Mellinghoff, and Franziska Michor. A mathematical framework to determine the temporal sequence of somatic genetic events in cancer. *Proceedings of the National Academy of Sciences*, 107(41): 17604–17609, 2010. ISSN 0027-8424. doi: 10.1073/pnas.1009117107. URL https://www.pnas.org/content/107/41/17604.

[9] Sergio E. Baranzini, Joann Mudge, Jennifer C. van Velkinburgh, Pouya Khankhanian, Irina Khrebtukova, Neil A. Miller, Lu Zhang, Andrew D. Farmer, Callum J. Bell, Ryan W. Kim, Gregory D. May, Jimmy E. Woodward, Stacy J. Caillier, Joseph P. McElroy, Refujia Gomez, Marcelo J. Pando, Leonda E. Clendenen, Elena E. Ganusova, Faye D. Schilkey, Thiruvarangan Ramaraj, Omar A. Khan, Jim J. Huntley, Shujun Luo, Pui-yan Kwok, Thomas D. Wu, Gary P. Schroth, Jorge R. Oksenberg, Stephen L. Hauser, and Stephen F. Kingsmore. Genome, epigenome and rna sequences of monozygotic twins discordant for multiple sclerosis. *Nature*, 464:1351, Apr 2010. URL `https://doi.org/10.1038/nature08990`.

[10] Jordi Barretina, Giordano Caponigro, Nicolas Stransky, Kavitha Venkatesan, Adam A. Margolin, Sungjoon Kim, Christopher J. Wilson, Joseph Lehar, Gregory V. Kryukov, Dmitriy Sonkin, Anupama Reddy, Manway Liu, Lauren Murray, Michael F. Berger, John E. Monahan, Paula Morais, Jodi Meltzer, Adam Korejwa, Judit Jane-Valbuena, Felipa A. Mapa, Joseph Thibault, Eva Bric-Furlong, Pichai Raman, Aaron Shipway, Ingo H. Engels, Jill Cheng, Guoying K. Yu, Jianjun Yu, Peter Aspesi, Melanie de Silva, Kalpana Jagtap, Michael D. Jones, Li Wang, Charles Hatton, Emanuele Palescandolo, Supriya Gupta, Scott Mahan, Carrie Sougnez, Robert C. Onofrio, Ted Liefeld, Laura MacConaill, Wendy Winckler, Michael Reich, Nanxin Li, Jill P. Mesirov, Stacey B. Gabriel, Gad Getz, Kristin Ardlie, Vivien Chan, Vic E. Myer, Barbara L. Weber, Jeff Porter, Markus Warmuth, Peter Finan, Jennifer L. Harris, Matthew Meyerson, Todd R. Golub, Michael P. Morrissey, William R. Sellers, Robert Schlegel, and Levi A. Garraway. The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, 483(7391):603–607, Mar 2012. ISSN 0028-0836. doi: 10.1038/nature11003. URL `http://dx.doi.org/10.1038/nature11003`.

[11] Chandra Bartholomeusz, Xuemei Xie, Mary Kathryn Pitner, Kimie Kondo, Ali Dadbin, Jangsoon Lee, Hitomi Saso, Paul D. Smith, Kevin N. Dalby, and Naoto T. Ueno. Mek inhibitor selumetinib (azd6244; arry-142886) prevents lung metastasis in a triple-negative breast cancer xenograft model. *Molecular Cancer Therapeutics*, 14(12):2773–2781, 12 2015. ISSN 1535-7163. doi: 10.1158/1535-7163.MCT-15-0243.

[12] Stephen B. Baylin and Peter A. Jones. A decade of exploring the cancer epigenome – biological and translational implications. *Nature Reviews Cancer*, 11:726, Sep 2011. URL `https://doi.org/10.1038/nrc3130`. Perspective.

[13] Jacques S. Beckmann and Daniel Lew. Reconciling evidence-based medicine and precision medicine in the era of big data: challenges and opportunities. *Genome Medicine*, 8(1):134, Dec 2016. ISSN 1756-994X. doi: 10.1186/s13073-016-0388-7. URL `https://doi.org/10.1186/s13073-016-0388-7`.

[14] N. Beerenwinkel and S. Sullivant. Markov models for accumulating mutations. *Biometrika*, 96(3):645–661, 06 2009. ISSN 0006-3444. doi: 10.1093/biomet/asp023. URL `https://dx.doi.org/10.1093/biomet/asp023`.

[15] Niko Beerenwinkel, Jörg Rahnenführer, Martin Däumer, Daniel Hoffmann, Rolf Kaiser, Joachim Selbig, and Thomas Lengauer. Learning multiple evolutionary pathways from cross-sectional data. In *Proceedings of the Eighth Annual International Conference on Resaerch in Computational Molecular Biology*, RECOMB

'04, pages 36–44, New York, NY, USA, 2004. ACM. ISBN 1-58113-755-9. doi: 10.1145/974614.974620. URL http://doi.acm.org/10.1145/974614.974620.

[16] Niko Beerenwinkel, Jörg Rahnenführer, Martin Däumer, Daniel Hoffmann, Rolf Kaiser, Joachim Selbig, and Thomas Lengauer. Learning multiple evolutionary pathways from cross-sectional data. *Journal of Computational Biology*, 12(6):584–598, 2005. doi: 10.1089/cmb.2005.12.584. URL https://doi.org/10.1089/cmb.2005.12.584. PMID: 16108705.

[17] Niko Beerenwinkel, Jörg Rahnenführer, Rolf Kaiser, Daniel Hoffmann, Joachim Selbig, and Thomas Lengauer. Mtreemix: a software package for learning and using mixture models of mutagenetic trees. *Bioinformatics*, 21(9):2106–2107, 01 2005. ISSN 1367-4803. doi: 10.1093/bioinformatics/bti274. URL https://dx.doi.org/10.1093/bioinformatics/bti274.

[18] Niko Beerenwinkel, Nicholas Eriksson, and Bernd Sturmfels. Evolution on distributive lattices. *Journal of Theoretical Biology*, 242(2):409 – 420, 2006. ISSN 0022-5193. doi: https://doi.org/10.1016/j.jtbi.2006.03.013. URL http://www.sciencedirect.com/science/article/pii/S0022519306001159.

[19] Niko Beerenwinkel, Nicholas Eriksson, and Bernd Sturmfels. Conjunctive Bayesian Networks. *Bernoulli*, 13(4):893–909, 11 2007. doi: 10.3150/07-BEJ6133. URL https://doi.org/10.3150/07-BEJ6133.

[20] Bradley E. Bernstein, Alexander Meissner, and Eric S. Lander. The mammalian epigenome. *Cell*, 128(4):669 – 681, 2007. ISSN 0092-8674. doi: https://doi.org/10.1016/j.cell.2007.01.033. URL http://www.sciencedirect.com/science/article/pii/S0092867407001286.

[21] W.M. Bolstad and J.M. Curran. *Introduction to Bayesian Statistics*. Wiley, 2016. ISBN 9781118091562. URL https://books.google.ca/books?id=BxfkDAAAQBAJ.

[22] Jean-Philippe Brunet, Pablo Tamayo, Todd R. Golub, and Jill P. Mesirov. Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the National Academy of Sciences*, 101(12):4164–4169, 2004. ISSN 0027-8424. doi: 10.1073/pnas.0308531101. URL https://www.pnas.org/content/101/12/4164.

[23] Giulio Caravagna, Ylenia Giarratano, Daniele Ramazzotti, Ian Tomlinson, Trevor A. Graham, Guido Sanguinetti, and Andrea Sottoriva. Detecting repeated cancer evolution from multi-region tumor sequencing data. *Nature Methods*, 15(9):707–714, 2018. ISSN 1548-7105. doi: 10.1038/s41592-018-0108-x. URL https://doi.org/10.1038/s41592-018-0108-x.

[24] George Casella and Edward I. George. Explaining the gibbs sampler. *The American Statistician*, 46(3):167–174, 1992. doi: 10.1080/00031305.1992.10475878. URL https://www.tandfonline.com/doi/abs/10.1080/00031305.1992.10475878.

[25] Ethan Cerami, Jianjiong Gao, Ugur Dogrusoz, Benjamin E Gross, Selcuk Onur Sumer, Bülent Arman Aksoy, Anders Jacobsen, Caitlin J Byrne, Michael L Heuer, Erik Larsson, et al. The cbio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data, 2012.

[26] Ethan G Cerami, Benjamin E Gross, Emek Demir, Igor Rodchenkov, Özgün Babur, Nadia Anwar, Nikolaus Schultz, Gary D Bader, and Chris Sander. Pathway commons, a web resource for biological pathway data. *Nucleic acids research*, 39(suppl_1):D685–D690, 2010.

[27] Rui Chen and Michael Snyder. Promise of personalized omics to precision medicine. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 5(1):73–82, 2013. doi: 10.1002/wsbm.1198. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/wsbm.1198`.

[28] Dong-Yeon Cho and Teresa M. Przytycka. Dissecting cancer heterogeneity with a probabilistic genotype-phenotype model. *Nucleic Acids Research*, 41(17):8011–8020, 07 2013. ISSN 0305-1048. doi: 10.1093/nar/gkt577. URL `https://doi.org/10.1093/nar/gkt577`.

[29] Francis S. Collins. Exceptional Opportunities in Medical Science: A View From the National Institutes of Health. *JAMA*, 313(2):131–132, 01 2015. ISSN 0098-7484. doi: 10.1001/jama.2014.16736. URL `https://doi.org/10.1001/jama.2014.16736`.

[30] Simona Cristea, Jack Kuipers, and Niko Beerenwinkel. pathTiMEx: Joint inference of mutually exclusive cancer pathways and their progression dynamics. *Journal of Computational Biology*, 24(6):603–615, 2017. doi: 10.1089/cmb.2016.0171. URL `https://doi.org/10.1089/cmb.2016.0171`. PMID: 27936934.

[31] H X Dang, B S White, S M Foltz, C A Miller, J Luo, R C Fields, and C A Maher. ClonEvol: clonal ordering and visualization in cancer sequencing. *Annals of Oncology*, 28(12):3076–3082, 2017. doi: 10.1093/annonc/mdx517. URL `http://dx.doi.org/10.1093/annonc/mdx517`.

[32] L. De Mattos-Arruda, B. Weigelt, J. Cortes, H. H. Won, C. K. Y. Ng, P. Nuciforo, F.-C. Bidard, C. Aura, C. Saura, V. Peg, S. Piscuoglio, M. Oliveira, Y. Smolders, P. Patel, L. Norton, J. Tabernero, M. F. Berger, J. Seoane, and J. S. Reis-Filho. Capturing intra-tumor genetic heterogeneity by de novo mutation profiling of circulating cell-free tumor dna: a proof-of-principle. *Annals of Oncology*, 25(9):1729–1735, 2014. doi: 10.1093/annonc/mdu239. URL `http://dx.doi.org/10.1093/annonc/mdu239`.

[33] Amit G Deshwar, Shankar Vembu, Christina K Yung, Gun Ho Jang, Lincoln Stein, and Quaid Morris. PhyloWGS: reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome biology*, 16(1):35, 2015.

[34] Richard Desper, Feng Jiang, Olli-P Kallioniemi, Holger Moch, Christos H. Papadimitriou, and Alejandro A. Schäffer. Inferring tree models for oncogenesis from comparative genome hybridization data. *Journal of Computational Biology*, 6(1):37–51, 1999. doi: 10.1089/cmb.1999.6.37. URL `https://doi.org/10.1089/cmb.1999.6.37`. PMID: 10223663.

[35] Richard Desper, Feng Jiang, Olli-P. Kallioniemi, Holger Moch, Christos H. Papadimitriou, and Alejandro A. Schäffer. Distance-based reconstruction of tree models for oncogenesis. *Journal of Computational Biology*, 7(6):789–803, 2000. doi: 10.1089/10665270050514936. URL `https://doi.org/10.1089/10665270050514936`. PMID: 11382362.

[36] Nilgun Donmez, Salem Malikic, Alexander W Wyatt, Martin E Gleave, Colin C Collins, and S Cenk Sahinalp. Clonality inference from single tumor samples using low-coverage sequence data. *Journal of Computational Biology*, 24(6):515–523, 2017. doi: 10.1089/cmb.2016.0148. URL `https://doi.org/10.1089/cmb.2016.0148`.

[37] Pieter C. Dorrestein, Sarkis K. Mazmanian, and Rob Knight. Finding the missing links among metabolites, microbes, and the host. *Immunity*, 40(6):824 – 832, 2014. ISSN 1074-7613. doi: https://doi.org/10.1016/j.immuni.2014.05.015. URL `http://www.sciencedirect.com/science/article/pii/S1074761314001952`.

[38] Leo L Duan, John P Clancy, and Rhonda D Szczesniak. Bayesian ensemble trees (BET) for clustering and prediction in heterogeneous data. *Journal of Computational and Graphical Statistics*, 25(3):748–761, 2016.

[39] Steffen Durinck, Yves Moreau, Arek Kasprzyk, Sean Davis, Bart De Moor, Alvis Brazma, and Wolfgang Huber. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics*, 21(16):3439–3440, 06 2005. ISSN 1367-4803. doi: 10.1093/bioinformatics/bti525. URL `https://doi.org/10.1093/bioinformatics/bti525`.

[40] Gunilla Einecke, Jeff Reeve, Banu Sis, Michael Mengel, Luis Hidalgo, Konrad S. Famulski, Arthur Matas, Bert Kasiske, Bruce Kaplan, and Philip F. Halloran. A molecular classifier for predicting future graft loss in late kidney transplant biopsies. *Journal of Clinical Investigation*, 120(6):1862–1872, 2010.

[41] Mohammed El-Kebir, Layla Oesper, Hannah Acheson-Field, and Benjamin J Raphael. Reconstruction of clonal trees and tumor composition from multi-sample sequencing data. *Bioinformatics*, 31(12):i62–i70, 2015.

[42] Jianjiong Gao, Bülent Arman Aksoy, Ugur Dogrusoz, Gideon Dresdner, Benjamin Gross, S Onur Sumer, Yichao Sun, Anders Jacobsen, Rileen Sinha, Erik Larsson, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cbioportal. *Sci. Signal.*, 6(269):pl1–pl1, 2013.

[43] Moritz Gerstung, Michael Baudis, Holger Moch, and Niko Beerenwinkel. Quantifying cancer progression with conjunctive Bayesian networks. *Bioinformatics*, 25(21):2809–2815, 2009. doi: 10.1093/bioinformatics/btp505. URL `http://dx.doi.org/10.1093/bioinformatics/btp505`.

[44] Moritz Gerstung, Nicholas Eriksson, Jimmy Lin, Bert Vogelstein, and Niko Beerenwinkel. The temporal order of genetic and pathway alterations in tumorigenesis. *PLOS ONE*, 6(11):1–9, 11 2011. doi: 10.1371/journal.pone.0027136. URL `https://doi.org/10.1371/journal.pone.0027136`.

[45] Moritz Gerstung, Clemency Jolly, Ignaty Leshchiner, Stefan C. Dentro, Santiago Gonzalez Rosado, Daniel Rosebrock, Thomas J. Mitchell, Yulia Rubanova, Pavana Anur, Kaixan Yu, Maxime Tarabichi, Amit Deshwar, Jeff Wintersinger, Kortine Kleinheinz, Ignacio Vazquez-Garcia, Kerstin Haase, Lara Jerman, Subhajit Sengupta, Geoff Macintyre, Salem Malikic, Nilgun Donmez, Dimitri G. Livitz, Marek Cmero, Jonas Demeulemeester, Steve Schumacher, Yu Fan, Xiaotong Yao, Juhee Lee, Matthias

Schlesner, Paul C. Boutros, David D. Bowtell, Hongtu Zhu, Gad Getz, Marcin Imielinski, Rameen Beroukhim, S. Cenk Sahinalp, Yuan Ji, Martin Peifer, Florian Markowetz, Ville Mustonen, Ke Yuan, Wenyi Wang, Quaid D. Morris, Paul T. Spellman, David C. Wedge, and Peter Van Loo. The evolutionary history of 2,658 cancers. *bioRxiv*, 2018. doi: 10.1101/161562. URL `https://www.biorxiv.org/content/early/2018/09/12/161562`.

[46] Vladimir Gligorijevic, Noel Malod-Dognin, and Natasa Przulj. Patient-specific data fusion for cancer stratification and personalised treatment. In *Proceedings of Pacific Symposium on Biocomputing*, January 2016.

[47] Gene H Golub and Christian Reinsch. Singular value decomposition and least squares solutions. In *Linear Algebra*, pages 134–151. Springer, 1971.

[48] Mehmet Gönen and Samuel Kaski. Kernelized bayesian matrix factorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(10):2047–2060, Oct 2014. ISSN 0162-8828. doi: 10.1109/TPAMI.2014.2313125.

[49] Gavin J. Gordon, Roderick V. Jensen, Li-Li Hsiao, Steven R. Gullans, Joshua E. Blumenstock, Sridhar Ramaswamy, William G. Richards, David J. Sugarbaker, and Raphael Bueno. Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. *Cancer Research*, 62(17):4963–4967, 2002. ISSN 0008-5472. URL `http://cancerres.aacrjournals.org/content/62/17/4963`.

[50] Peter J. E. Goss and Jean Peccoud. Quantitative modeling of stochastic systems in molecular biology by using stochastic petri nets. *Proceedings of the National Academy of Sciences*, 95(12):6750–6755, 1998. ISSN 0027-8424. doi: 10.1073/pnas.95.12.6750. URL `https://www.pnas.org/content/95/12/6750`.

[51] Rebecca Graziani, Michele Guindani, and Peter F Thall. Bayesian nonparametric estimation of targeted agent effects on biomarker change to predict clinical outcome. *Biometrics*, 71(1):188–197, 2015.

[52] Iman Hajirasouliha, Ahmad Mahmoody, and Benjamin J Raphael. A combinatorial approach for analyzing intra-tumor heterogeneity from high-throughput sequencing data. *Bioinformatics*, 30(12):i78–i86, 2014.

[53] Bhavana Harsha, Chai Yin Kok, Charlotte G. Cole, David Beare, Elisabeth Dawson, Harry Boutselakis, Harry Jubb, John Tate, Laura Ponting, Mingming Jia, Nidhi Bindal, Peter J. Campbell, Raymund Stefancsik, Sally Bamford, Sam Thompson, Sari Ward, Tisham De, Zbyslaw Sondka, and Simon A. Forbes. COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Research*, 45(D1):D777–D783, 11 2016. ISSN 0305-1048. doi: 10.1093/nar/gkw1121. URL `https://dx.doi.org/10.1093/nar/gkw1121`.

[54] Paul Helman, Robert Veroff, Susan R. Atlas, and Cheryl Willman. A bayesian network classification methodology for gene expression data. *Journal of Computational Biology*, 11(4):581–615, 2004. doi: 10.1089/cmb.2004.11.581. URL `https://doi.org/10.1089/cmb.2004.11.581`. PMID: 15579233.

[55] Marcus Hjelm, Mattias Höglund, and Jens Lagergren. New probabilistic network models and algorithms for oncogenesis. *Journal of Computational Biology*, 13(4):853–865, 2006. doi: 10.1089/cmb.2006.13.853. URL `https://doi.org/10.1089/cmb.2006.13.853`. PMID: 16761915.

[56] Sepp Hochreiter, Ulrich Bodenhofer, Martin Heusel, Andreas Mayr, Andreas Mitterecker, Adetayo Kasim, Tatsiana Khamiakova, Suzy Van Sanden, Dan Lin, Willem Talloen, Luc Bijnens, Hinrich W. H. Göhlmann, Ziv Shkedy, and Djork-Arnè Clevert. FABIA: factor analysis for bicluster acquisition. *Bioinformatics*, 26(12):1520–1527, 04 2010. ISSN 1367-4803. doi: 10.1093/bioinformatics/btq227. URL `https://doi.org/10.1093/bioinformatics/btq227`.

[57] Matan Hofree, John P. Shen, Hannah Carter, Andrew Gross, and Trey Ideker. Network-based stratification of tumor mutations. *Nature Methods*, 10:1108, Sep 2013. URL `https://doi.org/10.1038/nmeth.2651`. Article.

[58] Yu Hou, Huahu Guo, Chen Cao, Xianlong Li, Boqiang Hu, Ping Zhu, Xinglong Wu, Lu Wen, Fuchou Tang, Yanyi Huang, and Jirun Peng. Single-cell triple omics sequencing reveals genetic, epigenetic, and transcriptomic heterogeneity in hepatocellular carcinomas. *Cell Research*, 26:304, Feb 2016. URL `https://doi.org/10.1038/cr.2016.23`. Original Article.

[59] D. Husmeier, R. Dybowski, and S. Roberts. *Probabilistic Modeling in Bioinformatics and Medical Informatics*. Advanced Information and Knowledge Processing. Springer London, 2006. ISBN 9781846281198. URL `https://books.google.ca/books?id=5A1xjUUsNpsC`.

[60] Katharina Jahn, Jack Kuipers, and Niko Beerenwinkel. Tree inference for single-cell data. *Genome biology*, 17(1):86, 2016.

[61] Wei Jiao, Shankar Vembu, Amit G Deshwar, Lincoln Stein, and Quaid Morris. Inferring clonal evolution of tumors from single nucleotide somatic mutations. *BMC bioinformatics*, 15(1):35, 2014.

[62] Segun Jung, Yingtao Bi, and Ramana V. Davuluri. Evaluation of data discretization methods to derive platform independent isoform expression signatures for multi-class tumor subtyping. *BMC Genomics*, 16(11):S3, Nov 2015. ISSN 1471-2164. doi: 10.1186/1471-2164-16-S11-S3. URL `https://doi.org/10.1186/1471-2164-16-S11-S3`.

[63] S. Khakabimamaghani and M. Ester. Bayesian biclustering for patient stratification. In *Proceedings of Pacific Symposium on Biocomputing*, January 2016.

[64] Sahand Khakabimamaghani, Dujian Ding, Oliver Snow, and Martin Ester. Uncovering the subtype-specific temporal order of cancer pathwaydys regulation. *bioRxiv*, 2019. URL `https://www.biorxiv.org/content/early/`.

[65] Sahand Khakabimamaghani, Yogeshwar D Kelkar, Bruno M Grande, Ryan D Morin, Martin Ester, and Daniel Ziemek. SUBSTRA: Supervised Bayesian Patient Stratification. *Bioinformatics*, 02 2019. ISSN 1367-4803. doi: 10.1093/bioinformatics/btz112. URL `https://doi.org/10.1093/bioinformatics/btz112`.

[66] Sahand Khakabimamaghani, Salem Malikic, Jeffrey Tang, Dujian Ding, Ryan Morin, Leonid Chindelevitch, and Martin Ester. Collaborative intra-tumor heterogeneity detection. *Bioinformatics*, 35(14):i379–i388, 07 2019. ISSN 1367-4803. doi: 10.1093/bioinformatics/btz355. URL https://doi.org/10.1093/bioinformatics/btz355.

[67] Purvesh Khatri, Silke Roedder, Naoyuki Kimura, Katrien De Vusser, Alexander A. Morgan, Yongquan Gong, Michael P. Fischbein, Robert C. Robbins, Maarten Naesens, Atul J. Butte, and Minnie M. Sarwal. A common rejection module (CRM) for acute rejection across multiple organs identifies novel therapeutics for organ transplantation. *The Journal of Experimental Medicine*, 210(11):2205–2221, 2013.

[68] Kyung In Kim and Richard Simon. Using single cell sequencing data to model the evolutionary history of a tumor. *BMC bioinformatics*, 15(1):27, 2014.

[69] Philip M Kim and Bruce Tidor. Subsystem identification through dimensionality reduction of large-scale gene expression data. *Genome research*, 13(7):1706–1718, 2003.

[70] Yuko Komiya and Raymond Habas. Wnt signal transduction pathways. *Organogenesis*, 4(2):68–75, 2008.

[71] J. Larry Jameson and Dan L. Longo. Precision medicine–personalized, problematic, and promising. *Obstetrical & Gynecological Survey*, 70(10), 2015. ISSN 0029-7828. URL https://journals.lww.com/obgynsurvey/Fulltext/2015/10000/Precision_Medicine_Personalized,_Problematic,_and.7.aspx.

[72] Laura Lazzeroni and Art Owen. Plaid models for gene expression data. *Statistica Sinica*, 12(1):61–86, 2002. ISSN 10170405, 19968507. URL http://www.jstor.org/stable/24307036.

[73] Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788, 1999.

[74] Daniel D. Lee and H. Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788, Oct 1999. URL http://dx.doi.org/10.1038/44565.

[75] Daniel D. Lee and H. Sebastian Seung. Algorithms for non-negative matrix factorization. In T. K. Leen, T. G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems 13*, pages 556–562. MIT Press, 2001. URL http://papers.nips.cc/paper/1861-algorithms-for-non-negative-matrix-factorization.pdf.

[76] Juhee Lee, Peter Müller, Subhajit Sengupta, Kamalakar Gulukota, and Yuan Ji. Bayesian inference for intratumour heterogeneity in mutations and copy number variation. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 65(4):547–563, 2016. doi: 10.1111/rssc.12136. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssc.12136.

[77] Mihee Lee, Haipeng Shen, Jianhua Z Huang, and JS Marron. Biclustering via sparse singular value decomposition. *Biometrics*, 66(4):1087–1095, 2010.

[78] Xiyan Li, Tara A. Gianoulis, Kevin Y. Yip, Mark Gerstein, and Michael Snyder. Extensive in vivo metabolite-protein interactions revealed by large-scale systematic analyses. *Cell*, 143(4):639 – 650, 2010. ISSN 0092-8674. doi: https://doi.org/10.1016/j.cell.2010.09.048. URL `http://www.sciencedirect.com/science/article/pii/S0092867410011347`.

[79] Arthur Liberzon, Aravind Subramanian, Reid Pinchback, Helga Thorvaldsdóttir, Pablo Tamayo, and Jill P. Mesirov. Molecular signatures database (msigdb) 3.0. *Bioinformatics*, 27(12):1739–1740, 2011. doi: 10.1093/bioinformatics/btr260. URL `+http://dx.doi.org/10.1093/bioinformatics/btr260`.

[80] Xihui Lin and Paul C Boutros. *NNLM: Fast and Versatile Non-Negative Matrix Factorization*, 2018. URL `https://CRAN.R-project.org/package=NNLM`. R package version 0.4.2.

[81] Hongfu Liu, Rui Zhao, Hongsheng Fang, Feixiong Cheng, Yun Fu, and Yang-Yu Liu. Entropy-based consensus clustering for patient stratification. *Bioinformatics*, 33(17): 2691–2698, 2017.

[82] Loes Olde Loohuis, Giulio Caravagna, Alex Graudenzi, Daniele Ramazzotti, Giancarlo Mauri, Marco Antoniotti, and Bud Mishra. Inferring tree causal models of cancer progression with probability raising. *PLOS ONE*, 9(10):1–14, 10 2014. doi: 10.1371/journal.pone.0108358. URL `https://doi.org/10.1371/journal.pone.0108358`.

[83] Joseph Loscalzo and Albert-Laszlo Barabasi. Systems biology and the future of medicine. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 3(6):619–627, 2011. doi: 10.1002/wsbm.144. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/wsbm.144`.

[84] Friedrich C. Luft. Personalizing precision medicine. *Journal of the American Society of Hypertension*, 9(6):415–416, Jun 2015. ISSN 1933-1711. doi: 10.1016/j.jash.2015.03.289. URL `https://doi.org/10.1016/j.jash.2015.03.289`.

[85] Qiang Ma and Anthony Y. H. Lu. Pharmacogenetics, pharmacogenomics, and individualized medicine. *Pharmacological Reviews*, 63(2):437–459, 2011. ISSN 0031-6997. doi: 10.1124/pr.110.003533. URL `http://pharmrev.aspetjournals.org/content/63/2/437`.

[86] Salem Malikic, Andrew W. McPherson, Nilgun Donmez, and Cenk S. Sahinalp. Clonality inference in multiple tumor samples using phylogeny. *Bioinformatics*, 31(9): 1349–1356, 2015. doi: 10.1093/bioinformatics/btv003. URL `http://dx.doi.org/10.1093/bioinformatics/btv003`.

[87] Salem Malikic, Katharina Jahn, Jack Kuipers, Cenk Sahinalp, and Niko Beerenwinkel. Integrative inference of subclonal tumour evolution from single-cell and bulk sequencing data. *bioRxiv*, 2017. doi: 10.1101/234914. URL `https://www.biorxiv.org/content/early/2017/12/15/234914`.

[88] Salem Malikic, Simone Ciccolella, Farid Rashidi Mehrabadi, Camir Ricketts, Md Khaledur Rahman, Ehsan Haghshenas, Daniel Seidman, Faraz Hach, Iman Hajirasouliha, and S Cenk Sahinalp. PhISCS-a combinatorial approach for sub-perfect

tumor phylogeny reconstruction via integrative use of single cell and bulk sequencing data. *bioRxiv*, page 376996, 2018.

[89] N. Mantel. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemother Rep*, 50:163–170, 1966. URL `https://ci.nii.ac.jp/naid/10004996110/en/`.

[90] KV Mardia, JT Kent, and JM Bibby. Multivariate analysis. *Probability and mathematical statistics. Academic Press Inc*, 1979.

[91] Edward Meeds and Sam Roweis. Nonparametric Bayesian biclustering. Technical report, University of Toronto, Department of Computer Science, June 2007.

[92] Craig H. Mermel, Steven E. Schumacher, Barbara Hill, Matthew L. Meyerson, Rameen Beroukhim, and Gad Getz. Gistic2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biology*, 12(4):R41, Apr 2011. ISSN 1474-760X. doi: 10.1186/gb-2011-12-4-r41. URL `https://doi.org/10.1186/gb-2011-12-4-r41`.

[93] Navodit Misra, Ewa Szczurek, and Martin Vingron. Inferring the paths of somatic evolution in cancer. *Bioinformatics*, 30(17):2456–2463, 2014. doi: 10.1093/bioinformatics/btu319. URL `http://dx.doi.org/10.1093/bioinformatics/btu319`.

[94] Meeta Mistry and Paul Pavlidis. Gene ontology term overlap as a measure of gene functional similarity. *BMC Bioinformatics*, 9(1):327, Aug 2008. ISSN 1471-2105. doi: 10.1186/1471-2105-9-327. URL `https://doi.org/10.1186/1471-2105-9-327`.

[95] S. Monti, P. Tamayo, J. Mesirov, and T. Golub. Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. *Machine Learning*, 52:91–118, 2003.

[96] Toru Mukohara. Pi3k mutations in breast cancer: prognostic and therapeutic implications. *Breast Cancer (Dove Med Press)*, 7:111–123, May 2015. ISSN 1179-1314. doi: 10.2147/BCTT.S60696. URL `https://www.ncbi.nlm.nih.gov/pubmed/26028978`. PMC4440424[pmcid].

[97] Cancer Genome Atlas Network et al. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61, 2012.

[98] Ali Oghabian, Sami Kilpinen, Sampsa Hautaniemi, and Elena Czeizler. Biclustering methods: Biological relevance and application in gene expression analysis. *PLOS ONE*, 9(3):1–10, 03 2014. doi: 10.1371/journal.pone.0090801. URL `https://doi.org/10.1371/journal.pone.0090801`.

[99] Swapnali Pathare, Alejandro A. Schäffer, Niko Beerenwinkel, and Manoj Mahimkar. Construction of oncogenetic tree models reveals multiple pathways of oral cancer progression. *International Journal of Cancer*, 124(12):2864–2871, 2009. doi: 10.1002/ijc.24267. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/ijc.24267`.

[100] Beatriz Pontes, Raul Giraldez, and Jesus S. Aguilar-Ruiz. Biclustering on expression data: A review. *Journal of Biomedical Informatics*, 57:163 – 180, 2015. ISSN 1532-0464. doi: https://doi.org/10.1016/j.jbi.2015.06.028. URL `http://www.sciencedirect.com/science/article/pii/S1532046415001380`.

[101] Victoria Popic, Raheleh Salari, Iman Hajirasouliha, Dorna Kashef-Haghighi, Robert B West, and Serafim Batzoglou. Fast and scalable inference of multi-sample cancer lineages. *Genome biology*, 16(1):91, 2015.

[102] Amela Prelić, Eckart Zitzler, Lothar Thiele, Stefan Bleuler, Lars Hennig, Philip Zimmermann, Wilhelm Gruissem, Anja Wille, and Peter Bühlmann. A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics*, 22(9):1122–1129, 02 2006. ISSN 1367-4803. doi: 10.1093/bioinformatics/btl060. URL `https://doi.org/10.1093/bioinformatics/btl060`.

[103] Daniele Ramazzotti, Giulio Caravagna, Loes Olde Loohuis, Alex Graudenzi, Ilya Korsunsky, Giancarlo Mauri, Marco Antoniotti, and Bud Mishra. CAPRI: efficient inference of cancer progression models from cross-sectional data. *Bioinformatics*, 31(18):3016–3026, 2015. doi: 10.1093/bioinformatics/btv296. URL `http://dx.doi.org/10.1093/bioinformatics/btv296`.

[104] J. Ramey. A collection of small-sample, high-dimensional microarray data sets to assess machine-learning algorithms and models, 2011. URL `https://github.com/ramhiser/datamicroarray1`.

[105] Benjamin J. Raphael and Fabio Vandin. Simultaneous inference of cancer pathways and tumor progression from cross-sectional mutation data. In Roded Sharan, editor, *Research in Computational Molecular Biology*, pages 250–264, Cham, 2014. Springer International Publishing. ISBN 978-3-319-05269-4.

[106] Yordan P Raykov, Alexis Boukouvalas, Fahd Baig, and Max A Little. What to do when k-means clustering fails: A simple yet principled alternative algorithm. *PloS one*, 11(9):e0162259, 2016.

[107] Pedram Razavi, Matthew T. Chang, Guotai Xu, Chaitanya Bandlamudi, Dara S. Ross, Neil Vasan, Yanyan Cai, Craig M. Bielski, Mark T. A. Donoghue, Philip Jonsson, Alexander Penson, Ronglai Shen, Fresia Pareja, Ritika Kundra, Sumit Middha, Michael L. Cheng, Ahmet Zehir, Cyriac Kandoth, Ruchi Patel, Kety Huberman, Lillian M. Smyth, Komal Jhaveri, Shanu Modi, Tiffany A. Traina, Chau Dang, Wen Zhang, Britta Weigelt, Bob T. Li, Marc Ladanyi, David M. Hyman, Nikolaus Schultz, Mark E. Robson, Clifford Hudis, Edi Brogi, Agnes Viale, Larry Norton, Maura N. Dickler, Michael F. Berger, Christine A. Iacobuzio-Donahue, Sarat Chandarlapaty, Maurizio Scaltriti, Jorge S. Reis-Filho, David B. Solit, Barry S. Taylor, and José Baselga. The genomic landscape of endocrine-resistant advanced breast cancers. *Cancer Cell*, 34(3):427–438.e6, Sep 2018. ISSN 1878-3686. doi: 10.1016/j.ccell.2018.08.008. URL `https://www.ncbi.nlm.nih.gov/pubmed/30205045`. PMC6327853[pmcid].

[108] Edith M. Ross and Florian Markowetz. OncoNEM: inferring tumor evolution from single-cell sequencing data. *Genome Biology*, 17(1):69, Apr 2016. ISSN 1474-760X. doi: 10.1186/s13059-016-0929-9. URL `https://doi.org/10.1186/s13059-016-0929-9`.

[109] J. C. Ross, P. J. Castaldi, M. H. Cho, J. Chen, Y. Chang, J. G. Dy, E. K. Silverman, G. R. Washko, and R. S. José Estépar. A bayesian nonparametric model for disease subtyping: Application to emphysema phenotypes. *IEEE Transactions on Medical Imaging*, 36(1):343–354, Jan 2017. ISSN 0278-0062. doi: 10.1109/TMI.2016.2608782.

[110] Andrew Roth, Jaswinder Khattra, Damian Yap, Adrian Wan, Emma Laks, Justina Biele, Gavin Ha, Samuel Aparicio, Alexandre Bouchard-Côté, and Sohrab P. Shah. PyClone: statistical inference of clonal population structure in cancer. *Nat Methods*, 11(4):396–398, Apr 2014. ISSN 1548-7105. doi: 10.1038/nmeth.2883. URL https://www.ncbi.nlm.nih.gov/pubmed/24633410.

[111] Thomas Sakoparnig and Niko Beerenwinkel. Efficient sampling for Bayesian inference of conjunctive Bayesian networks. *Bioinformatics*, 28(18):2318–2324, 07 2012. ISSN 1367-4803. doi: 10.1093/bioinformatics/bts433. URL https://dx.doi.org/10.1093/bioinformatics/bts433.

[112] Sohrab Salehi, Adi Steif, Andrew Roth, Samuel Aparicio, Alexandre Bouchard-Côté, and Sohrab P. Shah. ddclone: joint statistical inference of clonal populations from single cell and bulk tumour sequencing data. *Genome Biology*, 18(1):44, Mar 2017. ISSN 1474-760X. doi: 10.1186/s13059-017-1169-3. URL https://doi.org/10.1186/s13059-017-1169-3.

[113] Subhajit Sengupta, Jin Wang, Juhee Lee, Peter Müller, Kamalkar Gulukota, Arunava Banerjee, and Yuan Ji. *BayClone: Bayesian Nonparametric Inference of Tumor Subclones Using NGS Data*, pages 467–478. 2015. doi: 10.1142/9789814644730_0044. URL https://www.worldscientific.com/doi/abs/10.1142/9789814644730_0044.

[114] Hossein Shahrabi Farahani and Jens Lagergren. Learning oncogenetic networks by reducing to mixed integer linear programming. *PLOS ONE*, 8(6):1–8, 06 2013. doi: 10.1371/journal.pone.0065773. URL https://doi.org/10.1371/journal.pone.0065773.

[115] Fanhua Shang, LC Jiao, and Fei Wang. Graph dual regularization non-negative matrix factorization for co-clustering. *Pattern Recognition*, 45(6):2237–2250, 2012.

[116] Ronglai Shen, Adam B. Olshen, and Marc Ladanyi. 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):2906–2912, 09 2009. ISSN 1367-4803. doi: 10.1093/bioinformatics/btp543. URL https://doi.org/10.1093/bioinformatics/btp543.

[117] Ronglai Shen, Qianxing Mo, Nikolaus Schultz, Venkatraman E. Seshan, Adam B. Olshen, Jason Huse, Marc Ladanyi, and Chris Sander. Integrative subtype discovery in glioblastoma using iCluster. *PLOS ONE*, 7(4):1–9, 04 2012. doi: 10.1371/journal.pone.0035236. URL https://doi.org/10.1371/journal.pone.0035236.

[118] Larry Smarr. Quantifying your body: A how-to guide from a systems biology perspective. *Biotechnology Journal*, 7(8):980–991, 2012. doi: 10.1002/biot.201100495. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/biot.201100495.

[119] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical bayesian optimization of machine learning algorithms. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 2951–2959. Curran Associates, Inc., 2012. URL http://papers.nips.cc/paper/

4522-practical-bayesian-optimization-of-machine-learning-algorithms.pdf.

[120] R. Srivastava, L. You, J. Summers, and J. Yin. Stochastic vs. deterministic modeling of intracellular viral kinetics. *Journal of Theoretical Biology*, 218(3):309 – 321, 2002. ISSN 0022-5193. doi: https://doi.org/10.1006/jtbi.2002.3078. URL http://www.sciencedirect.com/science/article/pii/S002251930293078X.

[121] Katherine Stemke-Hale, Ana Maria Gonzalez-Angulo, Ana Lluch, Richard M. Neve, Wen-Lin Kuo, Michael Davies, Mark Carey, Zhi Hu, Yinghui Guan, Aysegul Sahin, W. Fraser Symmans, Lajos Pusztai, Laura K. Nolden, Hugo Horlings, Katrien Berns, Mien-Chie Hung, Marc J. van de Vijver, Vicente Valero, Joe W. Gray, René Bernards, Gordon B. Mills, and Bryan T. Hennessy. An integrative genomic and proteomic analysis of PIK3CA, PTEN, and AKT mutations in breast cancer. *Cancer Research*, 68(15):6084–6091, 2008. ISSN 0008-5472. doi: 10.1158/0008-5472.CAN-07-6854. URL http://cancerres.aacrjournals.org/content/68/15/6084.

[122] Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, Scott L. Pomeroy, Todd R. Golub, Eric S. Lander, and Jill P. Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, 2005. doi: 10.1073/pnas.0506580102. URL http://www.pnas.org/content/102/43/15545.abstract.

[123] Jiangwen Sun, Jinbo Bi, and Henry R. Kranzler. Multi-view singular value decomposition for disease subtyping and genetic associations. *BMC Genetics*, 15 (1):73, Jun 2014. ISSN 1471-2156. doi: 10.1186/1471-2156-15-73. URL https://doi.org/10.1186/1471-2156-15-73.

[124] Yan V. Sun and Yi-Juan Hu. Chapter three - integrative analysis of multi-omics data for discovery and functional studies of complex human diseases. In Theodore Friedmann, Jay C. Dunlap, and Stephen F. Goodwin, editors, *Advances in Genetics*, volume 93, pages 147 – 190. Academic Press, 2016. doi: https://doi.org/10.1016/bs.adgen.2015.11.004. URL http://www.sciencedirect.com/science/article/pii/S0065266015000516.

[125] Patrick Suppes. *A probabilistic theory of causality.* North-Holland Pub. Co., 1970.

[126] Ewa Szczurek, Niko Beerenwinkel, Pejman Mohammadi, Simona Constantinescu, and Jörg Rahnenführer. TiMEx: a waiting time model for mutually exclusive cancer alterations. *Bioinformatics*, 32(7):968–975, 07 2015. ISSN 1367-4803. doi: 10.1093/bioinformatics/btv400. URL https://doi.org/10.1093/bioinformatics/btv400.

[127] Damian Szklarczyk, Andrea Franceschini, Michael Kuhn, Milan Simonovic, Alexander Roth, Pablo Minguez, Tobias Doerks, Manuel Stark, Jean Muller, Peer Bork, et al. The string database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic acids research*, 39(suppl_1):D561–D568, 2010.

[128] Amos Tanay, Roded Sharan, and Ron Shamir. Discovering statistically significant biclusters in gene expression data. *Bioinformatics*, 18(suppl_1):S136–S144, 07 2002. ISSN 1367-4803. doi: 10.1093/bioinformatics/18.suppl_1.S136.

[129] Terry M Therneau. *A Package for Survival Analysis in S*, 2015. URL https://CRAN.R-project.org/package=survival. version 2.38.

[130] Erming Tian, Fenghuang Zhan, Ronald Walker, Erik Rasmussen, Yupo Ma, Bart Barlogie, and John D. Jr. Shaughnessy. The role of the Wnt-signaling antagonist DKK1 in the development of osteolytic lesions in multiple myeloma. *New England Journal of Medicine*, 349(26):2483–2494, 2003. doi: 10.1056/NEJMoa030847. URL http://dx.doi.org/10.1056/NEJMoa030847. PMID: 14695408.

[131] Ali Tofigh, Erik Sjölund, Mattias Höglund, and Jens Lagergren. A global structural em algorithm for a model of cancer progression. *Adv. Neural Inform. Process. Syst.*, 24:163–171, 2011.

[132] U.S. National Library of Medicine. What is a genome?, 2019. URL https://ghr.nlm.nih.gov/primer/hgp/genome.

[133] U.S. National Library of Medicine. What are whole exome sequencing and whole genome sequencing?, 2019. URL https://ghr.nlm.nih.gov/primer/testing/sequencing.

[134] Gilmer Valdes, JosÃĨ Marcio Luna, Eric Eaton, Charles B. Simone, Lyle H. Ungar, and Timothy D. Solberg. Mediboost: a patient stratification tool for interpretable decision making in the era of precision medicine. *Scientific Reports*, 6, 2016.

[135] Oron Vanunu, Oded Magger, Eytan Ruppin, Tomer Shlomi, and Roded Sharan. Associating genes and protein complexes with disease via network propagation. *PLoS computational biology*, 6(1):e1000641, 2010.

[136] J. Craig Venter, Mark D. Adams, Eugene W. Myers, Peter W. Li, Richard J. Mural, et al. The sequence of the human genome. *Science*, 291(5507):1304–1351, 2001. ISSN 0036-8075. doi: 10.1126/science.1058040. URL https://science.sciencemag.org/content/291/5507/1304.

[137] Roel G.W. Verhaak, Katherine A. Hoadley, Elizabeth Purdom, Victoria Wang, Yuan Qi, Matthew D. Wilkerson, C. Ryan Miller, Li Ding, Todd Golub, Jill P. Mesirov, Gabriele Alexe, Michael Lawrence, Michael O'Kelly, Pablo Tamayo, Barbara A. Weir, Stacey Gabriel, Wendy Winckler, Supriya Gupta, Lakshmi Jakkula, Heidi S. Feiler, J. Graeme Hodgson, C. David James, Jann N. Sarkaria, Cameron Brennan, Ari Kahn, Paul T. Spellman, Richard K. Wilson, Terence P. Speed, Joe W. Gray, Matthew Meyerson, Gad Getz, Charles M. Perou, and D. Neil Hayes. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in pdgfra, idh1, egfr, and nf1. *Cancer Cell*, 17(1):98 – 110, 2010. ISSN 1535-6108. doi: https://doi.org/10.1016/j.ccr.2009.12.020. URL http://www.sciencedirect.com/science/article/pii/S1535610809004322.

[138] Guoli Wang, Andrew V. Kossenkov, and Michael F. Ochs. LS-NMF: A modified non-negative matrix factorization algorithm utilizing uncertainty estimates. *BMC Bioinformatics*, 7(1):175, Mar 2006. ISSN 1471-2105. doi: 10.1186/1471-2105-7-175. URL https://doi.org/10.1186/1471-2105-7-175.

[139] January Weiner. *tmod: Feature Set Enrichment Analysis for Metabolomics and Transcriptomics*, 2016. URL `https://CRAN.R-project.org/package=tmod`. R package version 0.31.

[140] Joanna C. D. Willis and Graham M. Lord. Immune biomarkers: the promises and pitfalls of personalized medicine. *Nature Reviews Immunology*, 15:323, Mar 2015. URL `https://doi.org/10.1038/nri3820`. Perspective.

[141] Yanxun Xu, Peter MÃijller, Yuan Yuan, Kamalakar Gulukota, and Yuan Ji. MAD Bayes for tumor heterogeneity-feature allocation with exponential family sampling. *Journal of the American Statistical Association*, 110(510):503–514, 2015. doi: 10.1080/01621459.2014.995794. URL `https://doi.org/10.1080/01621459.2014.995794`.

[142] Chen-Hsiang Yeang, Frank McCormick, and Arnold Levine. Combinatorial patterns of somatic gene mutations in cancer. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology*, 22:2605–22, 05 2008.

[143] K. Y. Yip, D. W. Cheung, and M. K. Ng. Harp: a practical projected clustering algorithm. *IEEE Transactions on Knowledge and Data Engineering*, 16(11):1387–1397, Nov 2004. ISSN 1041-4347. doi: 10.1109/TKDE.2004.74.

[144] Zhenhua Yu, Ao Li, and Minghui Wang. CLImAT-HET: detecting subclonal copy number alterations and loss of heterozygosity in heterogeneous tumor samples from whole-genome sequencing data. *BMC Medical Genomics*, 10(1):15, Mar 2017. ISSN 1755-8794. doi: 10.1186/s12920-017-0255-4. URL `https://doi.org/10.1186/s12920-017-0255-4`.

[145] Hamim Zafar, Anthony Tzen, Nicholas Navin, Ken Chen, and Luay Nakhleh. SiFit: inferring tumor trees from single-cell sequencing data under finite-sites models. *Genome biology*, 18(1):178, 2017.

[146] Shihua Zhang, Chun-Chi Liu, Wenyuan Li, Hui Shen, Peter W. Laird, and Xianghong Jasmine Zhou. Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. *Nucleic Acids Research*, 40(19):9379–9391, 08 2012. ISSN 0305-1048. doi: 10.1093/nar/gks725. URL `https://doi.org/10.1093/nar/gks725`.

[147] Ying Zhang, Tyler Pizzute, and Ming Pei. A review of crosstalk between MAPK and Wnt signals and its impact on cartilage regeneration. *Cell and Tissue Research*, 358(3):633–649, Dec 2014. ISSN 1432-0878. doi: 10.1007/s00441-014-2010-x. URL `https://doi.org/10.1007/s00441-014-2010-x`.

Appendix A

Experimental Results Corresponding to Chapter 4

A.1 Results for OR Simulations

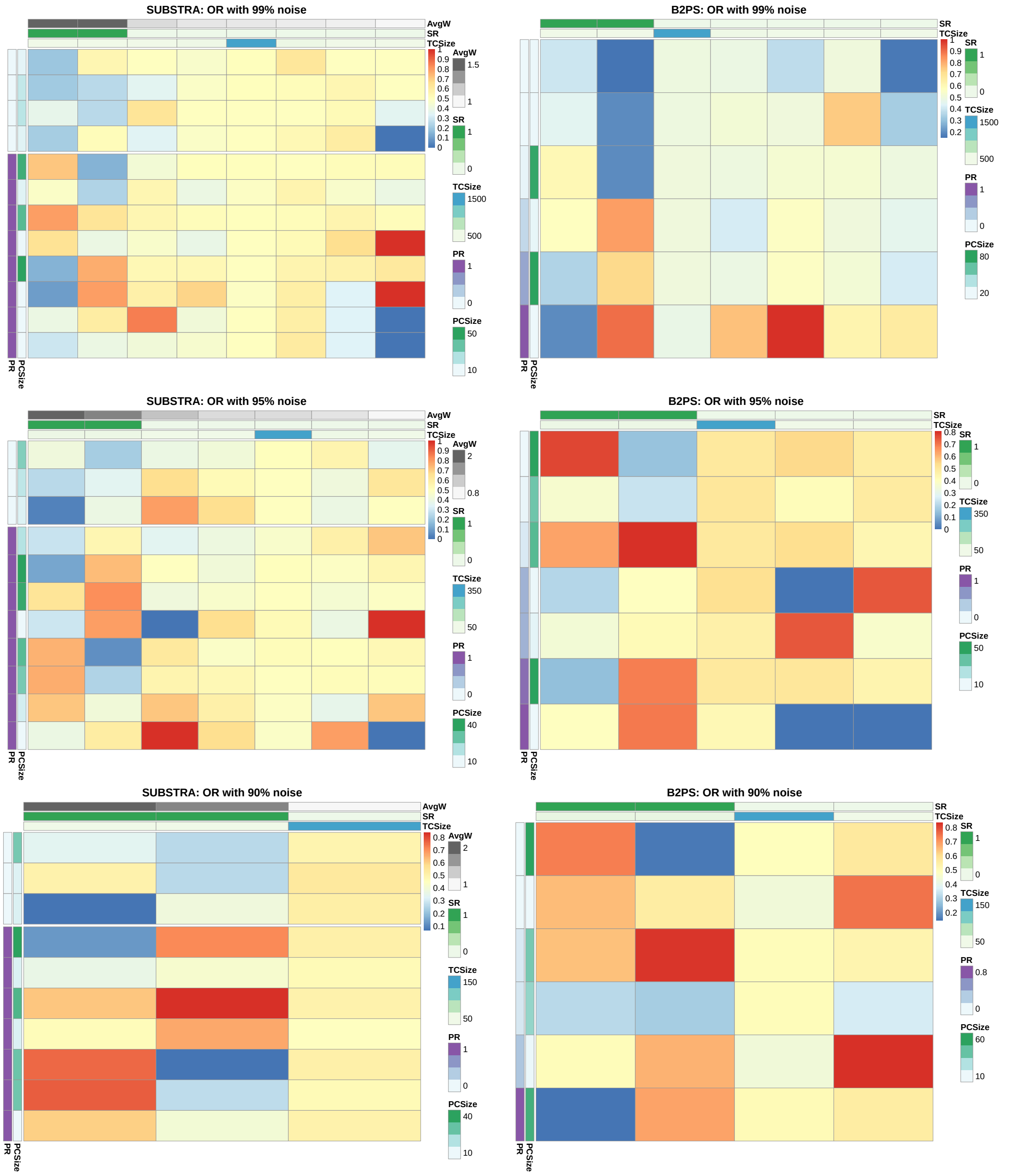


Figure A.1: Resulted heatmaps of SUBSTRA and B2PS for synthetic data for the OR relationship

A.2 Results for XOR Simulations

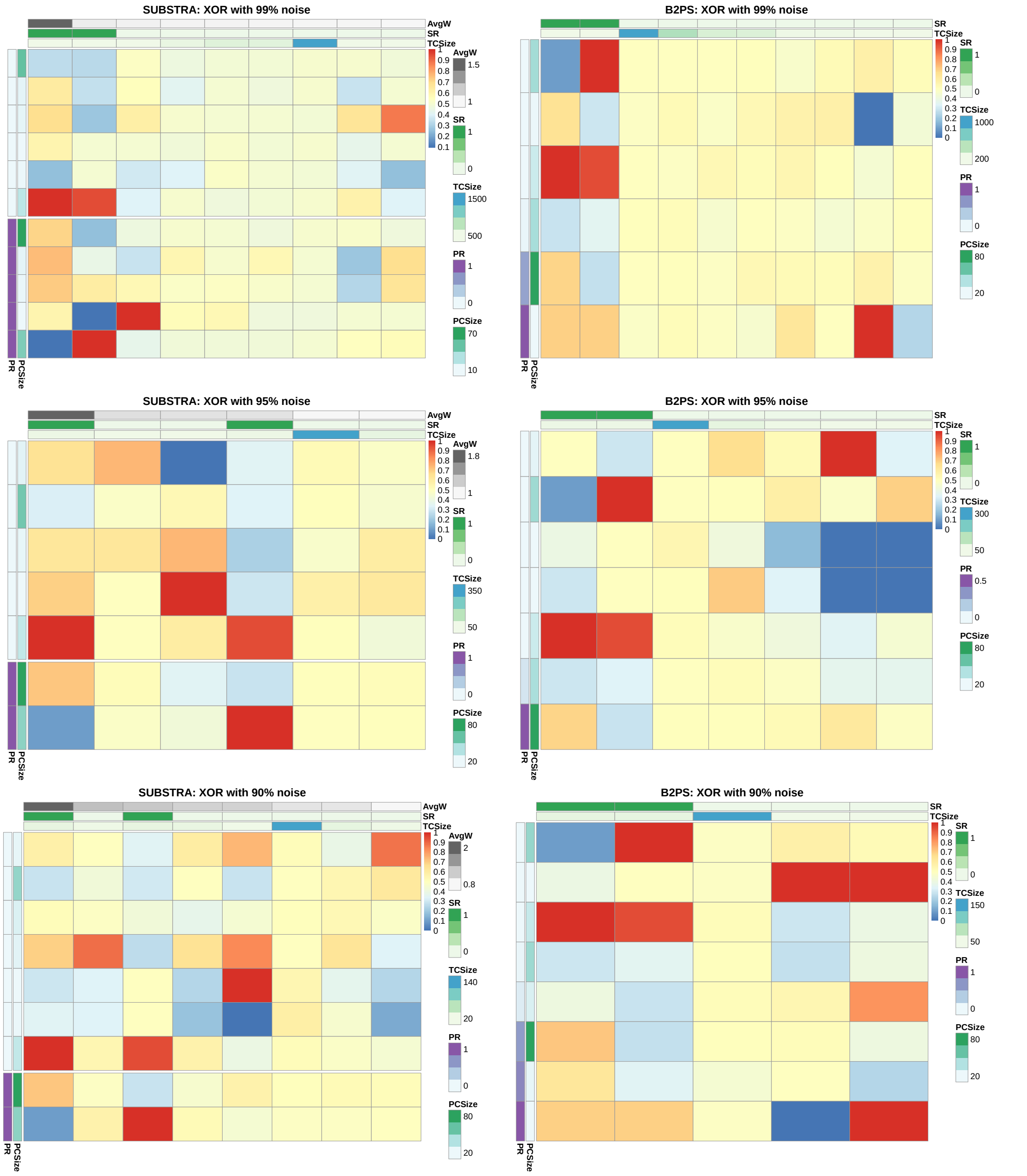


Figure A.2: Resulted heatmaps of SUBSTRA and B2PS for synthetic data for the XOR relationship

A.3 Results for Kidney 1 Dataset

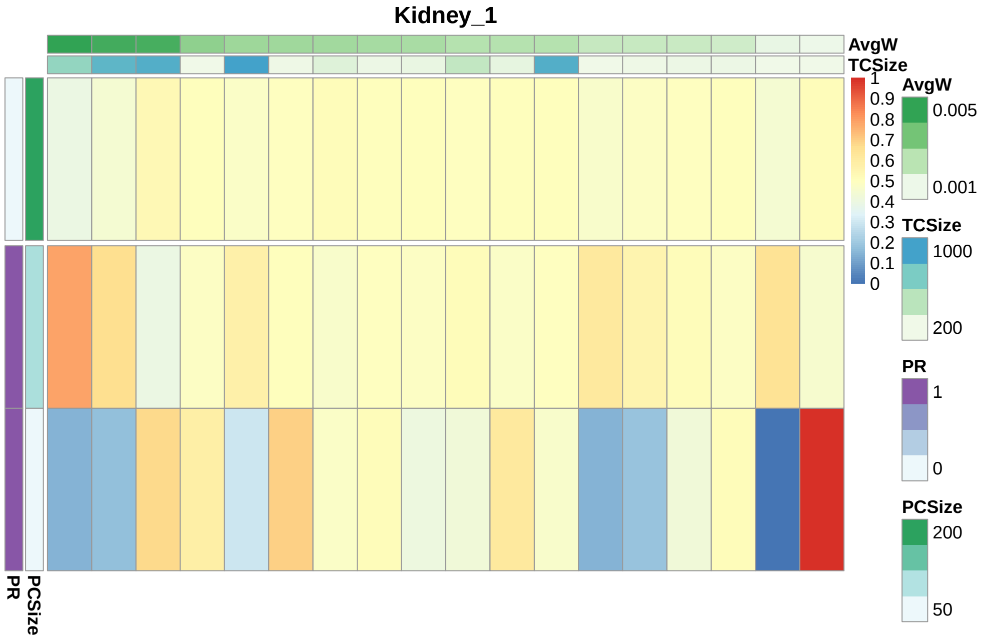


Figure A.3: Heatmap for Kidney 1 Dataset

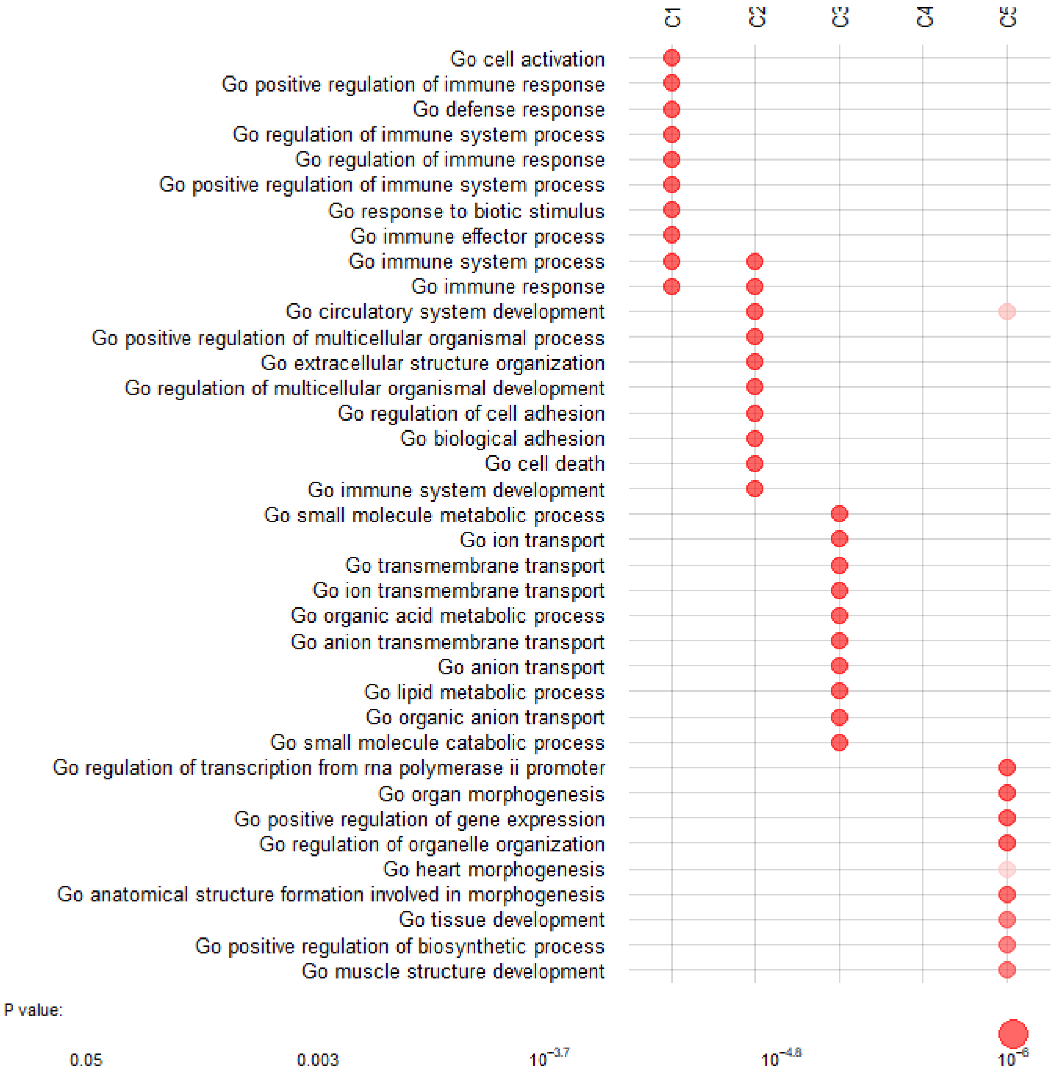


Figure A.4: GO Enrichment for Kidney 1 Dataset

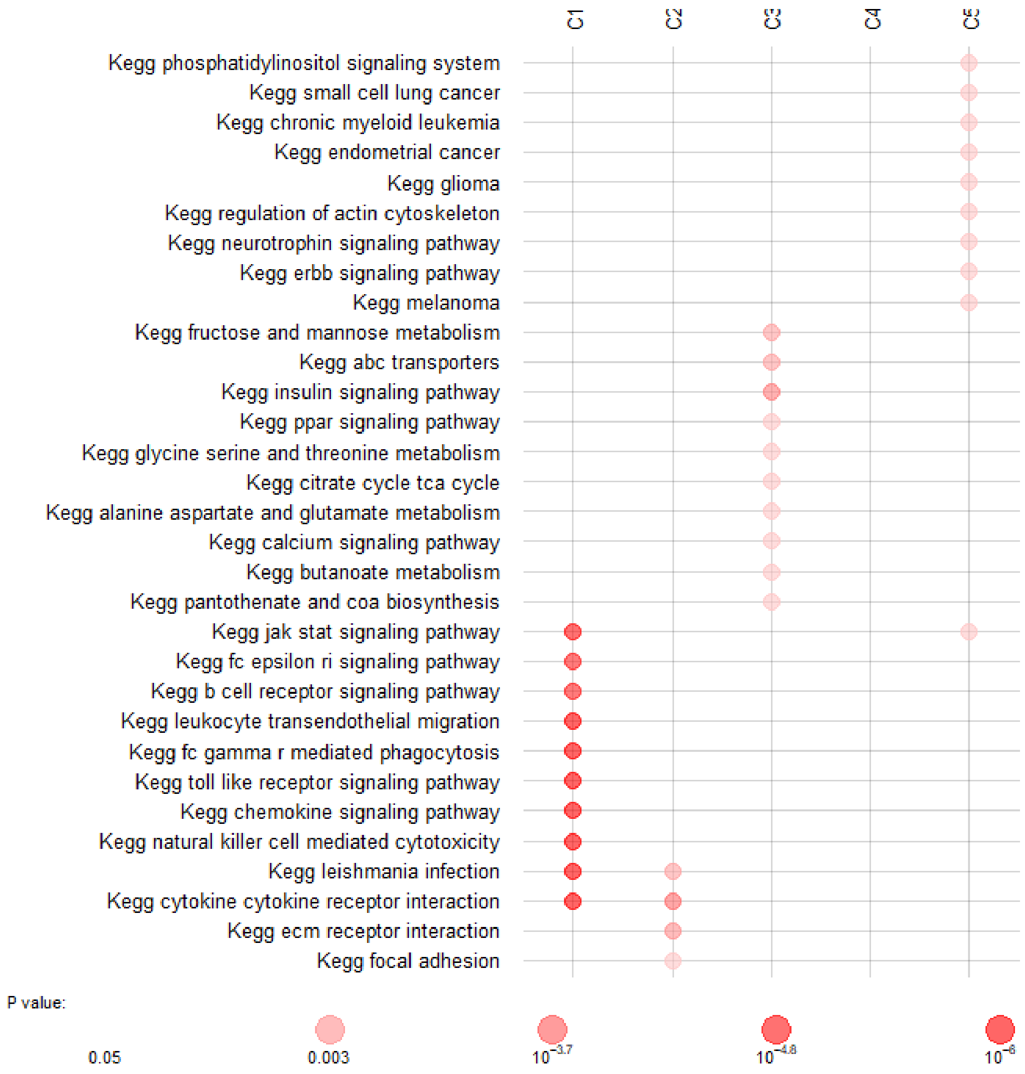


Figure A.5: Pathway Enrichment for Kidney 1 Dataset

A.4 Results for Drug Response Dataset

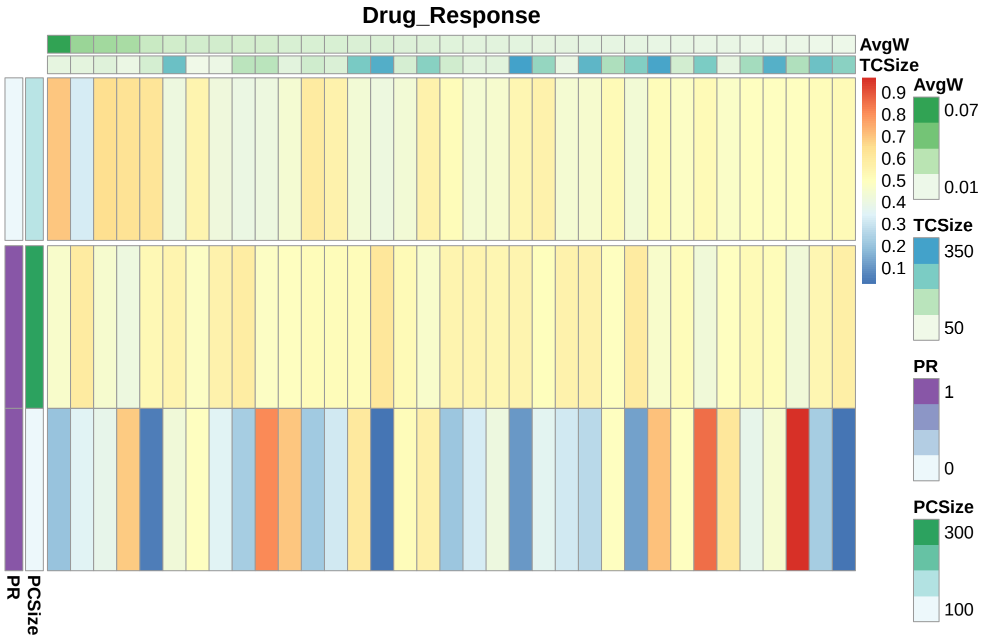


Figure A.6: Heatmap for Drug Response Dataset

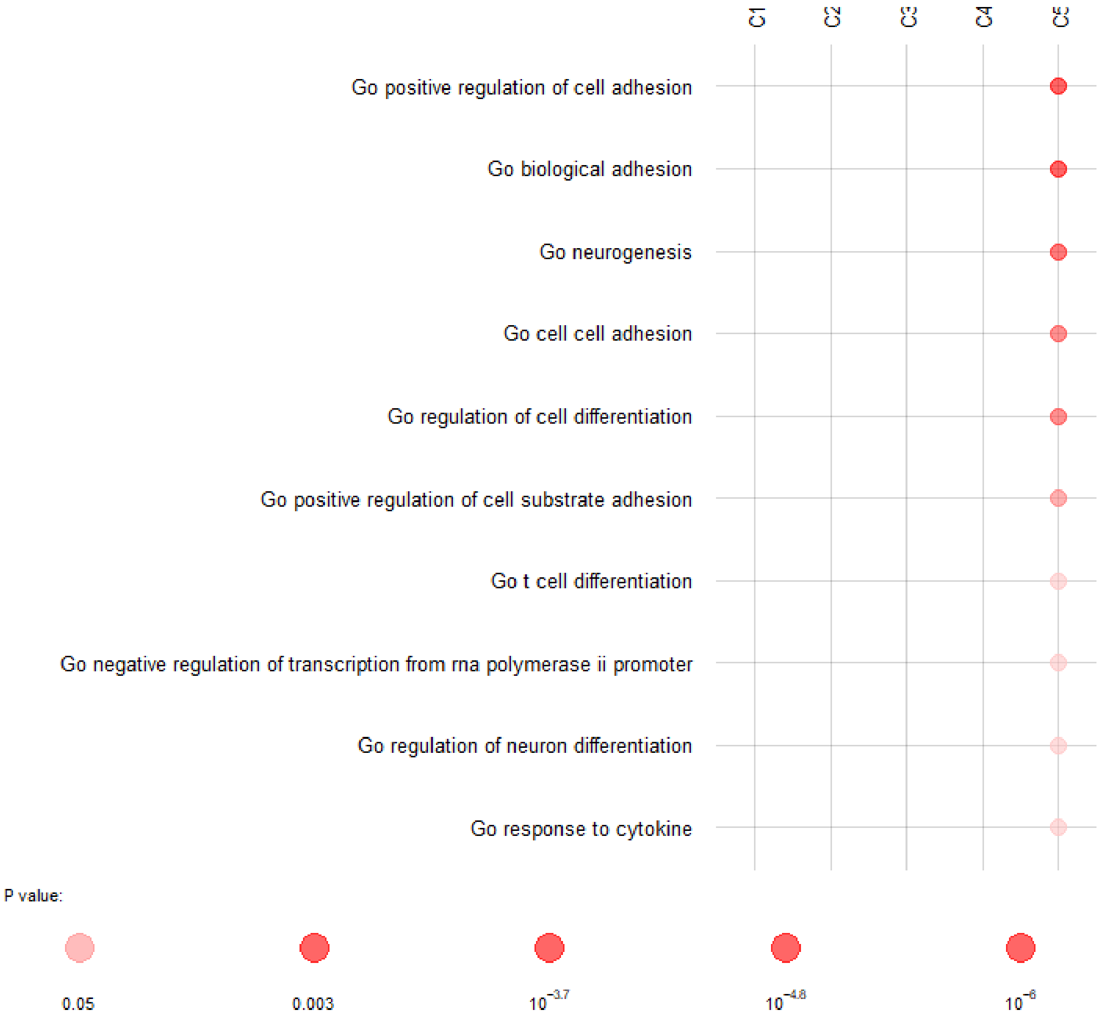


Figure A.7: GO Enrichment for Drug Response Dataset

A.5 Results for Multiple Myeloma Dataset

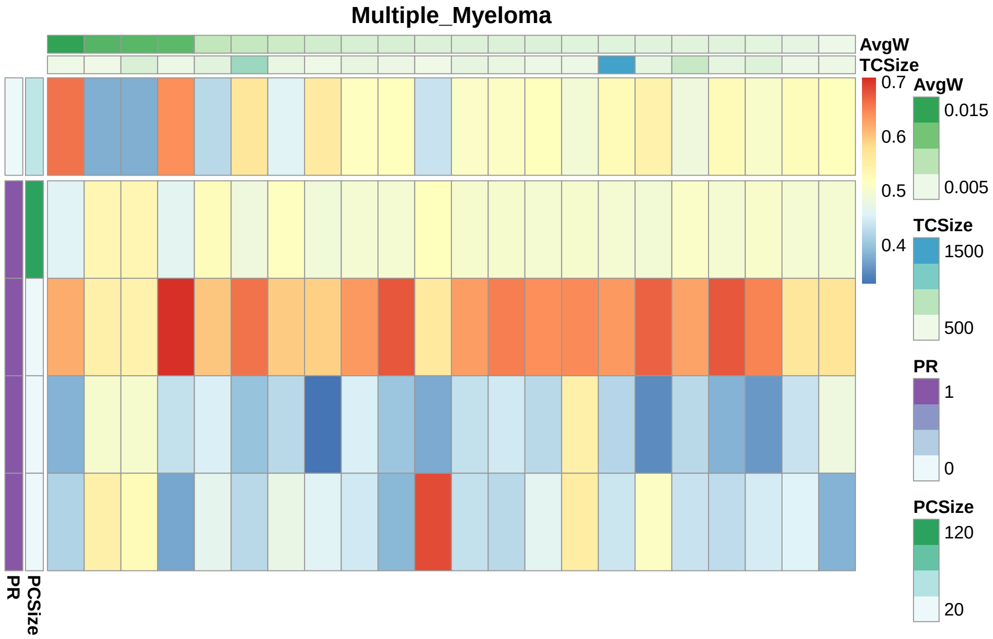


Figure A.8: Heatmap for Multiple Myeloma Dataset

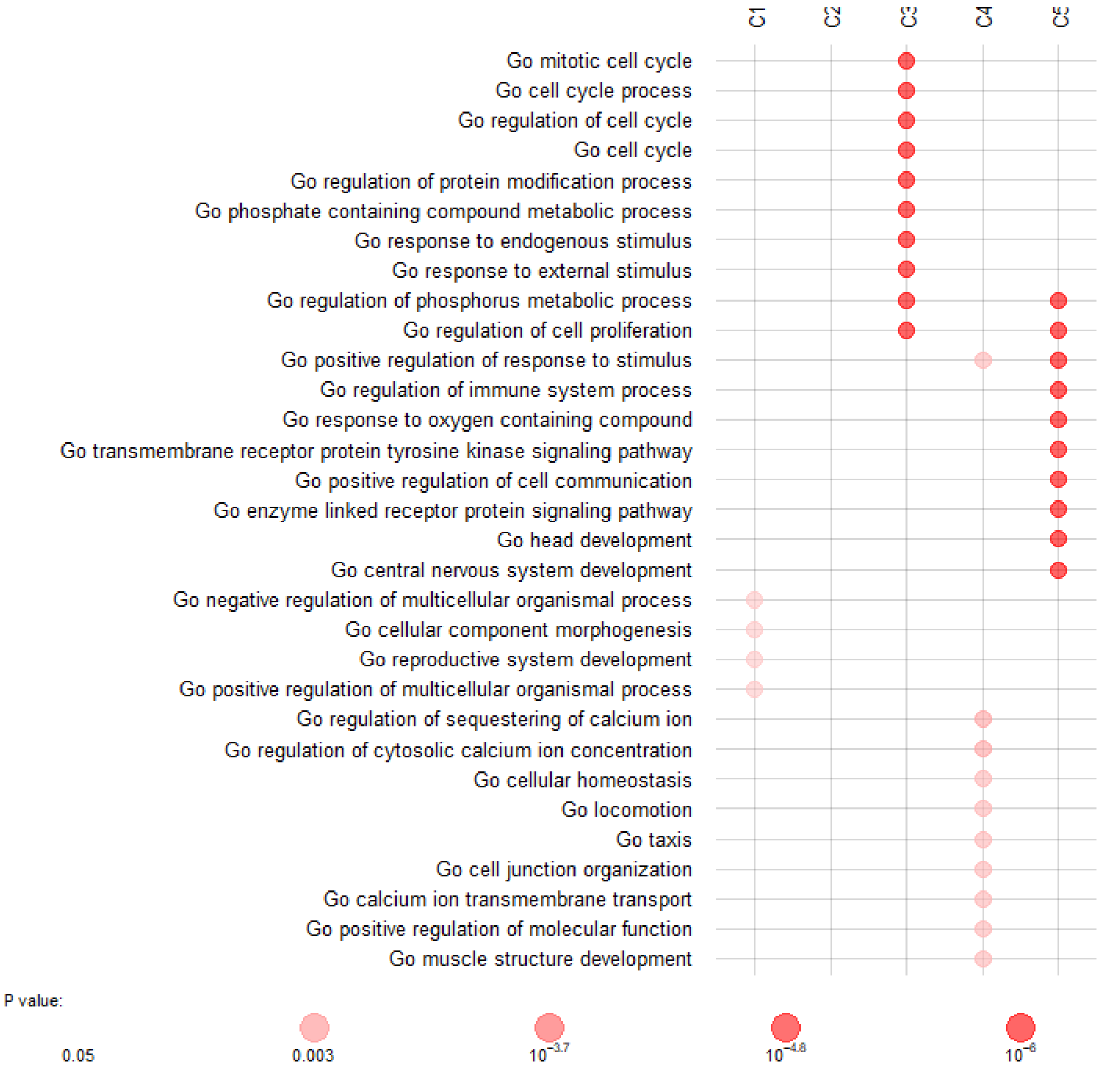


Figure A.9: GO Enrichment for Multiple Myeloma Dataset

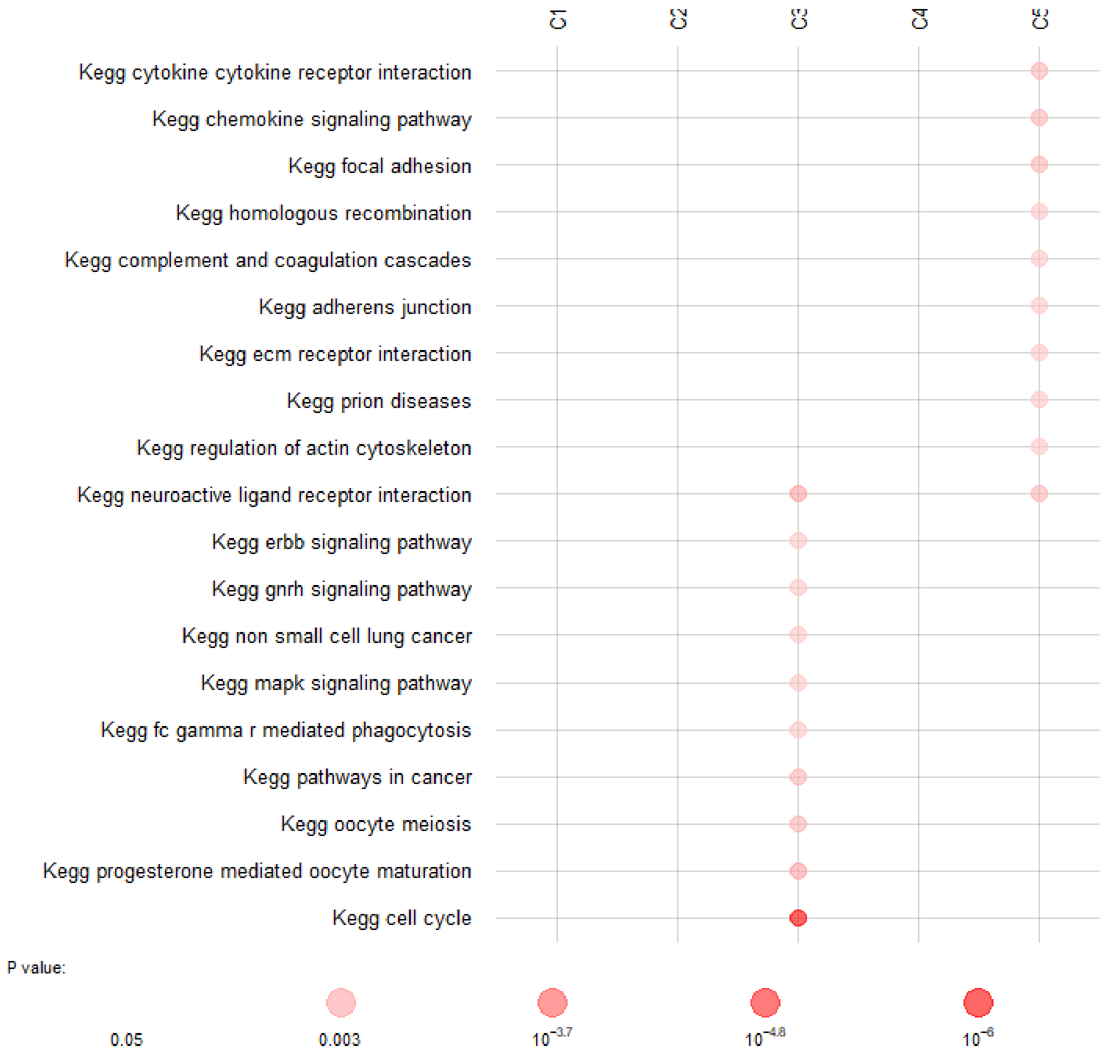


Figure A.10: Pathway Enrichment for Multiple Myeloma Dataset

A.6 Results for Lung Cancer Dataset

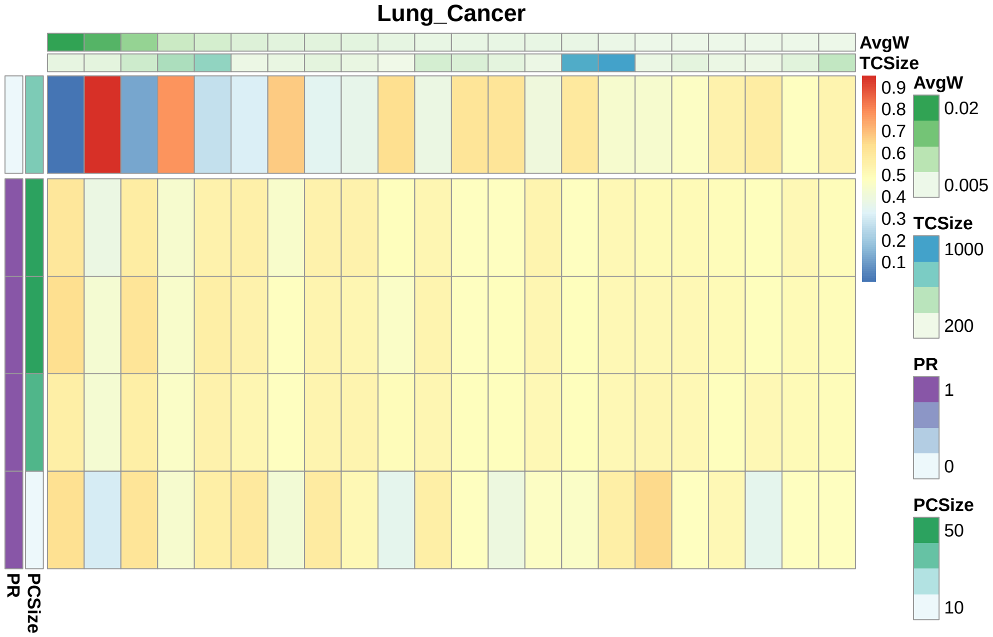


Figure A.11: Heatmap for Lung Cancer Dataset

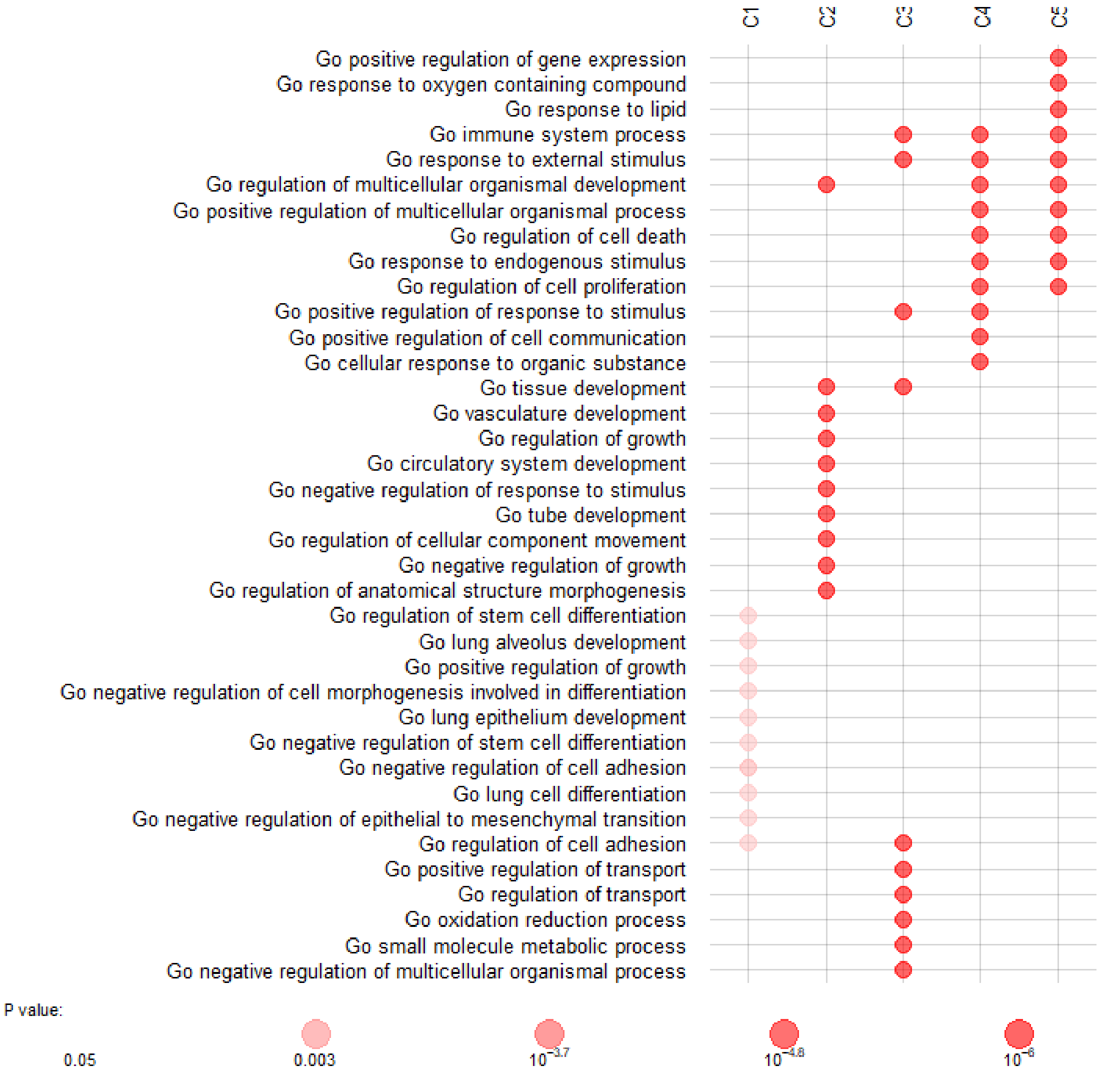


Figure A.12: GO Enrichment for Lung Cancer Dataset

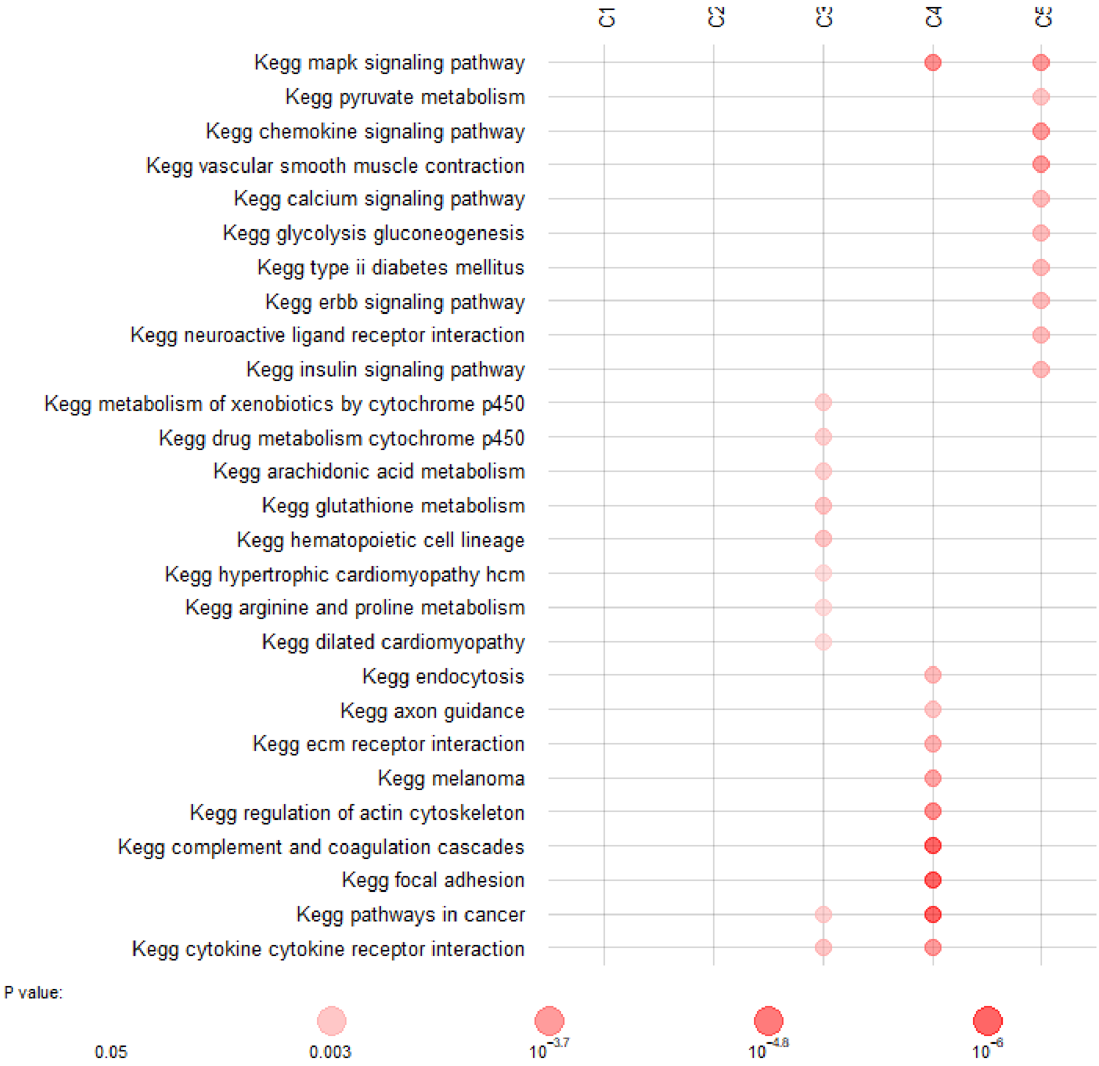


Figure A.13: Pathway Enrichment for Lung Cancer Dataset