Statistical Methods for Integrated Cancer Genomic Data Using a Joint Latent Variable Model

Esther Drill

Submitted in partial fulfillment of the
requirements for the degree
of Doctor of Public Health
in the Department of Biostatistics
at the Mailman School of Public Health

COLUMBIA UNIVERSITY

2018

ABSTRACT

Statistical Methods for Integrated Cancer Genomic Data Using a Joint Latent Variable Model

Esther Drill

Inspired by the TCGA (The Cancer Genome Atlas), we explore multimodal genomic datasets with integrative methods using a joint latent variable approach. We use *iCluster+*, an existing clustering method for integrative data, to identify potential subtypes within TCGA sarcoma and mesothelioma tumors, and across a large cohort of 33 different TCGA cancer datasets. For classification, motivated to improve the prediction of platinum resistance in high grade serous ovarian cancer (HGSOC) treatment, we propose novel integrative methods, *iClassify* to perform classification using a joint latent variable model. *iClassify* provides effective data integration and classification while handling heterogeneous data types, while providing a natural framework to incorporate covariate risk factors and examine genomic driver by covariate risk factor interaction. Feature selection is performed through a thresholding parameter that combines both latent variable and feature coefficients. We demonstrate increased accuracy in classification over methods that assume homogeneous data type, such as linear discriminant analysis and penalized logistic regression, and improved feature selection. We apply *i*Classify to a TCGA cohort of HGSOC patients with three types of genomic data and platinum response data. This methodology has broad applications beyond predicting treatment outcomes and disease progression in cancer, including predicting prognosis and diagnosis in other diseases with major public health implications.

Table of Contents

List of Figures

List of Tables

Acknowledgments

First and foremost, my most sincere thanks to my dissertation advisors, Dr. Yuanjia Wang and Dr. Ronglai Shen, for their guidance and assistance over the last several years. They have been understanding about my unique path and circumstances, and have been steadfastly patient, insightful, helpful, and wise.

And a huge and heartfelt thank you to the chair of my dissertation committee, Dr. Shuang Wang, and the other committee members, Dr. Iuliana Ionita-Laza and Dr. Jeanine Genkinger. This dissertation has been greatly improved as a result of their comments, questions and suggestions.

Also, a special thank you to Dr. Katherine Panageas at Memorial Sloan Kettering, who offered me a great job five years ago, and has been consistently supportive of my efforts to finish up my degree.

I couldn't have finished without the support of those closest to me. For logistical support, huge thanks to my stellar mother-in-law, Sharon Messitte, who was always ready to help me find some time to work. For emotional support, I thank my mother Ruth Ignatoff, ever interested and willing to listen. For support of all kinds, and often on a daily basis, I thank the universe for Dr. Rebecca Drill, truly one of the great sisters to walk the earth.

I am grateful to my inner circle of friends and family (looking at you, Dad!) for their unstinting encouragement and belief in my ability to "get it done" even when I was not so sure.

My deepest gratitude goes to Paul Greenberg, who has had to face the brunt of it all, and never wavered, and always understood. Last and littlest, but not least, a special thanks to my son Luke Greenberg, who has primarily seen this dissertation as an inconvenience and distraction, and in that has forced me to always keep perspective.

June, 2018

Chapter 1

Introduction

1.1 Overview

The TCGA (The Cancer Genome Atlas) project has made widely available for the first time multiple modes of genomic data from the same large number of samples. This has spurred the development of integrative methods that attempt to improve the power and efficiency of both clustering and classification methods by integrating data from multiple platforms into a unified analysis. Here, we focus on integrative methods using a joint latent variable approach and apply them to TCGA datasets.

We use *iCluster+*, an existing clustering method for integrative data using a joint latent variable model, to identify potential subtypes **within** TCGA sarcoma and mesothelioma tumors, and **across** a large cohort of 33 different TCGA cancer datasets.

For classification, we propose novel integrative methods to leverage data across multiple genomic platforms to perform classification using a joint latent variable model. This approach provides effective dimension reduction while handling heterogeneous, diverse data types of different scale and variance structure. It also provides a natural framework to incorporate important clinical or environmental covariates, and importantly to investigate

interactions with these covariates. Effective feature selection is performed through a thresholding parameter that combines both effects from latent variables and observed feature variables. We term this new methodology *iClassify*.

We demonstrate that this method leads to increased accuracy in prediction over methods that assume homogeneous data type. Moreover, we show a marked improvement in feature selection over commonly used methods. We then apply *iClassify* to the classification problem of predicting response to platinum chemotherapy in a TCGA ovarian cancer cohort.

We organize the material as follows. In Section 1.2 we provide an overview of the mutimodal genomic TCGA project, focusing on key questions about cancer subtyping and classification, then review current methodological approaches to integrative genomic analysis both in terms of clustering (Section 1.3.2) and classification (Section 1.2.2). In Chapter 2 we present the iCluster method and our results from integrative clustering in three published TCGA studies on sarcoma, mesothelioma, and a pancancer cohort of 33 cancer types. In Chapter 3, we introduce the methodological framework of *iClassify* , present and simulation studies demonstrating its capabilities in Chapter 4. In Chapter 5 we apply *iClassify* to the TCGA ovarian cancer dataset and present results of genomic-only classification and genomic-covariate interaction analysis. We conclude the paper with discussion and future extensions in Chapter 6.

1.2 Introduction to Motivating Studies

The Cancer Genome Atlas is providing researchers with unprecedented richness of genomic data, with each tumor being sequened for mutation data, genotyped for copy number data, and assayed for mRNA, noncoding RNA, and DNA methylation profiling. Integrated analyses of these data could yield important contributions to knowledge about cancer mechanisms

as well as improve prediction of drug response and overall prognosis. Integrative genomics starts conceptually from the idea that biological mechanisms are comprised of multiple molecular layers, and that understanding each of these layers will inform a more comprehensive understanding of mechanisms that lead to cancer. In fact, one of the key questions posed by the National Cancer Institute in relation to TCGA data is: "How can investigators effectively integrate data from multiple modes of genomic analysis into a unified view of oncogenic pathways?" ([National Cancer Institute, 2018]).

In particular, we are interested in two objectives using integrative genomic analysis. The first is finding individuals or samples that have similar mechanisms of disease within a particular cancer type, as these "subtypes" of cancer may have differential prognoses or responses to treatment. Moreover, we are interested in whether tumors may have similar mechanisms of disease across cancer types, so that an effective treatment for a subtype in one cancer may actually have relevance for a similar subtype in a different cancer. Additionally, we wish to leverage the integrative genomic information available to be able to perform better classification of tumors for new patients either by outcome or response to treatment.

1.2.1 Molecular subtyping in cancer

Cancer, even within a particular type, is an exceedingly heterogenous disease, with a myriad of driver mutations, chromosomal alterations and key pathway disruptions. Thus, subtyping cancer is widely understood to be essential to an improved and more personalized prognosis/treatment.

NCI defines a cancer subtype as a "smaller group that a type of cancer can be divided into, based on certain characteristics of the cancer cells." [National Cancer Institute, 2018]. A molecular subtype is specifically a group of samples that have a similar molecular mecha-

nism as the origin of the carcinogenesis [Le Van *et al.*, 2016]. These molecular mechanisms can be subtype-specific mutations or copy number alterations or expression features which may point to "disease-perturbed networks" [Hood and Friend, 2011] that may provide important new drug targets.

Prior to the genomic era, subtyping had mainly been accomplished by classic immunohistologic technique. For example, breast cancer was stratified into three subtypes using traditional immunohistochemistry techniques: hormone-receptor-positive, triple negative, and HER2-positive, each of which has it own treatment approach. With the availability of genome-wide expression profiles, researchers used hierarchical clustering to describe additional breast cancer subtypes, including luminal A, luminal B, lumnal C, HER2-enriched, basal-like, claudin-low, and normal breast-like [Glueck *et al.*, 2013]. There are now genomic assays meant to categorize breast cancer tumors into one of these subtypes, and in so doing provide better recurrence risk and prognosis estimates.

Other examples of successful, replicated subtyping performed by clustering of gene expression profiles are in Diffuse large B-cell lymphoma (DLBCL) with "oxidative phosphorylation," "B- cell receptor/proliferation," and "host response" subtypes [Monti *et al.*, 2005]; and glioblastoma with proneural, neural, classical and mesenchymal subtypes [Verhaak *et al.*, 2010].

As the multiple genomic data platforms of TCGA have become available, one of the major challenges in cancer research has been to use these integrative data to provide fuller insight into identifying clinically meaningful cancer subtypes, in the continued hopes of finding new stratified, effective treatments. Many of the clustering approaches commonly used are only capable of dealing with single data types at a time and results are often integrated manually. However, post hoc integration of results from individual genomic data

sets will likely not be able to capture multiple relationships that exist between different levels of the data, and thus may fail to realize the potential inherent in the multi-modal data.

It is in this context that we became involved in the TCGA studies of sarcoma and mesothelioma tumors. While sarcoma has many histologically diverse malignancies, the question of subtypes within these diverse sarcoma types had never been systematically looked at. And in mesothelioma, we were interested in whether there were molecular subtypes that would be histology-independent.

Additionally, an intermediate analysis of 12 TCGA cancer types and 3,527 tumors from 2014 had presented some results that suggested that molecular subtypes might provide an alternative to current organ- and tissue-histology-based classification, and had estimated that "at least one in ten cancer patients might be classified (and perhaps treated) differently using such a molecular taxonomy, rather than the current histopathology-based classification" [Hoadley *et al.*, 2014]. The question of interest in this larger pan-cancer cluster analysis of all 33 tumor types in TCGA was whether we would find further "convergent integrated molecular subypes."

Addressing these questions

1.2.2 Classification in cancer: response to platinum chemotherapy

In a natural followup to molecular subtyping based on integrative data, The Center for Cancer Genomics (CCG), which is the successor to TCGA, is interested in finding classifiers through integrative genomic classification that can facilitate prediction of newly-discovered clinically meaningful subtypes.

And, in general, the natural corollary to the challenge of this molecular subtyping is the

challenge of integrative genomic classification: How can we use the multi-modes of genomic data available in TCGA to find models that effectively predict outcomes or response?

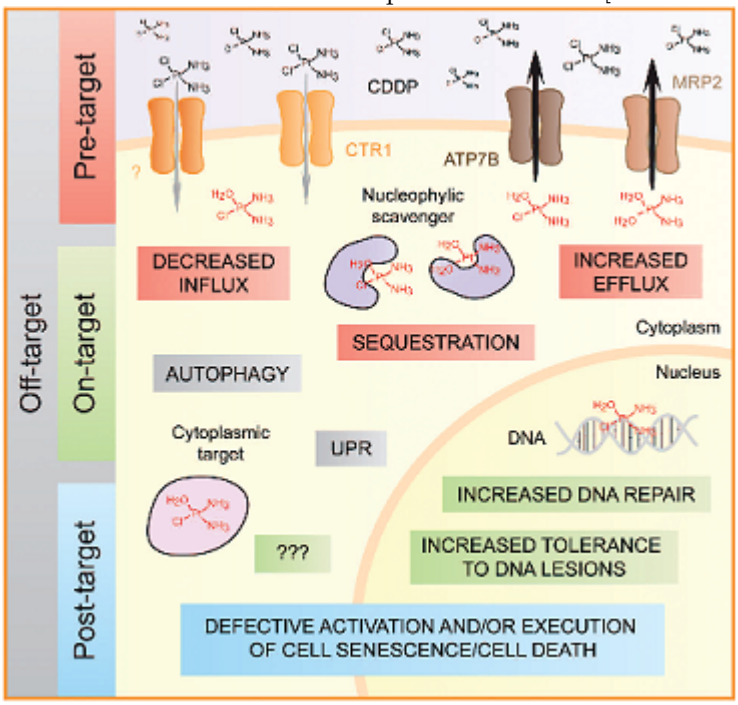
In this area, we were motivated by one problem in particular: current investigations into platinum resistance in high grade serous ovarian cancer (HGSOC), the most agressive subtype of ovarian cancer which makes up most of the advanced stage cases of epithelial ovarian cancer. HGSOC has the highest mortality rate among all gynecological cancers and its 5-year survival rate of 35-40% has shown little improvement for decades [Kamieniak *et al.*, 2015].

This is likely due to the fact that resistance to standard-of-care platinum chemotherapy eventually emerges in 70% of patients, who will go on to develop recurrent cancer. Platinum resistance is formally defined as a tumor progression within six months after completion of first-line platinum therapy. Despite the high percentage of platinum-resistant patients, the treatment remains the standard of care [Matsuo *et al.*, 2010]. Being able to predict those patients who are at high risk for platinum resistance could have important consequences for treatment, including an increased monitoring protocol as well as a complete change in treatment approach [Gonzalez Bosquet *et al.*, 2016].

1.2.2.1 Molecular mechanisms

Platinum resistance is known to be multifactorial, relying on the "activation of multiple, non-redundant molecular or cell circuitries". In a couple of papers, [Galluzzi *et al.*, 2012] and [Galluzzi *et al.*, 2012] classify these molecular mechanisms into four types: (1) pre-target resistance, which interferes with the binding of cisplatin to DNA; (2) on-target resistance, which disrupts DNA-cisplatin binding; (3) post-target resistance, which perturbs Cisplatin-initiated signaling pathways mean to lead to cancer cell death; and (4) off-target resistance,

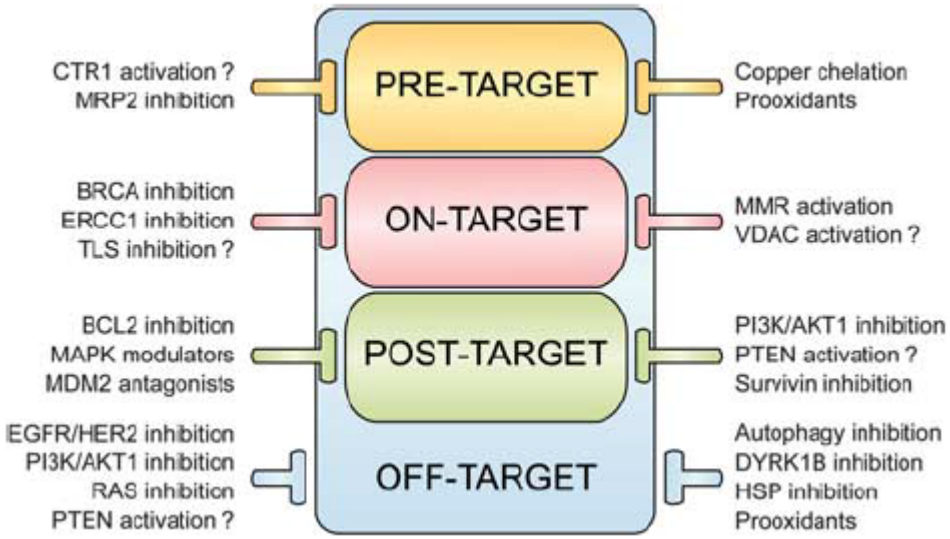
Figure 1.1: Molecular mechanisms of cisplatin resistance [Galluzzi *et al.*, 2014].



with no obvious links to mechanisms of cisplatin-mediated damage. The authors synthesize genomic and functional studies done in the last decade that have contributed to this knowledge and summarize findings that detail specific genes and pathways in each mechanism type in Figures 1.1 and 1.2.

Our substantial increase in understanding, however, has so far not led to an ability to be able to identify ovarian tumors that are likely to be platinum-resistant, although there is a growing body of work attempting to do just that. In a recent systematic review of prediction of resistance to chemotherapy in ovarian cancer, [Lloyd *et al.*, 2015] concluded that "A clinically applicable gene signature capable of predicting patient response to chemotherapy has not yet been identified". Of the 1298 genes that were identified by 32 prognostic/predictive models reviewed, 1214 (94%) were found by only one study. And the gene most frequently selected was only selected by 4 of the models. The vast majority of these studies have

Figure 1.2: Strategies for reverting cisplatin resistance [Galluzzi *et al.*, 2012].



looked at one genomic platform at a time, in most cases gene expression.

Because these efforts have not produced persuasive and replicable models, and because of the multifactorial nature of platinum resistance, there has been increasing interest in using additional and multiple genomic modalities. To this point, the Ovarian Cancer Action (OCA) 2015 meeting designated "Understanding drug response" as one of seven key areas for future research, and the emphasis was on moving to an "integrated view" that brings together all genomics data on individual samples [Bowtell *et al.*, 2015].

1.2.2.2 Important covariates

Additionally, there are some well known covariates that are associated with platinum resistance and HGSOC prognosis. For instance, residual disease is known to be one of the most influential factors in HGSOC prognosis. In the TCGA dataset of HGSOC tumors surgically resected before treatment with platinum chemotherapy, [Tucker *et al.*, 2014] reported that survival was significantly better for patients with no residual disease compared to those who had any residual disease at all (including <10 mm, which previously had been considered

to be optimal).

Recently, there has been increasing interest in and accumulating evidence for the idea that the tumor microenvironment has a large role to play in platinum resistance, particularly characteristics of the extracellular matrix that may preferentially support cancer stem cells or residual disease cells [Chien *et al.*, 2013].

There is also evidence that certain genetic mutations play a role in platinum resistance. For instance, TP53 mutations have been reported in the literature to be associated with platinum resistance in some studies but not in others ([Agarwal and Kaye, 2003]).

The evidence for BRCA mutations is much more consistent. Mutations in BRCA1 and BRCA2 (breast cancer susceptibility gene types 1 and 2) are known to significantly increase the chances of a person developing ovarian cancer. (BRCA1 confers a 39-40% and BRCA2 an 11-18% lifetime risk.) Because of BRCA's role in the DNA repair pathway, however, tumors with BRCA mutations have proven to be more sensitive to platinum drugs than tumors with wildtype BRCA. This leads to consistently better prognosis for patients with BRCA1/2-mutated ovarian cancer compared with non-carriers, if they receive platinum-based therapy ([Mylavarapu *et al.*, 2018], [Swisher *et al.*, 2008]).

Interestingly, this sensitivity can be modulated by other genomic events/factors. For example, [Norquist *et al.*, 2011] reported that 46.2% of platinum resistant tumors have secondary mutations that restored the function of BRCA1/2 as compared with 5.3% that are sensitive to platinum. Thus, researchers believe that secondary mutations in BRCA1/2 may be overriding the BRCA1/2 tumors' sensitivity to platinum, and perhaps there are other genomic factors as well that could be modulating the BRCA1/2 tumors' response to platinum therapy.

While it is impractical to test interaction between every genomic feature and the BRCA1/2

genotype, a model that could make a genomic features x BRCA1/2 genotype possible would certainly be of interest.

We propose our joint latent variable method iClassify to address the problem of predicting platinum response in high-grade serous ovarian cancer, and will use it to perform fully integrative genomic analyses that incorporate mRNA and miRNA expression, and methylation data, as well as interactions with clinically relevant mutations and residual disease.

1.3 Introduction to Statistical Methods

The TCGA (The Cancer Genome Atlas) project has made widely available for the first time multiple modes of genomic data from the same large number of samples. This has motivated the development of methods that attempt to improve the power of estimation and prediction of genomic effects on cancer outcomes by integrating data from these multiple platforms into a unified analysis. Methods development for this kind of approach fall roughly within a couple of categories that try to address two basic challenges: 1) how to reduce dimensionality of datasets where the number of variables is much greater than the number of samples ($p \gg n$, "the curse of high dimensionality") and 2) how to sensibly integrate such diverse data types.

1.3.1 Introduction to Integrative Clustering Methods

Clustering high-throughput multimodal genomic data is a challenge that has recently arisen as The Cancer Genome Atlas has accrued its rich, multiplatform cancer cohorts.

Early efforts to subtype TCGA cancer sets fell under the category of "clustering of clustering," where results of single platform clustering were then provided as input to a second-

level cluster analysis. The cluster-of-cluster assignments (COCA) method ([Hoadley *et al.*, 2014]) takes binary vectors of cluster assignments from each single platform as input and then performs Consensus clustering on those vectors. This allows for data to be combined without the challenge of non-trivial normalization. It also gives each single platform an influence on the final result proportional to how many clusters that platform produced. In this way, the number of features in a single platform does not carry undue weight. Another early method was SuperCluster ([Hoadley *et al.*, 2014]), an algorithm that took a similar approach to derive overall subtypes based on cluster memberships of single platform data types, but treats cluster assignments as nominal rather than binary, and assumes an equal contribution for each data type.

These methods can work well if most or all single platform clustering yields concordant results, but is less powerful otherwise. Also, assuming equal or near-equal contribution from all data types may not reflect true mechanisms of disease. More to the point, both of these methods perform integration of cluster membership values rather than at the data level, thereby losing valuable and essential information.

Pathway clustering is another common approach to integrative data subtyping. PARADIGM ([Vaske *et al.*, 2010]) is a pathway approach that has been used to cluster TCGA samples in a number of studies using copy number and mRNA expression data along with pathway interaction data found in public databases. The method infers integrated pathway levels (IPLs) for each gene that reflects a genes activity in a tumor sample relative to the median activity across all tumors, then runs consensus clustering on the most varying features. Other related approaches focus on integrating data into specific biological pathways, e.g. gene expression regulation ([Teo *et al.*, 2015], [Cheng *et al.*, 2015]) or drug pathways ([Li *et al.*, 2015a]).

These approaches can be very effective at answering the specific research questions they are designed for, but are inherently limited in a global sense by the focus on specific data types and networks. Also, as these methods depend, in large part, on prior biological knowledge, so can miss out on discovery of new mechanisms or previously unknown relationships.

Methods development for integrative clustering, spurred on by TCGA, has been a very active area for the last decade. There have been a number of"horizontal analyses" ([Tseng *et al.*, 2015]) approaches which focus on variable to variable relationships within and across data types . Examples of this are correlation network analysis ([Adourian *et al.*, 2008], [Li *et al.*, 2015b]), correlation motifs ([Ji *et al.*, 2015]) and multiple canonical correlation analysis ([Witten *et al.*, 2009]). Again, these methods are developed for specific questions and inherently do not focus on global modes of variation.

Another active area for clustering methods development has been in vertical integrative analyses, which focus on generalized dimensionality reduction ([Tseng *et al.*, 2015]). Common and early dimension reduction approaches such as singular value decomposition (SVD; [Alter *et al.*, 2000]; [Holter *et al.*, 2000] and non-negative matrix factorization (NMF; Brunet2004) work well for a single data type but do not accommodate multiple heterogenous data types. More recent developments specifically for multimodal data sets include Bayesian methods with penalization [Liu *et al.*, 2015], decomposition of variation ([Lock *et al.*, 2013]), and joint factor analysis ([Li and Jung, 2017], [Shen *et al.*, 2009]).

iCluster+ is specifically a joint latent variable model-based approach to integrative clustering with several advantages. In this formulation, the common set of latent variables are proposed to represent distinct driving factors of disease. Multiple genomic data types with different scale and variance structure can be incorporated into a single model. Biological relationships among platforms do not need to be specified. This acknowledges the incom-

plete nature of our understanding of genomic relationships, at the same time allowing for the discovery of novel pathways or drivers. A penalized likelihood method still allows for feature selection.

1.3.2 Introduction to Integrative Classification Methods

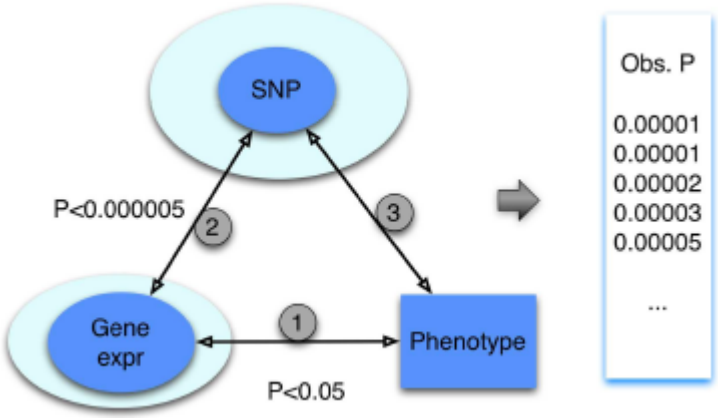
Methods for integrative classification are not as well developed as those for integrative clustering, and there is a real need for new approaches in this area.

The idea that looking at more than one type of genomic data for classification at once is more powerful is borne out by studies that have looked into the question. For example, [Fuchs *et al.*, 2013] found that in most cases, classifiers based on a combination of mRNA and miRNA data yield equivalent or greater accuracy than classifiers based on just one type of data. [Daemen *et al.*, 2009] similarly found that prediction of all outcomes improved when more than one type of genome-wide data set is used.

Sequential/filtering techniques consider different platforms one after the other, reducing the dimensionality of the feature space to those that have known relationships among at least two data types. For example, [Mankoo *et al.*, 2011] filtered out features that did not meet certain correlation criteria based on mRNA regulation in serous ovarian tumors, then used the greatly reduced set of features to predict survival.

This is the same logic used in studies of other complex diseases that seek to incorporate transcriptomic information into analyses to increase the power to uncover mechanisms of disease, spurred on by the demonstration by [Nicolae *et al.*, 2010] that single nucleotide polymorphisms (SNPs) discovered through GWAS are more likely than frequency-matched SNPs to be expression quantitative trait loci (eQTLs). In simple cases, eQTL studies have provided lists of single nucleotide variants that contribute to gene expression variation

Figure 1.3: The 'triangle" method for integration of genotype, expression, and phenotype data [Gamazon *et al.*, 2013].



(eSNPs) that can then be compared to significant disease-specific GWAS-associated SNPs [Cookson *et al.*, 2009]. This approach has proved successful in several studies, and is now routinely used to select candidate genes for exploring disease mechanisms [Montgomery and Dermitzakis, 2011]. For example, [Dubois *et al.*, 2010] highlighted 20 of the 38 associated celiac disease risk loci that are also correlated with expression variation in a nearby gene, and Anttila *et al* [Anttila *et al.*, 2010] used correlation of migraine-associated genotypes from GWAS with gene expression data to point to a potential regulatory mechanism.

More formally, genotype and gene expression data on the same samples can be integrated with phenotype in a "triangle" approach [Gamazon *et al.*, 2013], illustrated in Figure 1.3:

1. Identify a set of genes whose differential expression is associated with the phenotype with an arbitrarily chosen p-value, $p < p_{expression-phenotype}$.

2. Identify SNPs (eQTLs) that are associated with the selected gene in the previous step at an arbitrarily chosen p-value, $p < p_{expression-SNP}$.

3. SNPs from the previous step are then tested for association with phenotype at an

arbitrarily chosen p-value, $p < p_{SNP-phenotype}$.

Gamazon *et al* show that estimating the significance of these SNP-phenotype associations with a simple FDR leads to false positive results because the null distribution of p-values in this case is not uniform, and propose simulating a null distribution of p-values by:

1. Permuting phenotype labels Y_i, for each simulation i, while retaining the correlation structure of gene expression

2. For each permuted phenotype list, deriving a set of differentially expressed genes, g_i, that exceeds $p < p_{expression-phenotype}$

3. For each gene g_{ij}, retrieving set of eQTLs, S_{ijk} from the SCAN database that exceeds $p < p_{expression-SNP}$

4. Derive subset of S_{ijk} that exceed $p < p_{SNP-phenotype}$. These p-values are thus the null distribution.

Using the above null distribution of p-values, they follow the approach of [Storey and Tibshirani, 2003] to estimate an accurate FDR.

This method can incorporate other types of "omics" datasets (such as methylation, microRNA, and protein abundance, among others), and is theoretically not limited to using only two types of data at a time. The primary advantage of this approach is that each step of filtering accomplishes dimensionality reduction, which reduces the number of tests, potentially improving power. The performance of the method, however, is very dependent on the filtering thresholds chosen, although the overall p-value should remain relatively consistent regardless of thresholds. Practically speaking, including more than two data types becomes methodologically complicated and computing an overall p-value could quickly

become computationally unfeasible. Moreover, software is available for the case of only two data types.

The main drawback with these sequential or filtering methods is that they are based on an assumption of strong linear correlation between the same features in different data types, an assumption that could be violated either by weak correlation across two or more data types, or by nonlinear associations.

Other integration techniques focus on ways to analyze the different types of data simultaneously or within the same framework.

1.3.2.1 Concatenation/matrix factorization

The simplest form of this approach is to concatenate the separate matrices for each data type into one large matrix. Methods for dimensionality reduction could then include penalized likelihood, some kind of matrix decomposition e.g. principal components analysis (PCA) or singular value decomposition (SVD), or a combination of both.

[Barretina *et al.*, 2012] combined multiple types of genomic data from the Cancer Cell Line Encyclopedia (gene expression, gene copy number, gene mutation values, and others) together into one large matrix of genomic features: $X \in {}^{N,p}$, where N is the number of cell lines, and p is the number of predictive features. They then predicted continuous drug response with an elastic net regression algorithm [Zou and Hastie, 2005; Friedman *et al.*, 2010] which combines $L1$ and $L2$ regularized regression penalty terms in order to, on the one hand, find a parsimonious model selecting the most influential features (with $L1$), and on the other, to account for and include corelated features (with $L2$).

Combining different modes of data (e.g. continuous and binary), however, raises issues of scale and normalization that are not trivial, while including all genomic types in one

matrix puts equal weight on all data types which may not reflect the true disease mechanisms. Further, this kind of approach does not have a framework to sensibly incorporate environmental variables.

Decomposition of a concatenated matrix brings up similar issues. Performing decomposition requires an assumption that all data types share a common variance, which is not borne out in practice with different modes of genomic data. [Shen *et al.*, 2013] show that SVD on a concatenated matrix does not achieve effective integration both with simulated and real data.

Kernel methods map data into a feature space by a kernel function, which defines generalized similarity relationships between pairs of features by computing the inner product. Classification is then done on the kernel most often using a Support Vector Machine (SVM).

While methods were originally developed for homogeneous data, extensions for analyzing different data types simultaneously have been explored. [Pavlidis *et al.*, 2002] integrated two types of data for gene function classification using three approaches: concatenation of the data into one matrix (early integration), summing of the separately computed kernel matrices (intermediate integration), and summing of the discriminant values resulting from separate SVMs for each data type (late integration). The intermediate integration provided the best-performing classification. However, the authors also found that SVM with integrated data performs worse than SVM on a single data type in cases where one data type provides significantly more information than the other.

A disadvantage of this method is that kernel representations of effects of genomic markers are not directly interpretable, though methods for feature selection do exist. In terms of our purpose, if sparse environmental variables were incorporated into this framework, they could be easily overwhelmed by the high dimensional genomic data. And, in general,

computational burden becomes an issues as more data types are included.

[Li, 2013] proposes a two-stage formal integrated model for clinical outcome incorporating transcript expression and genotype data based on the idea that genetic variation affects gene regulation which affects disease probability.

Let $Y_i = 0, 1$ be disease status for the ith subject, let $Z_{ik} = 0, 1, 2$ be the number of minor alleles at the kth SNP for $k = 1, ..., K$, let X_{ij} be the expression of the jth transcript for $j = 1, ..., p$, and let $W_{il} = 1, ..., q$ be environmental covariates.

Disease risk is modeled as a function of gene expression and covariates, with gene expression modeled as a function of genetic variation and the same covariates:

$$\text{outcome model: logit } \mathrm{P}(Y_i = 1|X_i, Z_{ik}, W_i) = \alpha_0^{int} + X_i^\tau \alpha_0 + W_i^\tau \xi_0$$

$$\text{transcript model: } X_i^\tau \alpha_0 = \beta_0^{int} + Z_{ik}\beta_0 + W_i^\tau \nu_0 + \epsilon_i, \epsilon_i \sim N(0, \sigma^2)$$

An approach similar to this that integrates multiple genomic data types has been developed by [Jennings *et al.*, 2013], known as iBAG (integrative Bayesian analysis of genomics data). In iBag, the outcome model is known as the clinical model, and the transcript model is known as the mechanistic model. In the mechanistic model, the gene's expression gets partitioned into factors explained by methylation, copy number variation and other causes using principal-component-based regression. These principal components are then included in the clinical model, which they propose can find the gene expression related to the clinical outcome as well as the components that affect gene expression. (Sparsity is induced with Bayesian shrinkage.)

The main drawback here is that the appropriateness of the mechanistic model is dependent on understanding the biological relationships between data types. If the biological relationships are not well understood, or end up being more complex than the model

would indicate, the information that exists in the data will not be extracted. For example, [VanderKraats *et al.*, 2013] make the point that only modest correlations have been found between differential methylation at gene promoters and gene expression and hypothesize that this is because existing analysis methods oversimplify the representation of the data. Additionally, the method is not well-suited to incorporate binary or categorical genetic data that is also important to disease etiology. There also does not appear to be publicly available software to explore this approach.

A joint latent variable approach to classification, such as the one underlying *iCluster+*, thus offers advantages that the preceding methods do not. Namely, while accomplishing effective dimensionality reduction through latent variables, it: 1) accommodates different data types at the model level, 2) provides feature selection that takes feature effects and data type effects into account, and 3) is not reliant on prior biological knowledge or assumptions. We propose to build a classification method, *iClassify*, using this approach that will further accommodate clinical and environmental covariates and allow for genomic-covariate interaction terms.

1.4 Summary of Introduction

In this dissertation, we apply a joint latent variable approach to multimodal genomic data to gain insight into cancer studies.

For subtyping integrative data, we use the already existing method, *iCluster+*, an integrative clustering method using a joint latent variable approach, to three TCGA data sets. *iCluster+* is a generalized vertical integrative clustering method that formulates latent variables as distinct drivers of disease and accommodates heterogeneity of genomic data type. It offers penalized feature selection that allows identification of features that

directly contribute to clustering solutions.

Building on *iCluster+*'s model formulation, *iClassify* integrates different genomic platforms at the data level and uses a joint latent variable model for the purposes of prediction and classification. It additionally allows for covariate terms and genomic-covariate interaction.

The development of this method can provide a refined prediction tool for patient outcomes with systematic applications to large-scale sequencing studies, such as the TCGA ovarian cancer data set we investigate here. Moreover, this methodology has broad applications beyond predicting treatment outcomes and disease progression in cancer, including predicting prognosis and diagnosis in other diseases with major public health implications. For example, prediction of conversion to Alzheimer's disease in subjects with mild cognitive impairment and future diagnosis of post-traumatic stress disorder in patients exposed to violence are two areas of potential future investigation.

Chapter 2

iCluster: Methods and Applications

2.1 Methods

As we have noted, *iCluster+* takes a joint latent variable model approach to multimodal genomic data, jointly modeling all genomic features with a common set of unobserved latent variables that we propose represents distinct driving factors of cancer, e.g. molecular etiology and genetic pathways. These latent variables can be thought to collectively capture the major biological variations observed across cancer genomes. [Mo *et al.*, 2013]. To identify genomic features that contribute most to the biological variation and thus to proposed clustering solutions, an L_1 penalized likelihood approach is used to induce sparsity.

We will provide a more detailed formulation of *iCluster+* and then apply this method to three different TCGA studies.

Following [Mo *et al.*, 2013],let $x_{ijt}, i = 1, \cdots, n, j = 1, \cdots, p_t, t = 1, \cdots, m$ denote the genomic variables associated with the jth genomic feature in the ith subject of the tth data type. A genomic feature can be a variable such as mutation status, gene expression level or

methylation level, depending on the data type. Let $\mathbf{z}_i = (z_{i1}, \cdots, z_{ik})'$ be a column vector consisting of k unobserved latent variables.

We assume $\mathbf{z}_i$ are continuous variables that represent continuous spectrums of driver activation and follow a standard multivariate normal distribution $N(\mathbf{0}, \boldsymbol{I}_k)$. The genomic variables $x_{ijt}(j = 1, \cdots, p_t, t = 1, \cdots, m)$ are connected to the latent process via a parametric joint model in which different genomic variables are correlated through $\mathbf{z}_i$. Consider

To model, if x_{ijt} is a continuous variable (e.g. mRNA value), we assume it follows a normal distribution and consider the standard linear regression:

$$X_{ijt} = \alpha_{jt} + \boldsymbol{\beta}_{jt}\mathbf{z}_i + \epsilon_{ijt}, \epsilon_{ijt} \sim N(0, \sigma^2_{jt}), \tag{2.1}$$

where the error terms are uncorrelated and σ^2_{jt} is the residual variance not accounted for by the common associations represented by $\mathbf{z}_i$s. α_{jt} is an intercept term; and $\boldsymbol{\beta}_{jt}$ is a length-k row vector of coefficients that determine the weights genomic variable j contributes to the latent variables.

Or if x_{ijt} is a binary variable (e.g., mutation status), we consider standard logistic regression:

$$\text{logit}\{\Pr(X_{ijt} = 1|\mathbf{z}_i)\} = \alpha_{jt} + \boldsymbol{\beta}_{jt}\mathbf{z}_i,$$

where $\Pr(X_{ijt} = 1|\mathbf{z}_i)$ is the probability of gene j mutated in tumor i given the value of the latent factor $\mathbf{z}_i$.

The model also accommodates other genomic data types, such as multicategory (e.g. copy number loss, normal, gain) and poisson (e.g. sequencing count data) variables.

Assuming conditional independence of x_{ijt} given $\mathbf{z}_i$, the joint likelihood can be expressed

as this summation:

$$\ell(x_{ijt}, \mathbf{z}_i; \alpha_{jt}, \boldsymbol{\beta}_{jt}) = \sum_{i=1}^{n} \sum_{t=1}^{m} \sum_{j=1}^{p_t} \big\{ \log f(x_{ijt}|\mathbf{z}_i, \alpha_{jt}, \boldsymbol{\beta}_{jt}) + \log f(\mathbf{z}_i) \big\},$$

To obtain a sparse model that allows identification of the genomic variables that contribute to the model, the following penalized likelihood estimation is performed with the lasso (L_1) penalty:

$$\max_{\alpha_{jk}, \boldsymbol{\beta}_{jt}} \ell(X_{ijt}, Z_i; \alpha_{jt}, \boldsymbol{\beta}_{jt}) - \sum_{jt} \lambda_t \|\boldsymbol{\beta}_{jt}\|_1$$

where $\|\boldsymbol{\beta}_{jt}\|_1 = \|\beta_{j1t}\| + \cdots + \beta_{jkt}$ is the L_1-norm (lasso) penalty and λ_ts are sparsity-inducing tuning parameters with different values for each data type. Heterogeneity of the different data types is thus accommodated through the different sparsity tuning parameters. If the entire vector $\boldsymbol{\beta}_{jt}$ is zero, then the genomic variable j in data type t is removed from the model. The values of λ_t are determined by using the Bayesian information criteria (BIC).

For estimation, a modified Monte Carlo Newton-Raphson algorithm is used. As $\mathbf{z}_i$ is not observed, its joint posterior distribution using a random walk Metropolis-Hasting algorithm:

$$\mathbf{z}_i^{(r+1)}|X_{ijt} \propto f(Z_i^{(r)}) \prod_{j,t} \mathbf{f}(x_{ijt}|\alpha_{jt}, \boldsymbol{\beta}_{jt}, \mathbf{z}_i^{(r)})$$

Parameter updates are then calculated by their sample averages over repeated draws.

Once latent variables $\mathbf{z}_i$ are estimated, K-means clustering divide the n samples into $k+1$ clusters using the k latent variables. If k is unknown, it is selected from a range of k's using the BIC.

2.2 Data considerations

Mulitple genomic platforms are used in *iCluster+* analyses. In our TCGA analyses so far, we have used full sets and subsets of SCNA copy number data, DNA methylation,

and mRNA, miRNA, and ncRNA expression as input. Data are pre-processed using the following procedures: For mRNA, ncRNA and mature-strand miRNA sequence data, poorly expressed genes are excluded based on median-normalized counts, and variance filtering leads to a list of reduced features for clustering. Expression features are log2 transformed, normalized, and scaled before using them as input to *iCluster+*.

For methylation data, the median absolute deviation was employed to select the top 4000 most variable CpG sites after beta-mixture quantile normalization [Pidsley *et al.*, 2013]. We removed methylation probes with >20% or more missing data and those corresponding to SNP and autosomal chromosomes. We normalized, and scaled before using them as input to *iCluster+*.

For copy number data, Circular Binary Segmented (CBS) segmented data based on Affymetrix SNP Array 6.0 was used. We further reduced these data to a matrix of samples by non-redundant regions by adapting a method described in [Van De Wiel and Van Wieringen, 2007]. Our algorithm forms genomic regions along a chromosome defined by consecutive positions with a maximum Euclidean distance (based on copy number log-ratio segmented values) between any adjacent two probes smaller than a parameter ϵ, which determines the number of non-redundant region. Each region is then represented by its medoid signature.

2.3 Analysis of TCGA Sarcoma cohort

Adult soft tissue sarcomas are malignancies of the connective tissue, including fat, muscles, cartilage, nerves, blood vessels, and deep skin tissues. While they comprise ≈1% of adult solid tumors, they account for a disproportionate share of young adult (ages 20-39) cancer mortality due to the highly aggressive nature of many sarcomas. They are generally

classified by the soft (mesenchymal) tissue they resemble most.

The number of sample and sarcoma types in the TCGA cohort are detailed in Table 2.1. Platforms and number of features selected after pre-processing are in Table 2.2. The stated goal of the TCGA study was to "understand the genomic diversity of oncogenic drivers, to refine clinical risk stratification, and to identify potential therapeutic targets." We present here relevant results related to our *iCluster+* analyses of sarcoma tumors.

Table 2.1: TCGA sarcoma analysis: types and sample sizes

Cancer type	Abbreviation	Characteristics	No. Samples
Dedifferentiated liposarcoma	DDLPS	Undifferentiated	50
Leiomyosarcoma	LMS	Smooth muscle differentiation	80
		53 soft tissue (STLMS) and 27 uterine (ULMS)	
Undifferentiated pleomorphic sarcoma	UPS	Undifferentiated	44
Myxofibrosarcoma	MFS	Fibroblastic differentiation	17
Malignant peripheral nerve sheath tumor	MPNST	Peripheral nerves	5
Synovial sarcoma	SS	Simple-karyotype	10

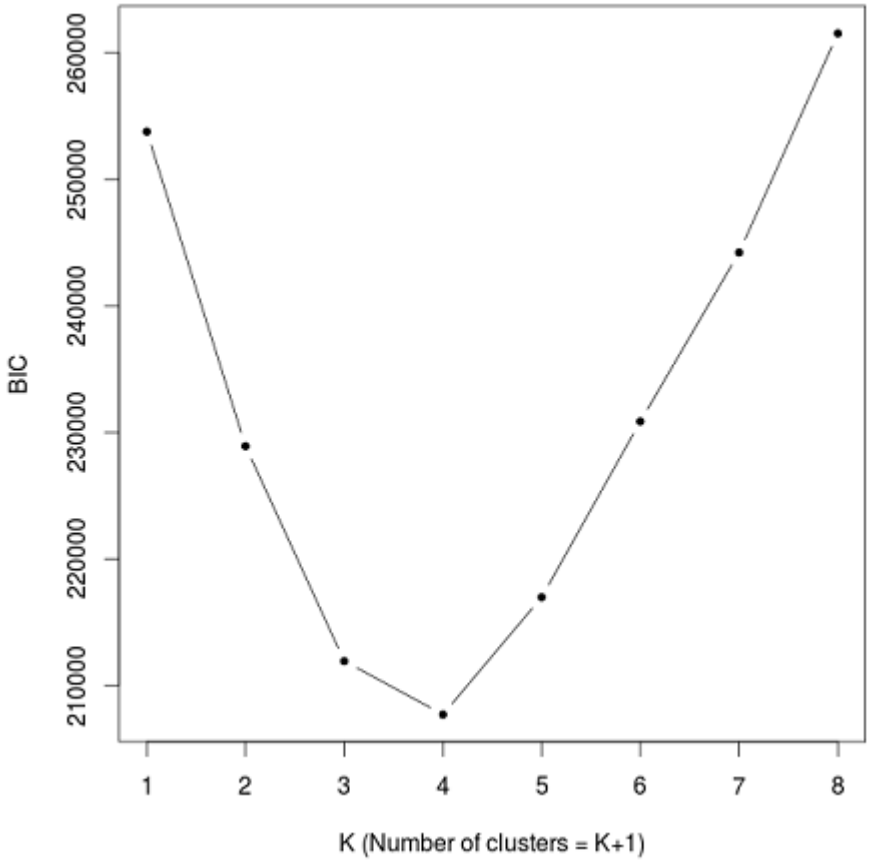
Table 2.2: TCGA sarcoma analysis: features and platforms

Genomic platform	No. of features
DNA copy number	1097
DNA methylation	1000
mRNA expression	1107
miRNA expression	171

Cross-sarcoma clustering We first performed clustering across all 206 samples and tried a range of cluster solutions from k=1 to 7. The plot of BIC by number of clusters showed a clear minimum at k=4, the 5 cluster solution (See Figure 2.1).

iCluster+ results were largely influenced by histology. SS was the most distinct sarcoma across all platforms, assigning all SS tumors into cluster C4, whose discriminatory features included high expression of FGFR3 (p = 7e-20) and miR-183 (p = 2e-25), methylation

Figure 2.1: *iCluster+* Sarcoma analysis: BIC by cluster number



of the PDE4A promoter (p = 1e-06) and partial or complete loss of chromosome 3p in 5 cases (45%). Full tables of differential features can be found in Supplementary materials of the full Sarcoma paper [Cancer Genome Atlas Network, 2017]. Unique patterns of DNA methylation, miRNA expression, and gene expression recapitulate single platform clustering for SS, and are consistent with an SS18-SSX fusion protein that is proposed to disrupt epigenetic regulation [Svejstrup, 2013].

iCluster C1 was dominated by LMS, 64 of 65 cases (98%), and was distinguished from other sarcomas largely by genes linked to myogenic differentiation, including high expression of MYLK, MYH11, ACTG2, miR-143, and miR-145 (all p ¡ 5e-39), low mRNA expression

Figure 2.2: *iCluster+* Sarcoma analysis

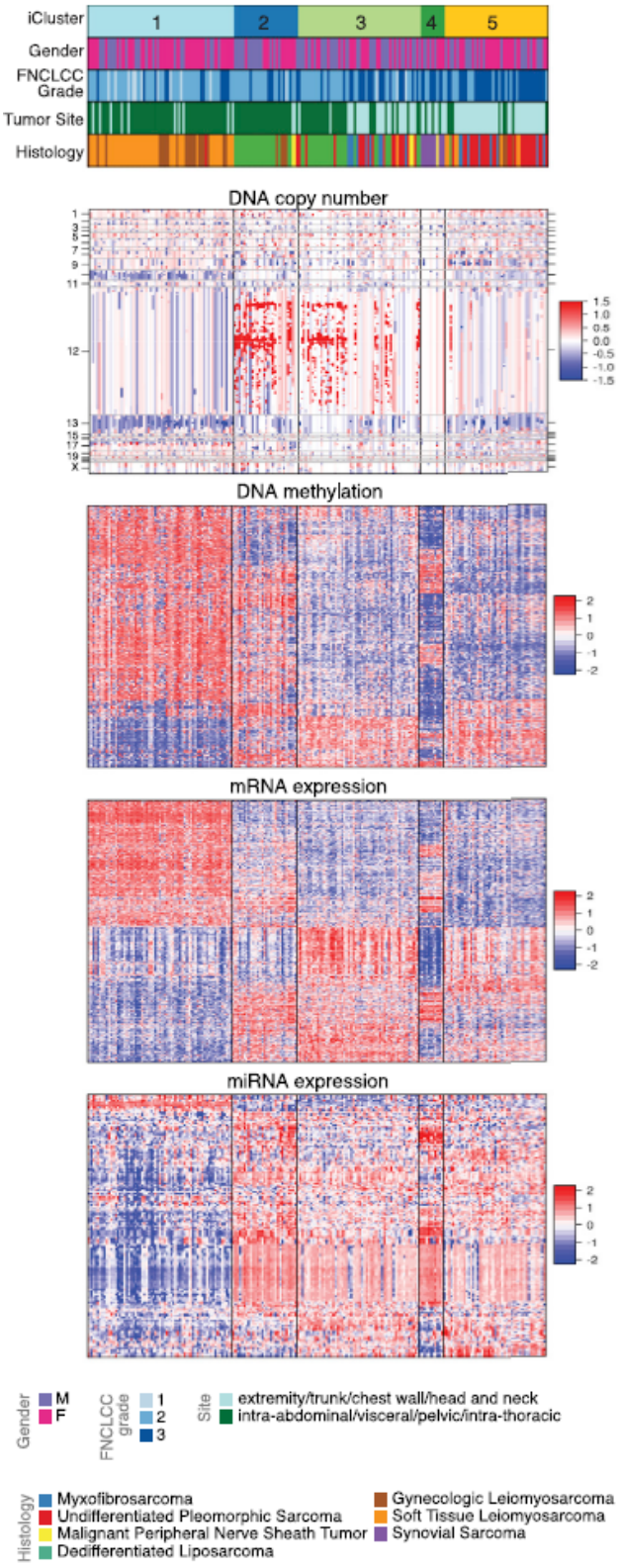


Figure 2.3: STLMS heatmap

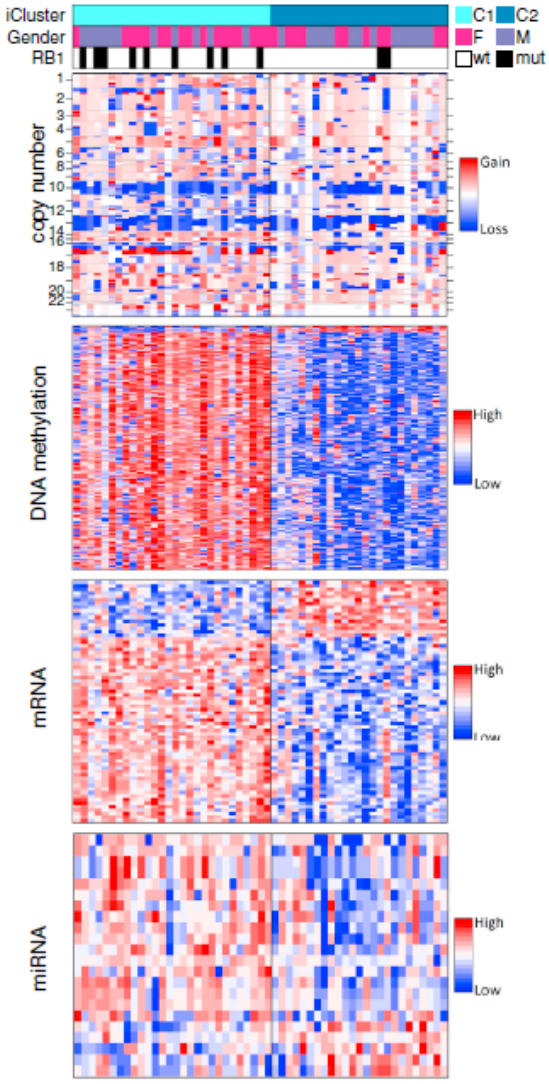
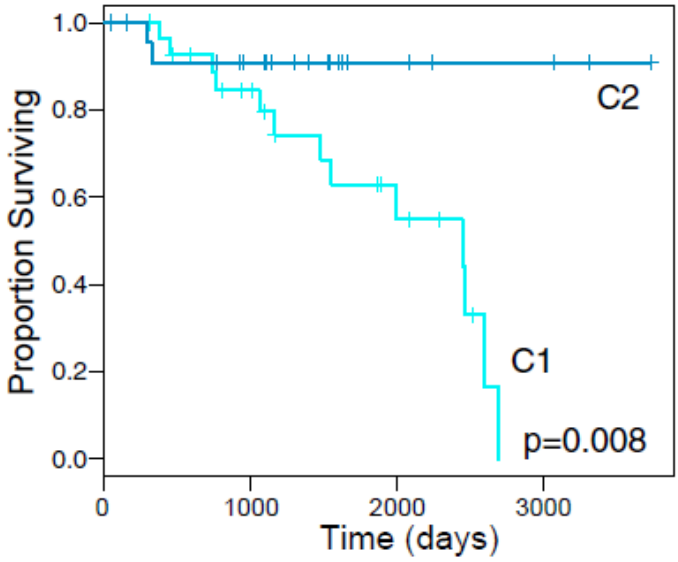


Figure 2.4: Disease Specific Survival by STLMS iCluster



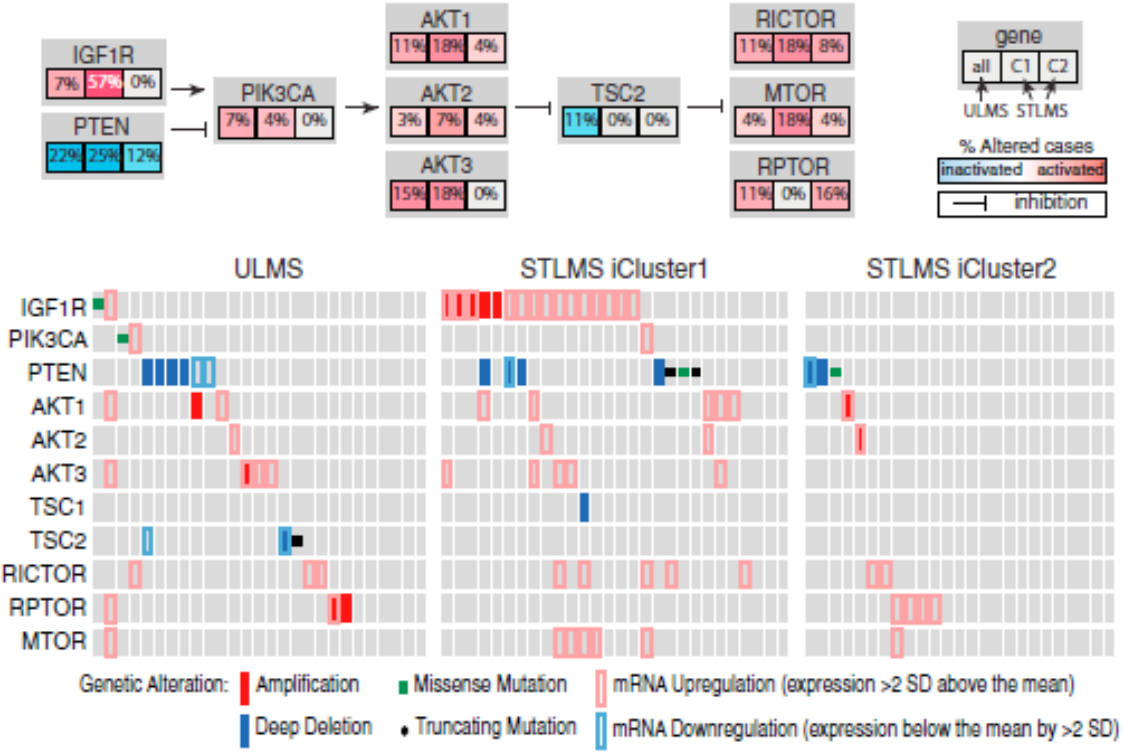
of inflammatory response genes, and low leukocyte fraction by methylation analysis. An association with grade was also noted, with iCluster C1 and C2 containing 11 of the 14 low-grade sarcomas (FNCLCC grade 1) compared to 3 in C3 and none in C4-5 ($p = 0.011$). However, this effect may be driven by iCluster separation by histologic type, as 12 of the 14 low-grade sarcomas were LMS, which was enriched in C1.

DDLPS was mainly broken up into iCluster C2 (44%) and C3 (54%), while UPS was mostly divided between C3 (36%) and C5 (59%). The five MPNST tumors were spread out over 3 iClusters.

Within-LMS clustering

We then performed clustering specifically on LMS, which has been reported to have 3 mRNA expression subtypes, i.e., a mostly uterine type and two mostly soft tissue types with very different prognoses [Guo *et al.*, 2015]. The all-LMS clustering had a minimum BIC at the 2-cluster solution, and resulted in one cluster highly associated with ULMS and the other with STLMS. To see whether we could replicate the two soft tissue types with

Figure 2.5: *iCluster+* Sarcoma analysis



different prognoses, we performed an STLMS-only clustering. Again the minimum BIC was at the2-cluster solution.

Indeed, our two clusters were consistent with reports from [Guo *et al.*, 2015] and we were able to contribute additional knowledge about inter-STLMS subtypes. STLMS C1 had worse recurrence-free survival (RFS; p = 0.0002) and DSS (p = 0.008; Figure 2.4). Compared with C2, C1 was hypermethylated (Figure 2.3) and showed higher expression of IGF1R and factors involved in cell-cycle control (CCNE2), DNA replication (MCM2), and DNA repair (FANCI) (all with adjusted $p \leq 0.03$). C1 also showed more frequent mutations of RB1 (p = 0.04) and amplification of 17p11.2-p12 (q = 0.022), a known alteration in LMS that notably includes MYOCD, encoding myocardin, a transcription factor involved

in smooth muscle differentiation. The hypomethylated STLMS C2 displayed prominent signatures of inflammatory cells, including NK cells (p = 0.004) and mast cells (p = 0.044).

The STLMS C1 cluster showed similarities with ULMS, including enrichment for PTEN deletion, mutation, or downregulation and for amplification or overexpression of AKT pathway members. Taken together, 46/55 (84%) of ULMS and STLMS iCluster C1 tumors contained alterations in the AKT pathway compared to 11/25 (44%) of STLMS iCluster C2 (p = 1e-04). Given recurrent deletion/mutation of PTEN along with frequent amplification and upregulation of IGF1R, AKT, RICTOR, and MTOR (Figure 2.5) and high AKT pathway scores by RPPA, aberrant PI3K-AKT-MTOR signaling may be crucial in LMS as a whole. Our collaborators suggest that while the effect of MTOR inhibitors such as everolimus and temsirolimus have been diminished by indirect upregulation of AKT, perhaps newer TORC1/TORC2 inhibitors and dual PI3K/ MTOR inhibitors may overcome this limitation and offer more effective therapy for LMS patients.

For other results and findings to emerge from the TCGA sarcoma study, see our report in Cell [Cancer Genome Atlas Network, 2017].

2.4 Analysis of TCGA Mesothelioma cohort

Table 2.3: TCGA Mesothelioma analysis: histology and sample sizes

Histology	No. Samples
Epithelioid	52
Biphasic	13
Sarcomatoid	3
Not otherwise specified	6

Malignant pleural mesothelioma (MPM) is a cancer of the mesothelial cells lining the pleural cavity. It was rare until the widespread use of asbestos in the mid-20th century

Table 2.4: TCGA Mesothelioma analysis: features and platforms

Genomic platform	No. of features
DNA copy number	1740
DNA methylation	4000
mRNA expression	4036
miRNA expression	304
lncRNA expression	1015

[Sekido, 2013]. Although reduction and strict regulation of asbestos use may be leading to a leveling off in new cases in Western countries, its long latency, together with continued use of asbestos in non-Western countries, ensures that MPM remains a global problem [Leong *et al.*, 2015]. MPM is almost universally lethal, with only modest survival improvements in the past decade [Yap *et al.*, 2017], suggesting that standard treatment is reaching a therapeutic plateau. Elucidating oncogenic genomic alterations in MPM is therefore essential for therapeutic progress.

To expand our understanding of the molecular landscape and biological subtypes of MPM, and provide insights that could lead to novel therapies, TCGA has conducted a comprehensive, multi-platform, genomic study of 74 MPM samples. Here, we provide *iCluster+* analysis of these 74 samples (with histology detailed in Table 2.3) and report prognostically relevant subsets of MPM with novel potential therapeutic targets. In this analysis, we used 5 data types, detailed in Table 2.4.

While the current classification of MPM into epithelioid, sarcomatoid and biphasic histologies is prognostically useful, there remains variability in clinical features and patient outcomes within histological subtypes. Previous analyses ([Bueno *et al.*, 2016], [De Reynies *et al.*, 2014]) based on mRNA expression alone have defined unsupervised clusters that largely recapitulate these histologic classes. To find out whether multi-platform molecu-

Figure 2.6: *iCluster+* Mesothelioma heatmap

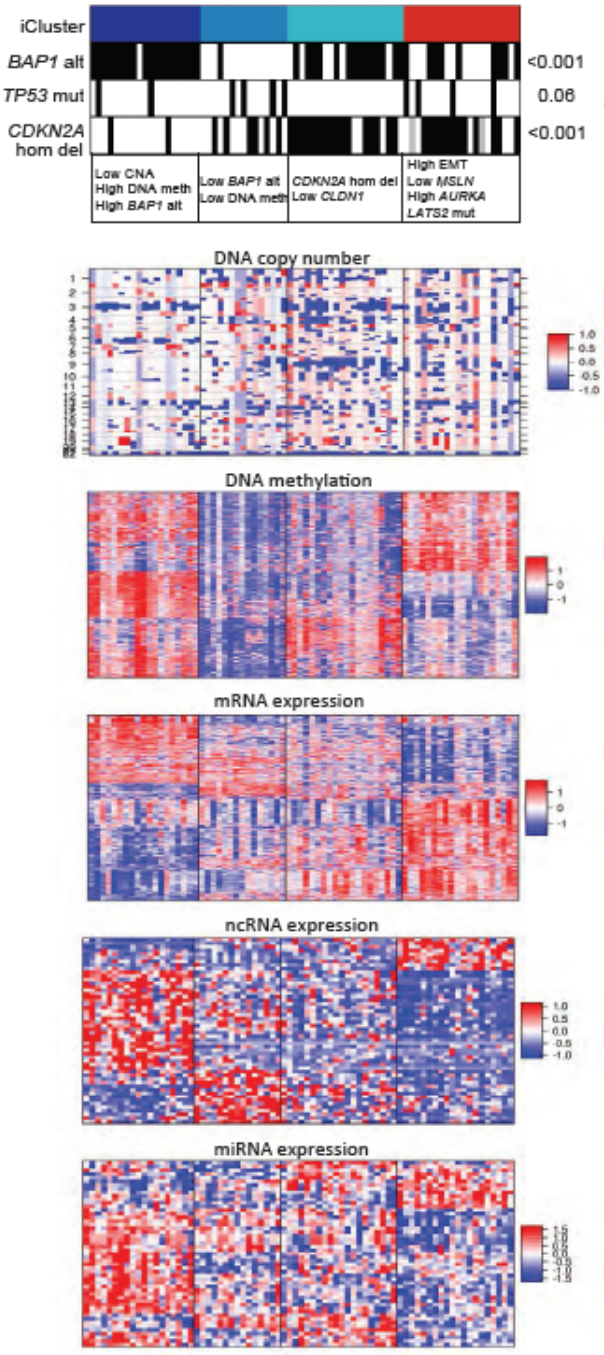


Figure 2.7: Epithelioid-only

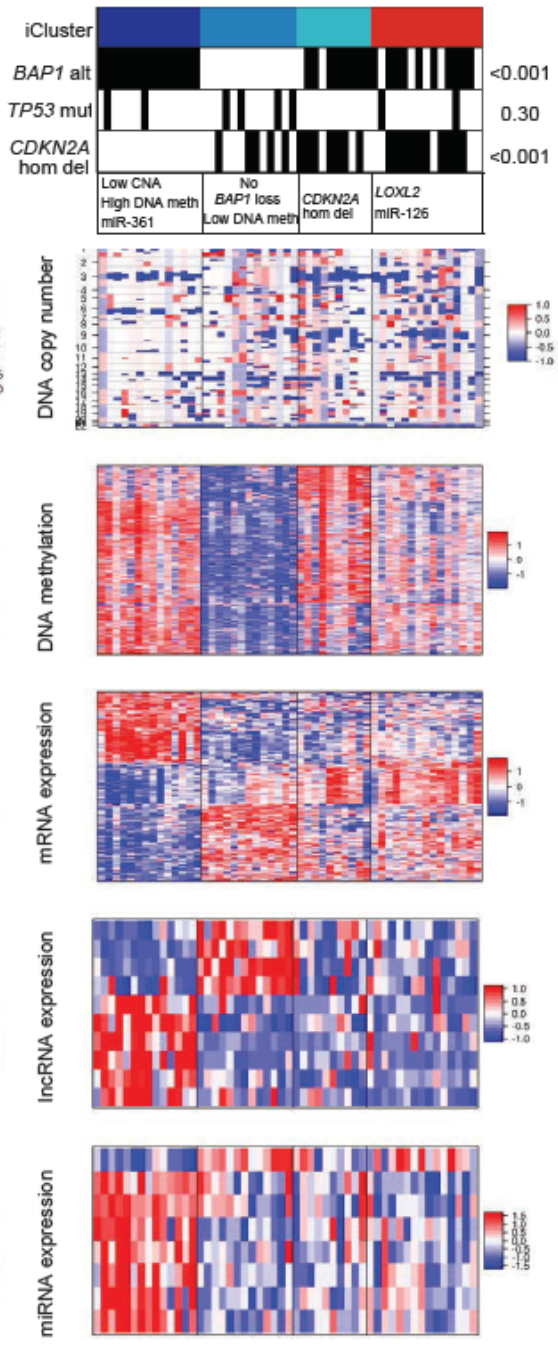


Figure 2.8: OS by MPM iClusters

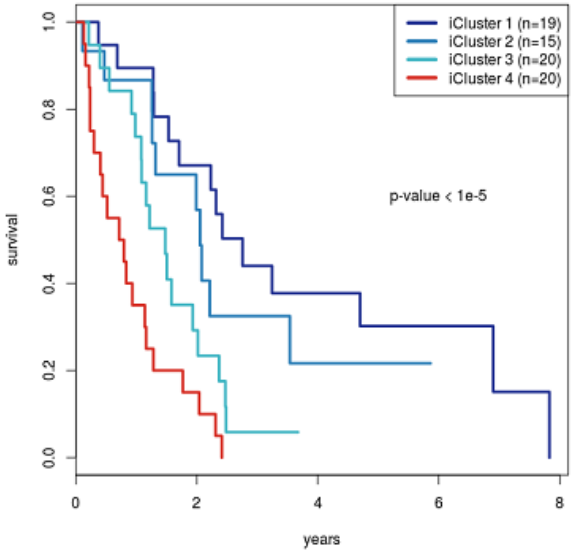
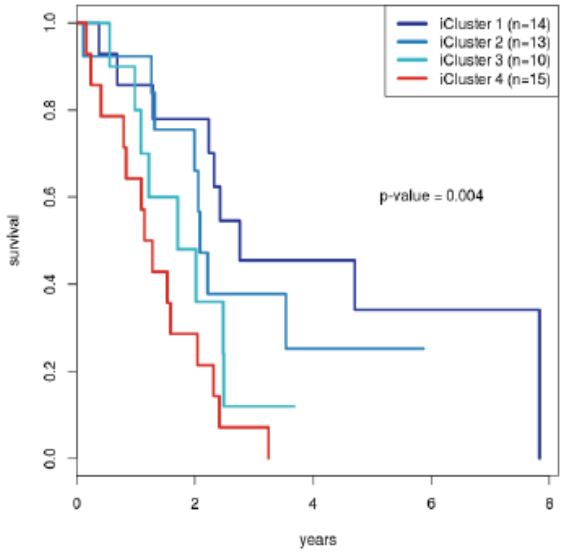


Figure 2.9: OS by MPM Epithelioid iClusters



lar profiling might provide additional resolution to define prognostic subsets of MPM, we used *iCluster+* to perform integrative clustering across multiple platforms. We identified four distinct integrated subtypes of MPM. Survival was significantly different across the 4 clusters (P¡0.001, Figure 2.8), and this survival difference remained significant (P=0.01) after adjusting for histology (epithelioid vs. non-epithelioid, Table 2.5) and age. iCluster 1 patients had the best prognosis, were likely to have undergone pneumonectomy, and were enriched for epithelioid histology. Molecularly, these tumors had low SCNA, relatively few CDKN2A homozygous deletions (11%), and a high level of methylation (Figure 2.6). All but one (95%) had BAP1 alterations: 26% had homozygous deletions and 53% had heterozygous loss with mutations.

The poor prognosis cluster (Cluster 4; red) had a high score for epithelial-mesenchymal transition (EMT) based on gene mRNA expression (P¡0.001) which was distinguished by high mRNA expression of VIM, PECAM1 and TGFB1, and low miR-200 family mRNA

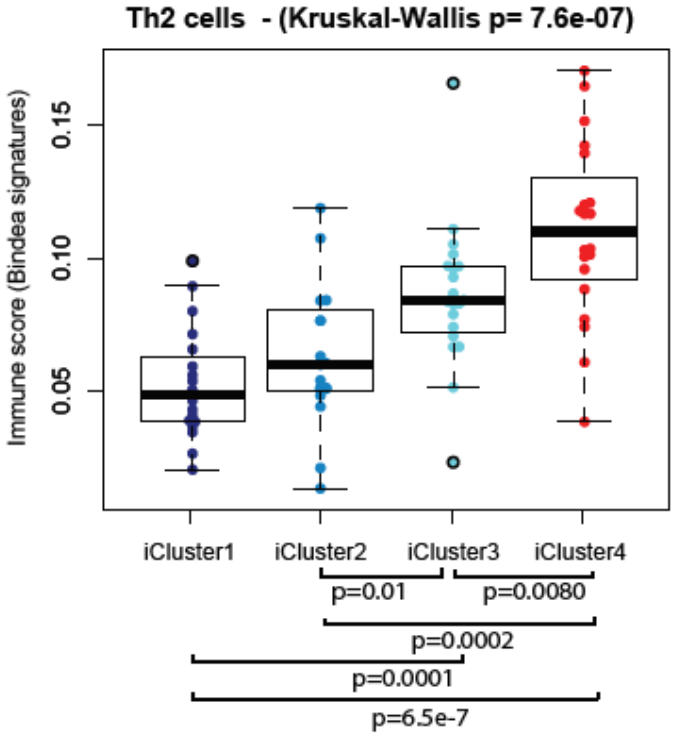
Table 2.5: All MPM: Multivariate Cox Regression

Characteristic	HR (95% CI)	P
iCluster group (ref iCluster 1)		4.60E-04
iCluster4	5.71 (2.49, 13.10)	
iCluster3	2.50 (1.14, 5.51)	
iCluster2	1.49 (0.63, 3.52)	
Histology (ref Epithelioid)		0.02
Non-Epithelioid	2.15 (1.17, 3.94)	
Age (continuous)	0.98 (0.95, 1.02)	0.37

expression. These tumors also displayed MSLN promoter methylation and consequent low mRNA expression of mesothelin, a marker of differentiated mesothelial cells, as noted previously in sarcomatoid MPM and the sarcomatoid components of biphasic MPM [Dacic *et al.*, 2008], [Tan *et al.*, 2010]. Overall, this poor prognosis cluster also showed enrichment of LATS2 mutations (30% compared to 4% in the rest of the cohort) and CDKN2A homozygous deletion (66%). Moreover, this cluster showed higher AURKA mRNA expression, higher leukocyte fraction (based on methylation), and elevated mRNA expression of E2F targets, G2M checkpoints, and DNA damage response genes. PI3K-mTOR and RAS/MAPK signaling were upregulated, based on both mRNA and protein mRNA expression. Additionally, several miRNAs were differentially expressed between the good and poor prognostic clusters, including miR-193a-3p, which has been proposed as a potential tumor suppressor [Williams *et al.*, 2015]. Finally, a comparison of immune gene mRNA expression signatures [Bindea *et al.*, 2013] across the four clusters revealed a significantly higher score for the Th2 cell signature in the poor prognosis cluster 4 compared to the other clusters (Figure 2.10). Coincidentally, it has been reported that Th2 cytokines secreted by immune cells upon exposure to asbestos may promote MPM [Mak *et al.*, 2016]).

Integrative clustering was also performed with PARADIGM ([Vaske *et al.*, 2010]). There

Figure 2.10: Th2 cells by MPM iCluster



was a strong concordance in subtype assignments between the two algorithms, especially for the best (Cluster 1) and worst (Cluster 4) prognosis clusters, indicating that integration of molecular data can identify distinct subgroups of MPM, independent of the specific statistical methodology.

While biphasic and sarcomatoid MPM are more aggressive, there remains a need for improved risk stratification of epithelioid MPM, for which clinical outcomes are more heterogeneous [Gill *et al.*, 2012]. Therefore, we performed an *iCluster+* analysis restricted to epithelioid MPM. The results for the 4-cluster epithelioid-only solution were highly similar to the 4-cluster all-MPM solution (Table 2.6), with only 7 of the 52 epithelioid samples reassigned to other clusters. This stability indicates that the features driving the all-MPM clustering are largely independent of histology. The epithelioid-only clusters share many of

the features defining the corresponding clusters in the all-MPM solution (2.7). The survival analysis also paralleled the all-MPM solution, with cluster 1 having the best outcomes and cluster 4 having the worst (2.9). Upregulation of AURKA mRNA expression in the poor prognosis epithelioid-only cluster 4 corroborated the results from the all-MPM analysis.

Table 2.6: Epithelioid cases: all-MPM vs Epithelioid-only iClusters

		Epithelioid-only iCluster			
		1	2	3	4
Epithelioid	1	14	0	1	3
cases from	2	0	13	0	0
All-MPM (n=74)	3	0	0	9	3
iCluster	4	0	0	0	9

Finally, we sought to independently validate the clinical correlations of clusters identified in the TCGA epithelioid cases using mRNA expression profiles from two published studies: 211 MPM analyzed by RNA-sequencing [Bueno *et al.*, 2016] and 52 MPM samples analyzed by mRNA expression microarrays [López-Ríos *et al.*, 2006]. Specifically, we assigned each mRNA expression profile to one of the integrative clusters based on the rules derived from the TCGA mRNA dataset. For the larger validation cohort (henceforth referred to as Bueno), we restricted our analysis to epithelioid samples and used the epithelioid-only gene signature to cluster samples. We found that the epithelioid-only samples assigned to iCluster 1 (good prognosis) had significantly better survival , even after adjusting for age (Figure 2.11). In the smaller cohort (referred to as Lopez-Rios), patient numbers were too small to split by histology. However, this analysis provided independent validation of the survival differences for the four all-MPM clusters (Figure 2.12). Taken together, these results suggest that the prognostically relevant molecular profiles defined by our analysis

Figure 2.11: Bueno OS Validation

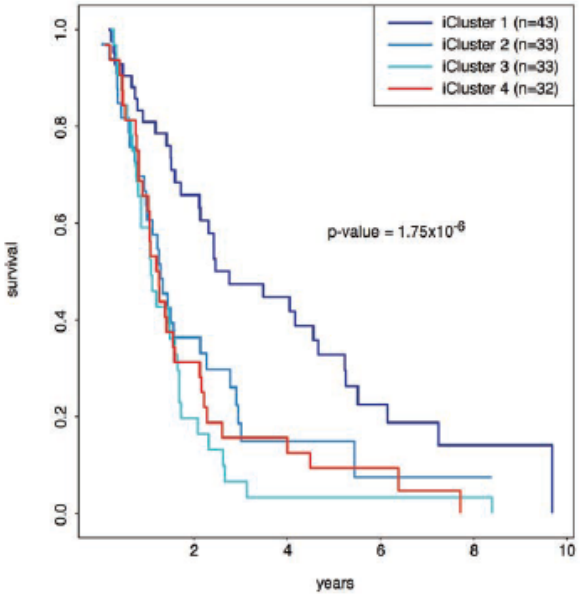
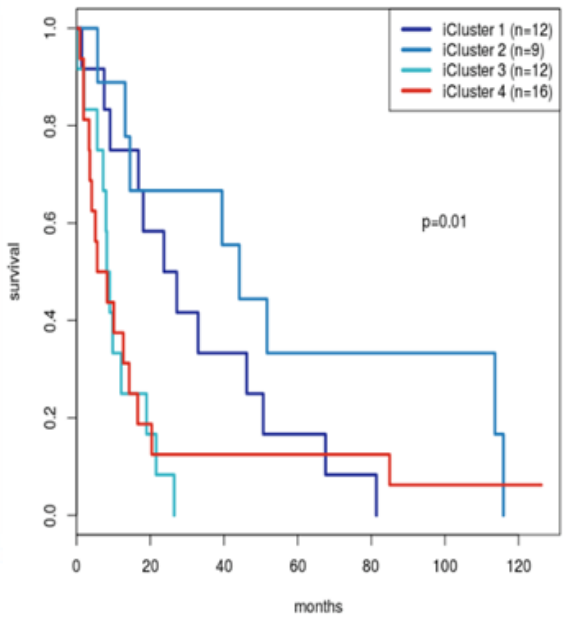


Figure 2.12: Lopez-Rios OS Validation



are robust and reproducible, and could be potentially used to improve risk stratification of patients with epithelioid MPM.

This manuscript is currently under review at Cell Cancer.

2.5 Analysis of TCGA Pancancer cohort

As discussed in Section 1.3.2, the question of interest for this pan-cancer TCGA 33 cancer type clustering analysis was whether our results would expand on those of the first pan-cancer TCGA study of 12 cancer types [Hoadley *et al.*, 2014] and find more molecular subtypes that would provide an alternative to the traditional organ and tissue-histology-based classification, or whether the majority of the molecular subtypes we would find would mirror the traditional ones.

This TCGA pancancer cohort was comprised of 9,759 tumors in TCGA, across 33 cancer

Table 2.7: TCGA pancancer analysis: cancer tumor types/categories

Cancer type	Abbreviation	No. Samples
Adrenocortical carcinoma	ACC	76
Bladder urothelial carcinoma	BLCA	399
Breast invasive carcinoma	BRCA	1031
Cervical squamous cell carcinoma	CESC	291
Cholangiocarcinoma	CHOL	36
Colon adenocarcinoma	COAD	405
Lymphoid neoplasmdiffuse large B cell lymphoma	DLBC	47
Esophageal carcinoma	ESCA	171
Glioblastoma multiforme	GBM	128
Head and neck squamous cell carcinoma	HNSC	506
Kidney chromophobe	KICH	65
Kidney renal clear cell carcinoma	KIRC	488
Kidney renal papillary cell carcinoma	KIRP	283
Acute myeloid leukemia	LAML	160
Brain lower-grade glioma	LGG	507
Liver hepatocellular carcinoma	LIHC	357
Lung adenocarcinoma	LUAD	490
Lung squamous cell carcinoma	LUSC	460
Mesothelioma	MESO	87
Ovarian serous cystadenocarcinoma	OV	294
Pancreatic adenocarcinoma	PAAD	176
Pheochromocytoma and paraganglioma	PCPG	161
Prostate adenocarcinoma	PRAD	484
Rectum adenocarcinoma	READ	148
Sarcoma	SARC	249
Skin cutaneous melanoma	SKCM	446
Stomach adenocarcinoma	STAD	407
Testicular germcell tumors	TGCT	149
Thyroid carcinoma	THCA	494
Thymoma	THYM	119
Uterine corpus endometrial carcinoma	UCEC	510
Uterine carcinosarcoma	UCS	55
Uveal melanoma	UVM	80

Table 2.8: TCGA pancancer analysis: features and platforms

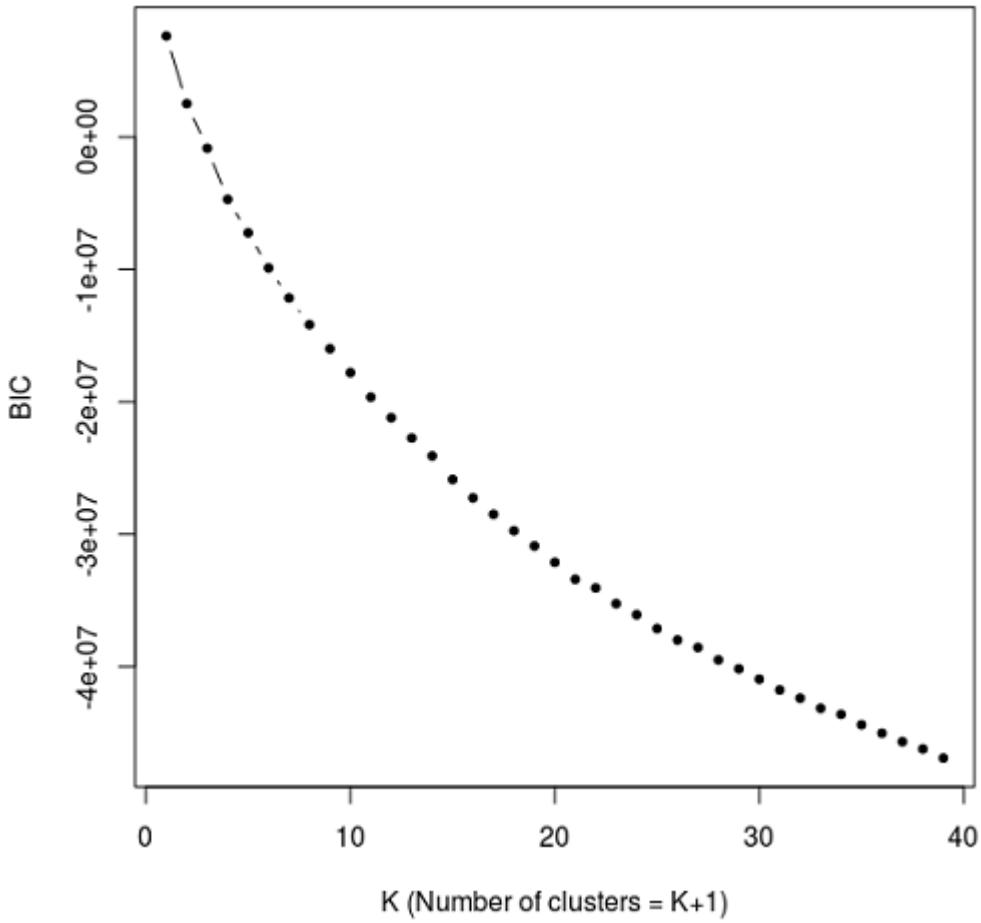
Genomic platform	No. of features
DNA copy number	3105
DNA methylation	3139
mRNA expression	3217
miRNA expression	382

types with complete data for four genomic platforms: mRNA, DNA methylation, miRNA and copy number alterations. See Table 2.7 for a breakdown of cancer types and Table 2.8 for the platforms and number of features that were input into the *iCluster+* algorithm.

We looked at a range of cluster solutions from k= 1 to 40 (number of clusters = k+1), and did not see a clear BIC minimum for any of the solutions (see Figure 2.13). There is a subtle flattening out of the curve at k=27, which pointed towards the 28-cluster solution, and further manual examination confirmed this choice. We quantified the relative contribution of each platform to the overall clustering by summing the platform-specific feature weights on the 27 iCluster latent variables. CNVs contributed 47%, mRNA and miRNA 42% and methylation 11%. Figure 2.14 shows heatmap of all features organized by iCluster, including 190 protein RPPA features that were missing in 2,808 samples (including all LAML samples) but are included in the heatmap for illustrative purposes.

Our *iCluster+* solution gave a two-pronged answer to the research question. While it emphasized the dominant role of cell-of-origin patterns, with the majority of clusters being comprised of a dominant cancer type, there were also clusters that brought together samples that had molecular similarities among histologically or anatomically related cancer types, for instance pan-kidney, pan-gastrointestinal, and pan-squamous clusters. And there were a few clusters that took members from a number of unrelated cancer types as well. As stated in the text,"Our analysis showed both divergences from and convergences with the

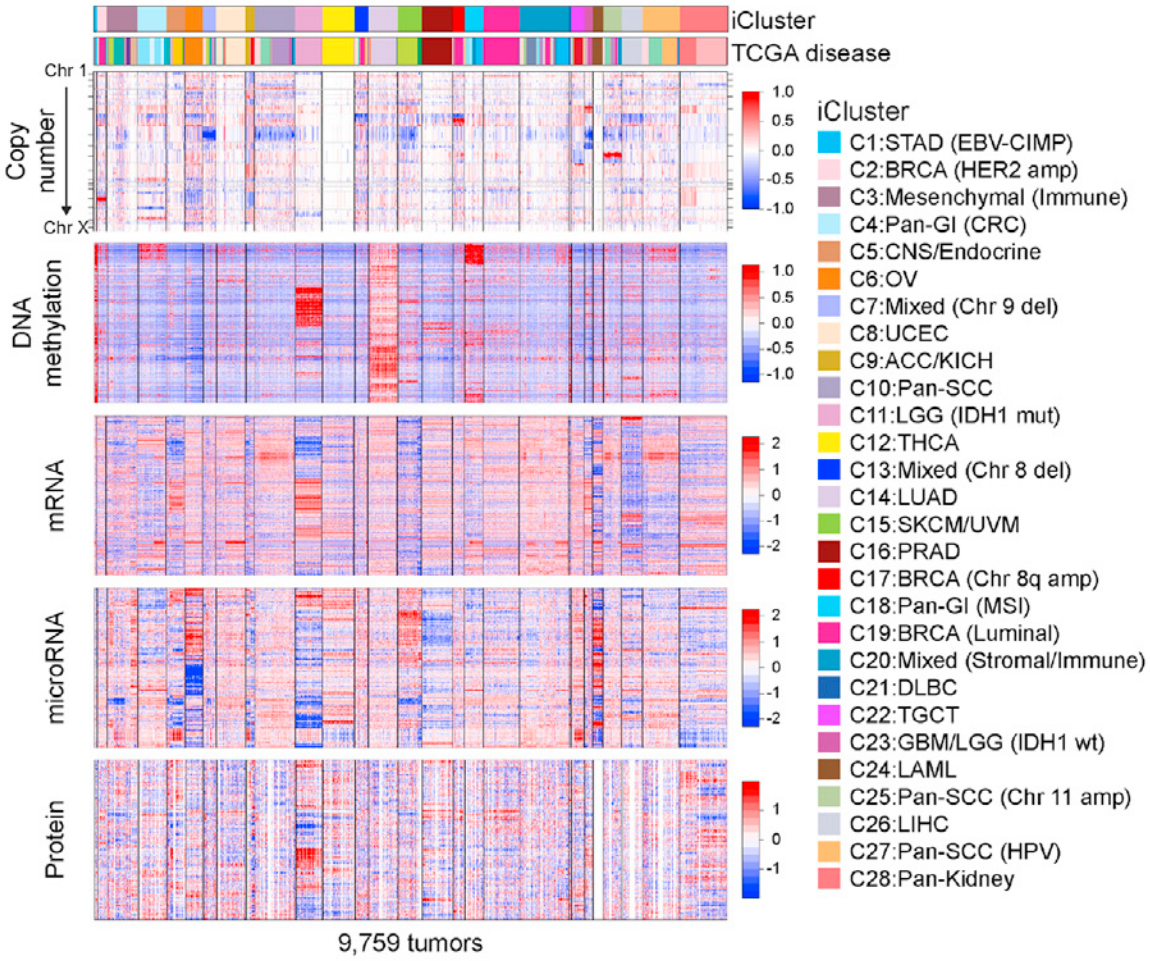
Figure 2.13: *iCluster+* Pancancer analysis: BIC by cluster number



routinely used clinical tumor classification system." The following highlights some of the interesting findings from our analysis, as reported in Cell [Hoadley *et al.*, 2018].

For 16 of the tumor types, over 80% of the samples grouped together in the same iCluster. Eight iClusters were characterized by a single tumor type (C24:LAML, C11:LGG [IDH1 mut], C6:OV, C8:UCEC, C12:THCA, C16:PRAD, C26:LIHC,C14:LUAD). Others contained tumors from similar or related cells or tissues:C28:pan-kidney (KIRC, KIRP), C15:SKCM/UVM-melanoma of the skin (SKCM) and eye (UVM), C23:GBM/LGG (IDH1wt), and C5:CNS/endocrine (PCPG). C9 (KICH, ACC) was comprised mainly of two tumor types that had high levels of hypodiploidism. Six tumor types had more diverse iCluster

Figure 2.14: *iCluster+* Pancancer analysis heatmap



membership,with less than 50% of tumors represented in a given iCluster (BLCA, UCS, HNSC, ESCA, STAD, and CHOL).

The pan-GI cohort separated into three iClusters (C1, C4, and C18), primarily driven by differences in DNA methylation profiles. C1:STAD (Epstein-Barr virus [EBV]-CIMP) consisted of hypermethylated EBV-associated tumors, and C18:pan-GI (MSI) consisted mostly of microsatellite instability (MSI) tumors of STAD and COAD. C4:pan-GI (CRC) was predominantly COAD and READ with chromosomal instability (CIN) and a distinct

aneuploidy profile (Figure 2.14). The pan-squamous cohort formed three iClusters (C10, C25, and C27). The majority of LUSC fell into C10:pan-SCC, and nearly all CESC fell into C27:pan-SCC (human papillomavirus [HPV]). Even though all squamous iClusters were characterized by chromosome 3q amplification, unique features defined C10:pan-SCC (9p deletion) and C25:pan-SCC (Chr11 amp) (Figure 2.14).

Among mixed tumor type iClusters, three were defined by copy-number alterations. C7:mixed was characterized by chr9 deletion, C2:BRCA (HER2 amp) mainly consisted of ERBB2-amplified tumors (BRCA, BLCA, and STAD), and C13:mixed (Chr8 del) contained highly aneuploid tumors, including a mixture of BRCA-Basal, UCEC (CN-high subtype), UCS, and BLCA. C3 and C20 were defined by their non-tumor-cell components including immune and stromal features.

We explored the non-tumor components of the iClusters in more detail. We estimated the stromal fraction as 1 minus tumor purity and the leukocyte fraction based on DNA methylation. Of the mixed tumor types, C20 had the highest median stromal fraction followed C3 (Figure 2.15). Each of these iClusters also displayed elevated leukocyte fractions (Figure 2.16). To estimate how much of the stromal fraction was due to immune cell infiltration, we investigated the stromal fraction versus the leukocyte fraction. In C3, more of the stromal fraction was defined by leukocytes than in C20. C3 contained predominately mesenchymal cancers, which we labeled C3:mesenchymal (immune). C20 tumors were predominately mixed epithelial cancers, which we labeled C20:mixed (stromal/immune).

To characterize composition and relative homogeneity of each iCluster, we computed the dominant-cancer-type proportion within each iCluster and plotted it against the mean iCluster silhouette width, a measure of within-group homogeneity (Figure 2.17). The silhouette widths ranged from -0.05 to 0.59, with the highest silhouette widths belonging

Figure 2.15: Stromal Proportion of Pancancer iClusters

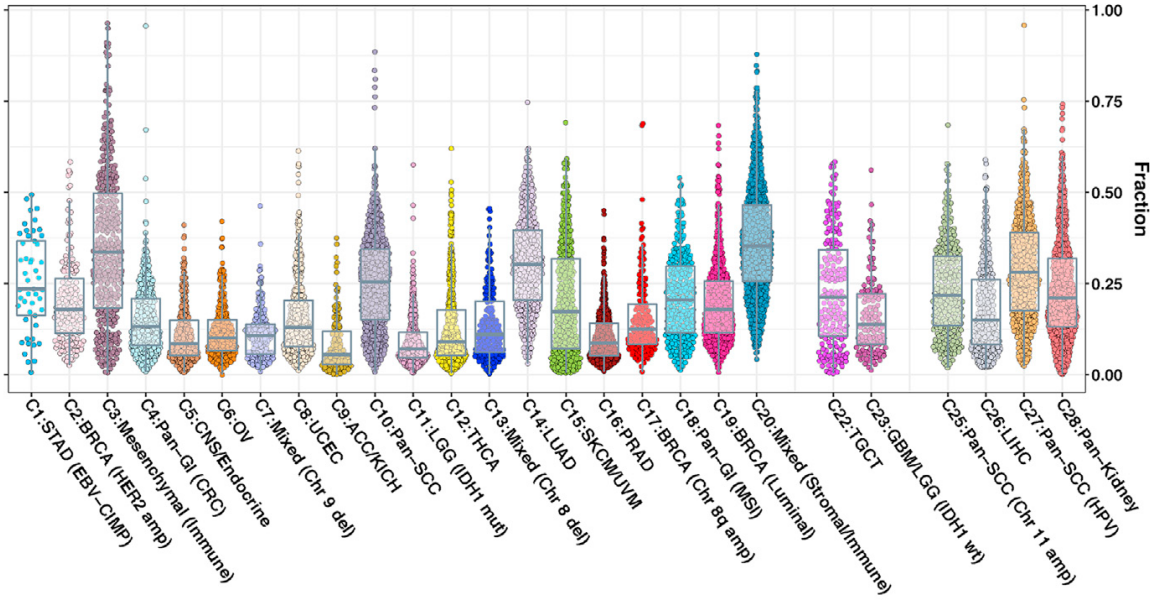


Figure 2.16: Leukocyte Proportion of Pancancer iClusters

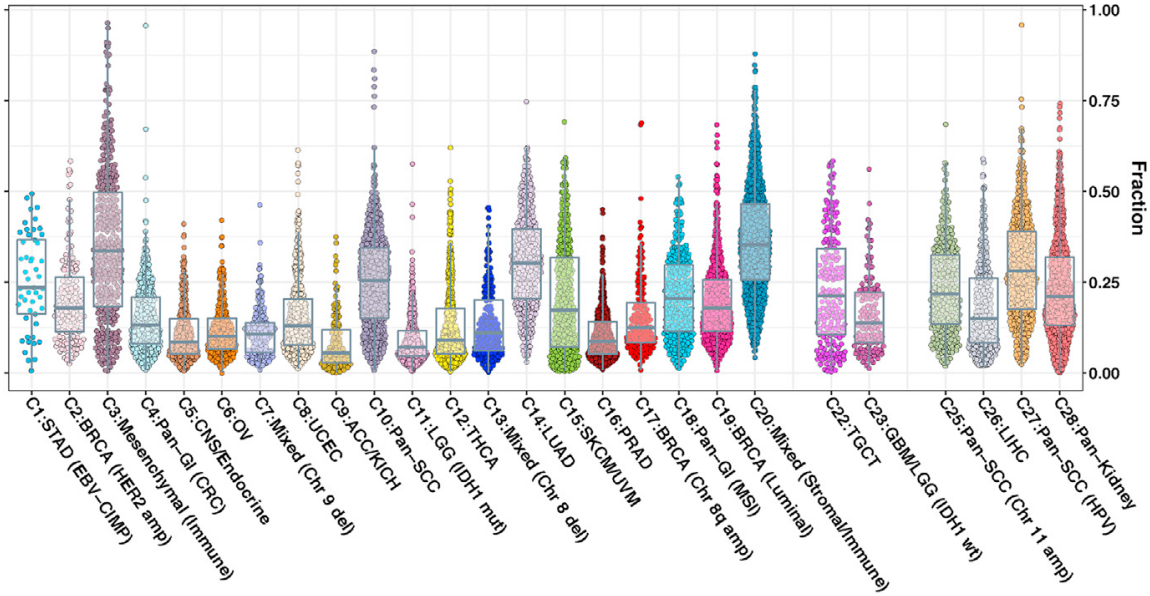
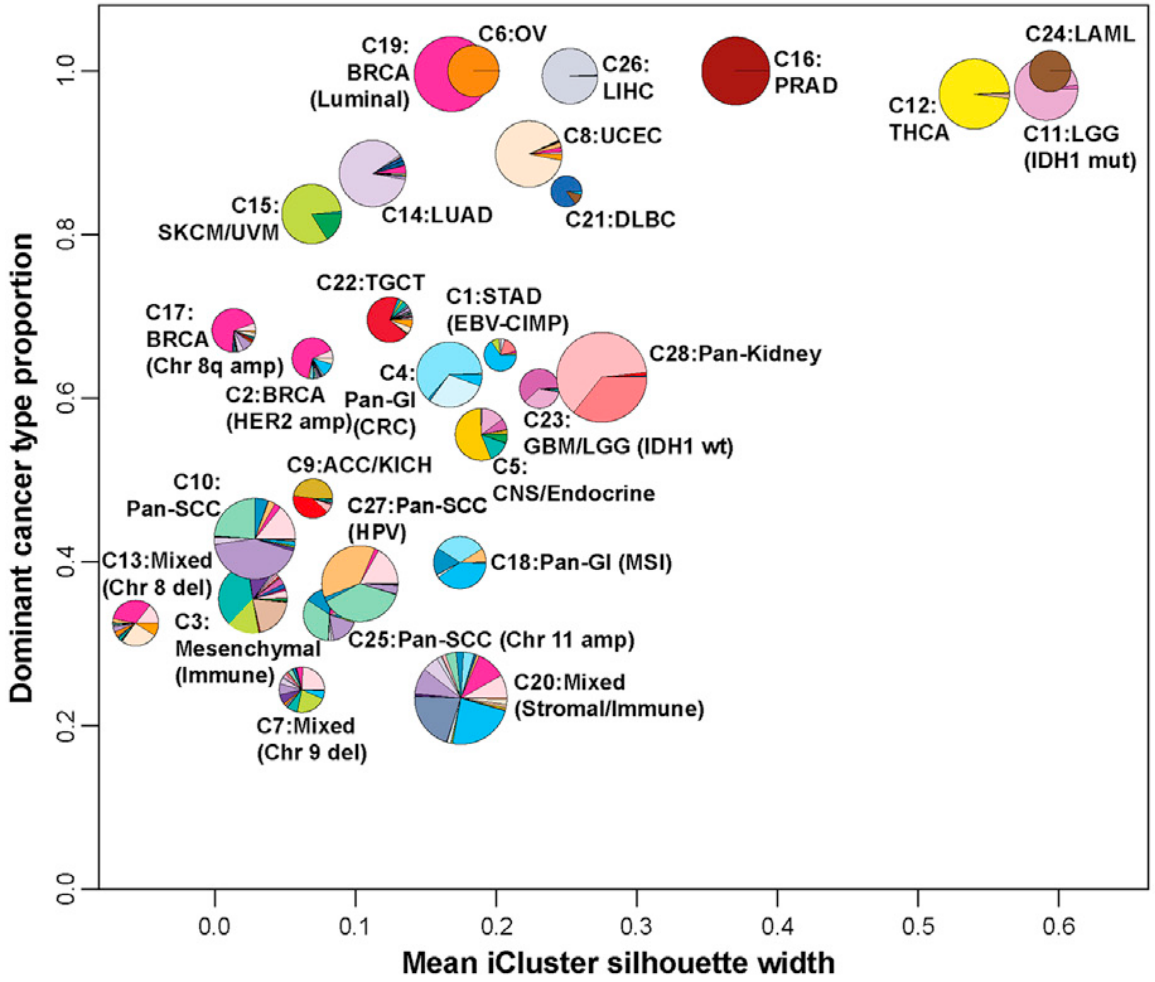


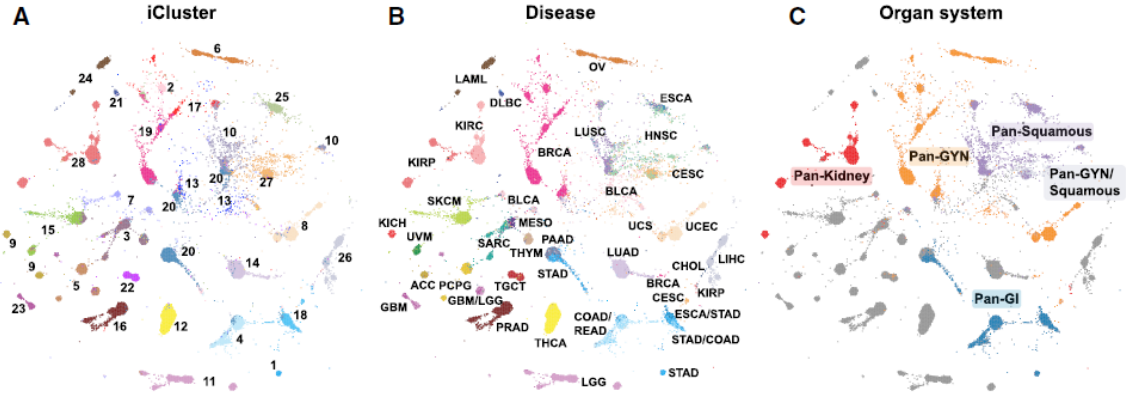
Figure 2.17: iCluster silhouette width vs. cancer type proportion



to single-cancer-type-dominant iClusters (C11:LGG [IDH1 mut], C12:THCA, C16:PRAD, and C24:LAML). Interestingly, 6 of the 7 pan-organ system iClusters (pan-GI: C1, C4, C18; pan-SCC: C25, C27, and pan-kidney: C28) had similar ranges of silhouette widths to those of single cancer-type dominant iClusters, suggesting that these were as robust as the cancer-type-dominant iClusters. iClusters driven by a shared specific chromosomal alteration (e.g., C13:mixed [chr8 del]) tended to compose multiple tumor types and appeared to have among the lowest silhouette widths, suggesting substantial molecular heterogeneity.

While a third of iClusters were mostly homogeneous for a single tumor type, the other two-thirds showed varying degrees of heterogeneity. The most diverse group, C20:mixed (stromal/immune), contained a remarkable 25 tumor types. Most of the heterogeneous iClusters, including C20:mixed (stromal/immune), contained tumor types that fell within four major cell-of-origin, or organ system, patterns: pan-GI, pan-gyn, pan-squamous, and pan-kidney.

Figure 2.18: iCluster TumorMaps



We visualized the samples by calculating Euclidean distances between the iCluster latent variables for all sample pairs and projecting the distances onto a 2D layout withTumorMap (Figure 2.18A) [Newton *et al.*, 2017]. We overlaid the tumor-type colors (Figure2.18B) which demonstrated that tumors systematically assembled along the major organ systems (Figure2.18C), lending further support and justification for the separate in-depth organ-systems-focused pan-gynecological [Berger *et al.*, 2018], pan-squamous [Campbell *et al.*, 2018], pan-gastrointestinal [Liu *et al.*, 2018], and pan-kidney [Ricketts *et al.*, 2018] TCGA reports. For the TCGA pan-squamous report, we additionally provided squamous-specific iCluster heatmaps and analyses.

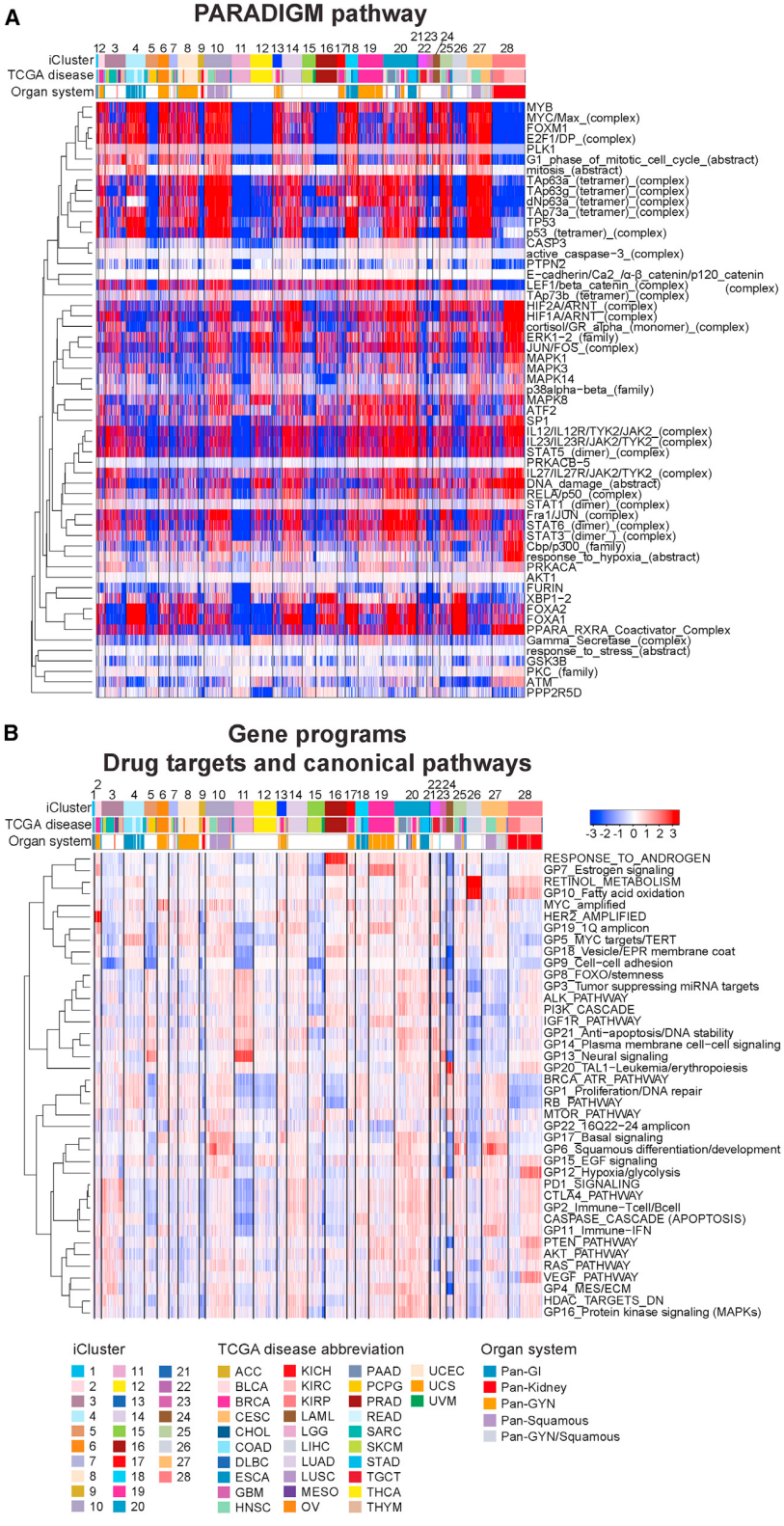
Compared to the seemingly discohesive groupings of the 17 heterogeneous iClusters, the

11 most homogeneous iClusters (C6:OV, C8:UCEC, C11:LGG [IDH1 mut], C12:THCA, C14:LUAD, C15:SKCM/UVM, C16:PRAD, C19:BRCA [luminal], C21:DLBC, C24:LAML, C26:LIHC) had higher silhouette widths, uniform tumor types, and histopathologies, but showed surprising degrees of spatial discohesion in the TumorMap, attesting to their underlying molecular heterogeneity, which has been the subject of many TCGA reports on individual cancer types.

Analysis of individual iClusters for their differentiating PARADIGM pathway features [Vaske *et al.*, 2010], canonical pathways, and gene programs amenable to drug targeting identified strong immune-related signaling features for both C3:mesenchymal (immune) and C20:mixed (stromal/immune) tumors, suggesting that they may share potential susceptibility to immunotherapy. C20:mixed (stromal/immune) and C3:mesenchymal (immune) tumors were commonly enriched for gene programs representing PD1, CTLA4, and GP2-T cell/B cell activation (Figure 2.19B), indicating that new therapies targeting these specific immune pathways might be appropriate. Relatedly, PARADIGM analysis (Figure 2.19A) showed that C3:mesenchymal (immune) and C20:mixed (stromal/immune) tumors shared upregulated JAK2/STAT1,3,6 signaling with C14:LUAD tumors and C10:pan-SCC, pointing to the possibility of treating these diverse iCluster tumors with JAK-STAT agents currently approved to treat rheumatoid arthritis, myelofibrosis, polycythemia vera, and other non-malignant diseases.

We investigated other characteristics of the 28 iClusters as well, including mutational assessment, cancer stemness, and immune subtypes. These analyses and additional description of results can be found in [Hoadley *et al.*, 2018].

Figure 2.19: iCluster pathway analyses



2.6 Summary of *iCluster+* analyses

iCluster+'s joint latent variable approach integrates mutiple genomic platforms at the data level and allows us to effectively model distinct global driving factors in tumor cells. Integrative cluster analyses performed on three different TCGA cohorts with *iCluster+* demonstrate its ability to reveal moleculary distinct groups of tumors.

In the TCGA sarcoma cohort (Section 2.3), we were able to recapitulate two soft tissue leiomyosarcoma (LMS) RNA-only subtypes and add significantly to the characterization and understanding of these prognostically-distinct groups. We also showed that one of these subtypes shared common aberrant PI3K-AKT-MTOR signaling with uterine LMS, which may be a crucial disease mechanism in LMS overall.

Our analysis of the TCGA mesothelioma cohort was able to go beyond histology to find four distinct malignant pleural mesothelioma (MPM) subtypes. We molecularly characterized these subtypes, which were stable both over all MPM samples and within only the epithelioid samples, and were able to replicate in two external data sets the differentially better prognosis of Cluster 1 relative to Cluster 4. These results could potentially be used to improve risk stratification, particularly in epithelioid MPM.

Our pancancer TCGA clustering of almost 10,000 cancer tumors over 33 different cancer types led to 28 different iClusters with a mix of clusters dominated by specific cancer type/tissue and those that were comprised by a wider range of cancer types. A number of our clusters were made up of samples from distinct organ systems and this provided justification for a a number of pan-organ-systems TCGA studies, including pan-kidney, pan-gynecological, pan-gastrointestinal, and pan-squamous. We demonstrated that two of our most mixed-type cancer clusters (C3 and C20) were enriched for immune-related

signaling features which might signal a potential for immunotherapy.

Chapter 3

iClassify: Statistical Methodologies

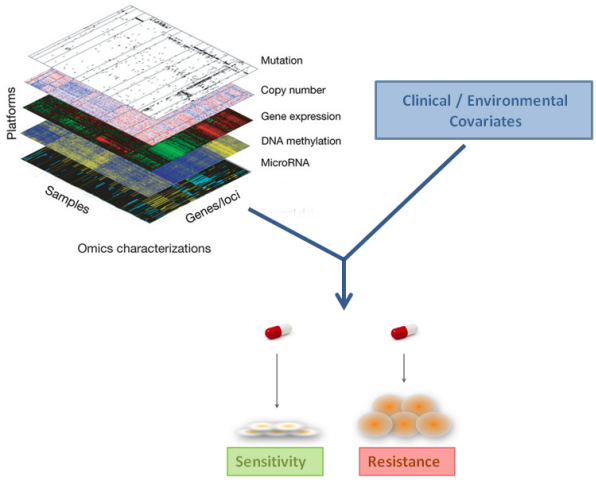
To solve the classification problem of predicting response to platinum chemotherapy in ovarian cancer, we developed a method based off of the joint latent variable modeling of iCluster.

Latent variables in *iClassify* represent distinct latent driving factors for each platform, which are predictive of the values of the original genomic variables. *iClassify* then models the likelihood of genomic factors, latent variables, and disease status jointly, and allow for environmental and clincial covariates. Our fitted model resultsare then used to predict disease or response status future subjects. A schematic representation of the overarching goals of *iClassify* is depicted in Figure 3.1.

3.1 Statistical framework for integrative genomics

Let $X_{ijt}, i = 1, \cdots, n, j = 1, \cdots, p_t, t = 1, \cdots, m$ denote genomic variables in the ith subject, jth genomic feature in the tth data type. A genomic feature can be a random variable such as gene expression, methylation, or copy number alteration level, depending on the data type. We introduce data-source-specific latent factors (e.g., $\boldsymbol{Z}_{it}, t = 1, ..., m$) to

Figure 3.1: Schematics of the Proposed Integrative Genomic Method, ***iClassify***



summarize information in each data type t. Let D_i be the disease status indicator, and let $\boldsymbol{Z}_i = \{\boldsymbol{Z}_{it}, t = 1, ..., m\}$.

Our approach is based on retrospective likelihood in a case-control design as used in many genomic studies. The conditional likelihood of the observed genomic features $\{X_{ijt}\}$ given D_i is

$$\begin{aligned} & \prod_i \ell(X_{i11}, \cdots, X_{ip_m m}|D_i) \\ = & \prod_i \int \ell(X_{i11}, \cdots, X_{ip_m m}|\mathbf{Z}_i, D_i)\ell(\mathbf{Z}_i|D_i)\mathrm{d}P(\mathbf{Z}_i), \end{aligned} \tag{3.1}$$

where $P(\mathbf{Z})$ denotes the probability measure of $\mathbf{Z}$. Assume that the latent genomic drivers fully summarize information in X_{ijt}, conditioning on $\mathbf{Z}_i$, X_{ijt} are independent of D_i. Thus, the likelihood (3.1) simplifies to

$$\prod_i \ell(X_{i11}, \cdots, X_{ip_m m}|D_i) = \prod_i \prod_{j,t} \int \ell(X_{ijt}|Z_{it})\ell(Z_{it}|D_i)\mathrm{d}P(\mathbf{Z}_{it}). \tag{3.2}$$

The first component of equation (3.2) is the product of the likelihood of every genomic feature over every data type. The model will vary based on data type. For continuous genomic variables (e.g., log-transformed gene expression, methylation logM), we use a stan-

dard linear regression model:

$$X_{ijt} = \alpha_{jt} + \beta_{jt} Z_{it} + \epsilon_{ijt}, \epsilon_{ijt} \sim N(0, \sigma^2_{jt}), \tag{3.3}$$

where β_{jt} represent coefficients for each genomic feature and ϵ_{ijt} is an independent error term with mean zero and variance σ^2_{jt}. The model can be extended to accommodate binary, multicategory and poisson genomic variables. The second term in (3.2) is the distribution of $\mathbf{Z}_i$ given D_i. We assume a normal distribution for the latent genomic drivers given disease status as:

$$Z_{it}|D_i = d \sim N(\mu_{dt}, \sigma^2), \quad \mu_{dt} = \gamma_{0t} + \gamma_{1t} d, \quad d = 0, 1, \tag{3.4}$$

where $\boldsymbol{\gamma}$'s represent coefficients for each data type, and in particular γ_{1t} represents the coefficient for data type effect on disease.

In *iClassify*, the first component of the likelihood in equation (3.2) summarizes information in genomic features and the second component performs disease prediction through latent genomic features. We will maximize the likelihood jointly across the data types. One distinction of the model (3.4) with traditional classification methods such as Fisher's linear discriminant analysis (LDA) is that $\mathbf{Z}_i$ is a latent unobserved variable.

3.2 Estimation procedure

For model estimation, we use a modified Monte Carlo Expectation-Maximization (EM) algorithm treating Z_{it} as missing data. The algorithm consists of an E-step and an M-step. As latent variable Z_{it} is not observed in our model, and the likelihood for different data types can vary (e.g., linear model and logistic model), direct computation of the E-step may be difficult. Thus, we adopt Monte-Carlo EM and repeatedly sample from the joint

posterior distribution of Z_{it} given observed data,

$$Z_{it}^{(r+1)}|X_{ijt}, D_i \propto f(Z_{it}^{(r)}|D_i) \prod_{j,t} \ell(X_{ijt}|Z_{it}^{(r)}) \tag{3.5}$$

using a random-walk Metropolis-Hasting algorithm. We then calculate parameter updates by their sample averages over the repeated draws of the latent variables. Typically, we draw 1000 times. For the prior distribution of $Z_{it}^{(r)}$, we draw samples based on μ_d obtained with updated γ. In the M-step, using the complete data likelihood, $\prod_i \prod_{j,t} \ell(X_{ijt}|Z_{it})\ell(Z_{it}|D_i)$, standard methods in linear regression models and generalized linear models yield updated parameter estimates. The algorithm continues to iterate through these steps until it reaches convergence.

3.3 Feature selection

To induce a sparse model, we perform hard-thresholding, where we estimate a threshold parameter λ that incorporates estimates for both β_{jt} and γ_t coefficients, thereby accommodating the heterogeneity of different genomic data types. For feature selection, we determine the value of our threshold parameter λ using cross-validated classification error as the criterion in our model selection process. In order to do this, we repeatedly partition the data into a training and a testing set. For each value of λ on a grid points of the range of $\beta_{jt}\gamma_t$, we first obtain parameter estimates from the training set and use these to make predictions (response or non-response) in the testing set, for which we calculate classification error. The λ that results in the lowest classification error is the threshold we select denoted as λ^*. Features with a combined effect of $|\widehat{\beta}_{jt}\widehat{\gamma}_t| < \lambda^*$ are set to have null effects. Since *iClassify* is a likelihood-based approach, alternative measures such as AIC or BIC can be combined with hard-thresholding for feature selection.

The tuning procedure for large datasets is computationally intensive. To minimize computing time, we have developed a parallel processing algorithm (as in *iCluster*) that can take advantage of a computing cluster.

3.4 Prediction of disease status for new subjects

For a new subject with observed genomic variables $\mathbf{X}^*$, but unobserved disease status, we first obtain the marginal probability for predicting disease status:

$$P(D^* = 1|\mathbf{X}^*) = \int_{\mathbf{z}} P(D^* = 1, \mathbf{z}|\mathbf{X}^*) dP(\mathbf{z})$$

$$= \left\{ \int_{\mathbf{z}} f(\mathbf{X}^*|D^* = 1, \mathbf{z}) f(\mathbf{z}|D^* = 1) P(D^* = 1) d\mathbf{z} \right\} / P(\mathbf{X}^*)$$

$$= \left\{ \int_{\mathbf{z}} f(\mathbf{X}^*|\mathbf{z}) f(\mathbf{z}|D^* = 1) P(D^* = 1) d\mathbf{z} \right\} / P(\mathbf{X}^*),$$

where the above second equation follows from conditional independence of $\mathbf{X}$ and D given $\mathbf{Z}$, and

$$P(\boldsymbol{X}^*) = \int_{\mathbf{z}} \sum_{d=0,1} f(\mathbf{X}^*|D^* = d, \mathbf{z}) f(\mathbf{z}|D^* = d) P(D^* = d).$$

We use Monte-Carlo integration, sampling from priors of $f(\mathbf{Z}_i^{(r)}|D^* = 1)$ or $f(\mathbf{Z}_i^{(r)}|D^* = 0)$, and taking averages to obtain the desired marginal probabilities.

If the predicted probability of disease/response is greater than 0.5, we will classify the subject as diseased.

3.5 Inclusion of covariates and interaction with genomic drivers

To extend the framework to covariates, we let $\boldsymbol{E}_i$ denote covariates (e.g. environmental, clinical, germline mutation).

Then the conditional likelihood of $\{X_{ijt}, \boldsymbol{E}_i\}$ given D_i is

$$\begin{aligned} & \prod_i \ell(X_{i11}, \cdots, X_{ip_m m}, \boldsymbol{E}_i | D_i) \\ = & \prod_i \int \ell(X_{i11}, \cdots, X_{ip_m m} | \mathbf{Z}_i, D_i, \boldsymbol{E}_i) \ell(\mathbf{Z}_i | D_i, \boldsymbol{E}_i) \ell(\boldsymbol{E}_i | D_i) \mathrm{d}P(\mathbf{Z}_i) \end{aligned} \tag{3.6}$$

The association between covariates and disease status is not of our primary interest here. Therefore we drop the term $l(\boldsymbol{E}_i|D_i)$ in the subsequent discussion, and the likelihood in Equation (3.6) simplifies to:

$$\prod_i \ell(X_{i11}, \cdots, X_{ip_m m}, \boldsymbol{E}_i | D_i) = \prod_i \prod_{j,t} \int \ell(X_{ijt} | Z_{it}) \ell(Z_{it} | D_i, \boldsymbol{E}_i) \mathrm{d}P(\mathbf{Z}_{it}). \tag{3.7}$$

While the covariates $\boldsymbol{E}_i$ do not bear on the first term in Equation (3.7), they can be naturally incorporated into the second term, which is now the distribution of $\mathbf{Z}_i$ given D_i and $\boldsymbol{E}_i$. With the same normality assumption,

$$Z_{it} | D_i = d, \boldsymbol{E}_i = \mathbf{e} \sim N(\mu_{d\mathbf{e}t}, \sigma^2)$$

$$\mu_{d\mathbf{e}t} = \gamma_{0t} + \gamma_{1t} d + \boldsymbol{\gamma}_{2t}^T \mathbf{e} + \boldsymbol{\gamma}_{3t}^T \mathbf{e}^* \times d, d = 0, 1.$$

Importantly, interaction between D_i and E_i can be modeled, and we may consider a subset (instead of the full set) of the covariates (denoted by $\mathbf{e}^*$) that interact with genomic drivers to influence disease. The estimation procedure incorporates these extended likelihoods containing $\boldsymbol{E}_i$. Note that the interaction between genomic features and covariate risk factors can be tested parsimoniously by $\boldsymbol{\gamma}_{3t}$, which has the same dimension as the covariates. Testing parameters $\boldsymbol{\gamma}_{3t}$ gives an overall strength of interaction between all genomic features in the same platform and covariate risk factors. Thus *i*Classify alleviates the multiple comparisons issue for testing interaction effects. The interaction effect of a specific genomic feature j with covariates can be estimated by $\beta_{jt} * \boldsymbol{\gamma}_{3t}$.

The marginal probability for predicting disease status now depends on observed covariates $\boldsymbol{E}^*$ as well as observed genomic variables $\mathbf{X}^*$:

$$P(D^* = 1|\mathbf{X}^*, \boldsymbol{E}^*) \tag{3.8}$$

$$= \left\{ \int_{\mathbf{z}} f(\mathbf{X}^*|\mathbf{z}, \boldsymbol{E}^*) f(\mathbf{z}|D^* = 1, \boldsymbol{E}^*) P(D^* = 1, \boldsymbol{E}^*) d\mathbf{z} \right\} / P(\mathbf{X}^*|E^*), and$$

$$P(\boldsymbol{X}^*|E^*) = \int_{\mathbf{z}} \sum_{d=0,1} f(\mathbf{X}^*|D^* = d, \mathbf{z}, \boldsymbol{E}^*) f(\mathbf{z}|D^* = d, E^*) P(D^* = d, E^*).$$

We now sample from priors of $f(\mathbf{Z}_i^{(r)}|D^* = 1, \boldsymbol{E}_i)$ or $f(\mathbf{Z}_i^{(r)}|D^* = 0, \boldsymbol{E}_i)$, and take averages to obtain the desired marginal probabilities. Again, a subject will be predicted to be as diseased or a responder if the corresponding probability is greater than 0.5.

Chapter 4

iClassify: Simulation Studies

To assess the performance of our method, we performed several simulation studies with differing sample size under different scenarios. In all cases, we simulated 100 replicates and measured prediction accuracy through three-fold cross-validation.

4.1 Integration of genomic platforms

Table 4.1: Simple scenarios ($n = 100$): parameter estimation

	Scenario A. Strong $\boldsymbol{\beta}_{jt}$'s (1-1.5), strong $\boldsymbol{\gamma}$'s (both 1.5)				Scenario B. Strong $\boldsymbol{\beta}_{jt}$'s (1-1.5), weak $\boldsymbol{\gamma}$'s (1, 0.5)				Scenario C. Weak $\boldsymbol{\beta}_{jt}$'s (0.5-1), weak $\boldsymbol{\gamma}$'s (1, 0.5)				Scenario D. Weak $\boldsymbol{\beta}_{jt}$'s (0.5-1), weakest $\boldsymbol{\gamma}$'s (both 0.5)			
Para	True	Est	Bias	MSE	True	Est	Bias	MSE	True	Est	Bias	MSE	True	Est	Bias	MSE
$\boldsymbol{\beta}_{1,1}$	1.5	1.43	-0.07	0.02	1.5	1.44	-0.06	0.02	-1	-0.97	0.04	0.02	-1	-0.84	0.16	0.09
$\boldsymbol{\beta}_{2,1}$	1	0.94	-0.06	0.01	-1	-0.96	0.04	0.02	0.5	0.49	-0.01	0.01	0.5	0.45	-0.05	0.04
$\boldsymbol{\beta}_{3,1}$	0	0	0	0.01	0	0	0	0.01	0	0	0	0.01	0	-0.01	-0.01	0.01
$\boldsymbol{\beta}_{4,1}$	0	-0.02	-0.02	0.01	0	0	0	0.01	0	0.02	0.02	0.01	0	0.01	0.01	0.01
$\boldsymbol{\beta}_{5,1}$	1.5	1.44	-0.06	0.02	1.5	1.43	-0.07	0.02	-1	-0.96	0.04	0.02	-1	-0.83	0.17	0.08
$\boldsymbol{\beta}_{1,2}$	1	1	0	0.02	1	0.95	-0.05	0.05	-0.5	-0.42	0.08	0.03	-0.5	-0.44	0.06	0.05
$\boldsymbol{\beta}_{2,2}$	0	0.01	0	0.01	0	0	0	0.01	0	0	0	0.01	0	-0.02	-0.02	0.02
$\boldsymbol{\beta}_{3,2}$	0	0	0	0.01	0	0	0	0.01	0	0	0	0.01	0	-0.02	-0.02	0.02
$\boldsymbol{\beta}_{4,2}$	1.5	1.38	-0.12	0.03	-1.5	-1.21	0.29	0.09	1	0.63	-0.37	0.06	1	0.67	-0.33	0.08
$\boldsymbol{\beta}_{5,2}$	0	-0.02	-0.02	0.01	0	-0.01	-0.01	0.01	0	0	0	0.01	0	-0.01	-0.01	0.02
$\boldsymbol{\gamma}_{1,1}$	1.5	1.63	0.13	0.1	1	0.99	-0.01	0.07	1	1.1	0.1	0.06	0.5	0.55	0.05	0.06
$\boldsymbol{\gamma}_{2,1}$	1.5	1.59	0.09	0.09	0.5	0.51	0.01	0.07	0.5	0.53	0.03	0.09	0.5	0.61	0.11	0.08

Table 4.2: Simple scenarios (n=100): prediction accuracy

Scenario	1 - (Bayes error)	iClass (SD)	LDA (SD)	LR (SD)
A. Strong $\boldsymbol{\beta}_{jt}$'s (1-1.5), strong $\boldsymbol{\gamma}$'s (both 1.5)	0.82	0.82 (0.04)	0.78 (0.04)	0.78 (0.04)
B. Strong $\boldsymbol{\beta}_{jt}$'s (1-1.5), weak $\boldsymbol{\gamma}$'s (1, 0.5)	0.69	0.67 (0.05)	0.64 (0.06)	0.64 (0.05)
C. Weak $\boldsymbol{\beta}_{jt}$'s (0.5-1), weak $\boldsymbol{\gamma}$'s (1, 0.5)	0.67	0.67 (0.04)	0.62 (0.05)	0.61 (0.05)
D. Weak $\boldsymbol{\beta}_{jt}$'s (0.5-1), weakest $\boldsymbol{\gamma}$'s (both 0.5)	0.61	0.58 (0.05)	0.55 (0.05)	0.55 (0.05)

Table 4.3: Simple scenarios (n=200): parameter estimation

	Scenario A. Strong $\boldsymbol{\beta}_{jt}$'s (1-1.5), strong $\boldsymbol{\gamma}$'s (both 1.5)				Scenario D. Weak $\boldsymbol{\beta}_{jt}$'s (0.5-1), weakest $\boldsymbol{\gamma}$'s (both 0.5)			
Para	True	Est	Bias	MSE	True	Est	Bias	MSE
$\boldsymbol{\beta}_{1,1}$	1.5	1.42	-0.08	0.01	-1	-0.93	0.07	0.02
$\boldsymbol{\beta}_{2,1}$	1	0.95	-0.05	0.01	0.5	0.47	-0.03	0.01
$\boldsymbol{\beta}_{3,1}$	0	0	0	0	0	0	0	0.01
$\boldsymbol{\beta}_{4,1}$	0	0	0	0	0	0	0	0.01
$\boldsymbol{\beta}_{5,1}$	1.5	1.42	-0.08	0.01	-1	-0.94	0.06	0.02
$\boldsymbol{\beta}_{1,2}$	1	0.98	-0.02	0.01	-0.5	-0.55	-0.05	0.01
$\boldsymbol{\beta}_{2,2}$	0	0.01	0.01	0	0	0	0	0.01
$\boldsymbol{\beta}_{3,2}$	0	0	0	0	0	0	0	0.01
$\boldsymbol{\beta}_{4,2}$	1.5	1.36	-0.14	0.01	1	0.8	-0.2	0.02
$\boldsymbol{\beta}_{5,2}$	0	-0.01	-0.01	0	0	0.01	0.01	0.01
$\boldsymbol{\gamma}_{1,1}$	1.5	1.64	0.14	0.04	0.5	0.53	0.03	0.04
$\boldsymbol{\gamma}_{2,1}$	1.5	1.65	0.15	0.05	0.5	0.56	0.06	0.04

In a simple setup, we created 100 replicates of a dataset comprised of 50 cases and 50 controls, with 10 features from two different genomic data types (e.g. 5 RNA features and 5 methylation features). Four of the features were noise variables which do not associate with disease status. Our interest was to determine both estimation and classification accuracy of our method.

We simulated 50% of the samples in each dataset as cases and 50% as controls, and then simulated latent variables $\mathbf{Z}_{it}$ using the true $\boldsymbol{\gamma}$ parameters according to equation (3.4). We then used these simulated $\mathbf{Z}_{it}$ and true α_{jt}, β_{jt}, and ϵ_{ijt} parameters to simulate features

X_{ijt}. Our simulated errors were were always $\sim N(0,1)$. This simulation process resulted in datasets with on average 40-50% of features correlated, levels of correlation which are common in genomic data sets.

We compared prediction accuracy of our method to prediction accuracies obtained through Linear Discriminant Analysis (LDA) and standard logistic regression (LR), methods with a unifying framework that concatenates all of the data features together into a single matrix without distinguishing between them. All features were included as observed data and no latent variables were considered in these alternative methods since they do not handle latent effects. Classification in those methods was performed either with a discriminant function $\delta_D(x)$ in the case of LDA or with prediction probabilities $Pr(D = 1|X = x)$ as in LR [Hastie *et al.*, 2009].

In these scenarios, we varied true β values from stronger (1-1.5) to weaker (0.5-1) , and true $\boldsymbol{\gamma}$ values from strong (both 1.5) to weaker (0.5 and 1) to weakest (both 0.5). In Table 4.1, we see that in all cases the parameters are being estimated with small MSE. The bias is within the range of variability of the estimator and stabilizes with increasing sample size.

Table 4.2 shows that prediction accuracy is higher with *iClassify* than LDA or LR. Interestingly, the prediction accuracy appears to be driven primarily by the $\boldsymbol{\gamma}$ values. Simple scenario A with the strongest $\boldsymbol{\gamma}$ values results in significantly better prediction accuracy across all methods than all other scenarios, and prediction accuracy decreases with $\boldsymbol{\gamma}$ values. Although scenario B has stronger β values than scenario C, this has no effect on prediction accuracy, which does not vary between the two. However, we assume that the 2% increase in Bayes error accuracy from Scenario C to B is driven by the those stronger β values. We did not simulate a scenario with weak β and strong $\boldsymbol{\gamma}$ values, but based on the similarity of prediction accuracies for Scenario B (strong $\boldsymbol{\beta}$, weak $\boldsymbol{\gamma}$) and Scenario C (strong $\boldsymbol{\beta}$, weak

$\boldsymbol{\gamma}$), we assume the prediction results would be close to Scenario A (strong β and strong $\boldsymbol{\gamma}$).

In order to see the effect of sample size, we re-simulated Scenarios A and D, this time with 100 cases and 100 controls (n=200). Table 4.3 shows the decreases in Bias and MSE that we would expect from the increase in sample size. Table 4.4 shows the effect of increasing the sample size on prediction accuracy. For Scenario A with large $\boldsymbol{\gamma}$ values, the increase in sample size does not affect iClassify's performance, but does lead to increased prediction accuracy from the comparative methods. When there are large genomic platform effects, iClassify performs well even at small sample sizes. For the smaller $\boldsymbol{\gamma}$ values of Scenario D, the increase in sample size boosts both iClassify's prediction accuracy and the prediction accuracy for LDA and LR by 2%.

Table 4.4: Simple scenarios ($n = 200$): prediction accuracy

Scenario	iClass	LDA	LR
A. Strong $\boldsymbol{\beta}_{jt}$'s (1-1.5), strong $\boldsymbol{\gamma}$'s (both 1.5)	0.82	0.81	0.81
D. Weak $\boldsymbol{\beta}_{jt}$'s (0.5-1), weakest $\boldsymbol{\gamma}$'s (both 0.5)	0.60	0.57	0.57

In more complex scenarios, we simulated 50 features across three data types (e.g. 20 RNA-seq, 20 miRNA seq, 10 CNAs), 25 of which were noise variables. Here, we varied β values from strong (1-2) to weak(0.5-1) and $\boldsymbol{\gamma}$ values from somewhat stronger (all 1) to somewhat weaker (1, 0.5, 1), and simulated with both 50 cases/50 controls and 100 cases/100 controls. Results in Table 4.5 show reasonable parameter estimation that improved for β estimation in particular as sample size improved.

In this more complex setting, logistic regression did not converge, and iClassify significantly outperformed LDA in prediction accuracy by 8-9% (Table 4.5). Again, we see the strong effect of $\boldsymbol{\gamma}$ values and the relatively weaker effect of β values on prediction accuracy. While only one $\boldsymbol{\gamma}$ value decreased from Scenario E to Scenario F, this presumably caused

a decrease of 4% in prediction accuracy. Conversely, a more significant increase in all $\boldsymbol{\beta}$ values from Scenario F to Scenario G resulted in only a 1% increase in prediction accuracy.

As the latent variable structure induces correlation in the genomic feature simulation, there was again correlation in 40-50% of features. This simulation setting demonstrates the robustness of *iClassify* in performing prediction in the presence of correlation and noise variables, due to using a likelihood framework to integrate latent effects of genomic features. In contrast, LDA's performance is deteriorated, probably in part because of the effect multicollinearity on the performance of LDA [Hastie *et al.*, 1995].

We also looked at a complex scenario with different genomic platforms contributing different weights to the model: strong ($\boldsymbol{\gamma}_{1,1} = 2$), medium ($\boldsymbol{\gamma}_{1,2} = 1.25$), weak ($\boldsymbol{\gamma}_{1,3} = .75$). $\boldsymbol{\beta}$ values were kept relatively weak (0.1-1) across data types. Table 4.6 shows a similar increase in performance of *iClassify* over LDA with the stronger $\boldsymbol{\gamma}$'s driving a higher classification accuracy of 89%.

Table 4.5: Three data type scenario (n=200): Estimation and Prediction Accuracy

Scenario	$M\bar{S}E_{\beta}$	$M\bar{S}E_{\gamma}$	iClass	LDA
E. Weak $\boldsymbol{\beta}_{jt}$'s (0.5-1), stronger $\boldsymbol{\gamma}$'s (all 1)	0.01	0.03	0.77	0.69
F. Weak $\boldsymbol{\beta}_{jt}$'s (0.5-1), weaker $\boldsymbol{\gamma}$'s (0.5-1)	0.01	0.03	0.74	0.65
G. Strong $\boldsymbol{\beta}_{jt}$'s (1-2), weaker $\boldsymbol{\gamma}$'s (0.5-1)	0.00	0.03	0.75	0.67

Table 4.6: Varying γ effects setup (n=200): Estimation and Prediction Accuracy

	$M\bar{S}E_{\beta}$	$M\bar{S}E_{\gamma}$	iClass	LDA
H. Weak $\boldsymbol{\beta}_{jt}$'s (0.2-1), varying $\boldsymbol{\gamma}$'s (2, 1.25, 0.75)	0.01	0.04	0.89	0.81

4.1.1 Feature selection

We used the "Varying $\boldsymbol{\gamma}$ effects" scenario from Table 4.6 to assess our feature selection method. 50 features were simulated across three data types, 25 of which were noise variables.

For feature selection, we determine the value of our threshold parameter λ using cross-validated classification error as the criterion. For each value of λ in a range of $\beta_{jt} * \gamma_t$, we obtain parameter estimates from the training set and use these to make predictions (response or non-response) on the testing set, for which we calculate classification error. The λ that results in the lowest classification error is the threshold we select. Since *iClassify* is a likelihood-based approach, alternative measures such as AIC or BIC can be combined with hard-thresholding for feature selection.

We compared results from our feature selection method to those from cross-validation lasso regression which maximizes a penalized version of the log likelihood:

$$\max_{\beta_0,\beta} \left\{ \sum_{i=1}^{N} \left[y_i(\beta_0 + \beta^T xi) - log(1 + e^{\beta_0 + \beta^T xi}) \right] - \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

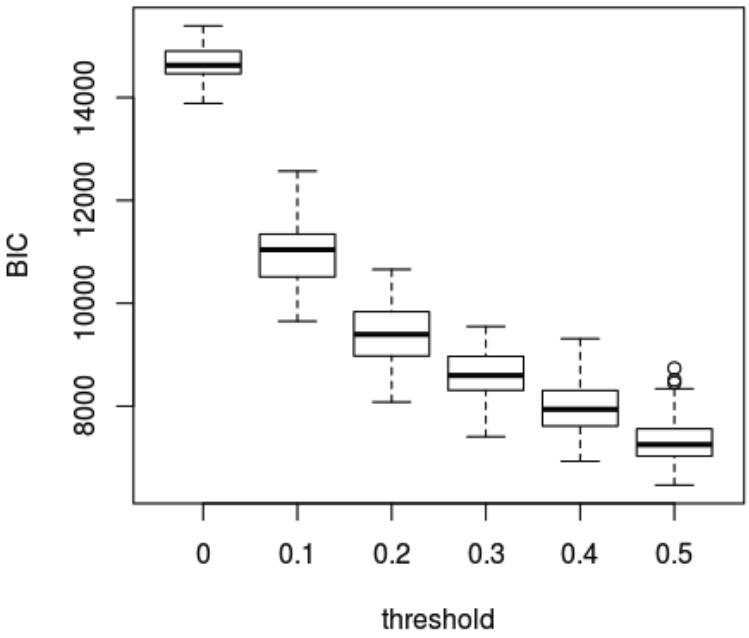
We used classification accuracy to determine the tuning parameter λ.

Table 4.7: Varying γ effects simulation: Hard thresholding and prediction accuracy

λ	iClass accuracy	false positive rate	false negative rate
0.0	0.890	1.00	0.00
0.1	0.890	0.22	0.00
0.2	0.891	0.02	0.02
0.3	0.889	0.00	0.10
0.4	0.888	0.00	0.28
0.5	0.888	0.00	0.45
	Lasso accuracy	false positive rate	false negative rate
	0.863	0.12	0.54

Table 4.7 shows that a λ value of 0.2 produces the highest *iClassify* prediction accuracy (0.891), which coincides with the lowest false positive and false negative rates of any λ. In comparison, the Lasso accuracy rate at the *L1* penalty parameter is 0.863, with a 10% higher false positive rate and notably a >50% higher false negative rate. Interestingly,

Figure 4.1: Varying γ effects simulation: BIC by threshold

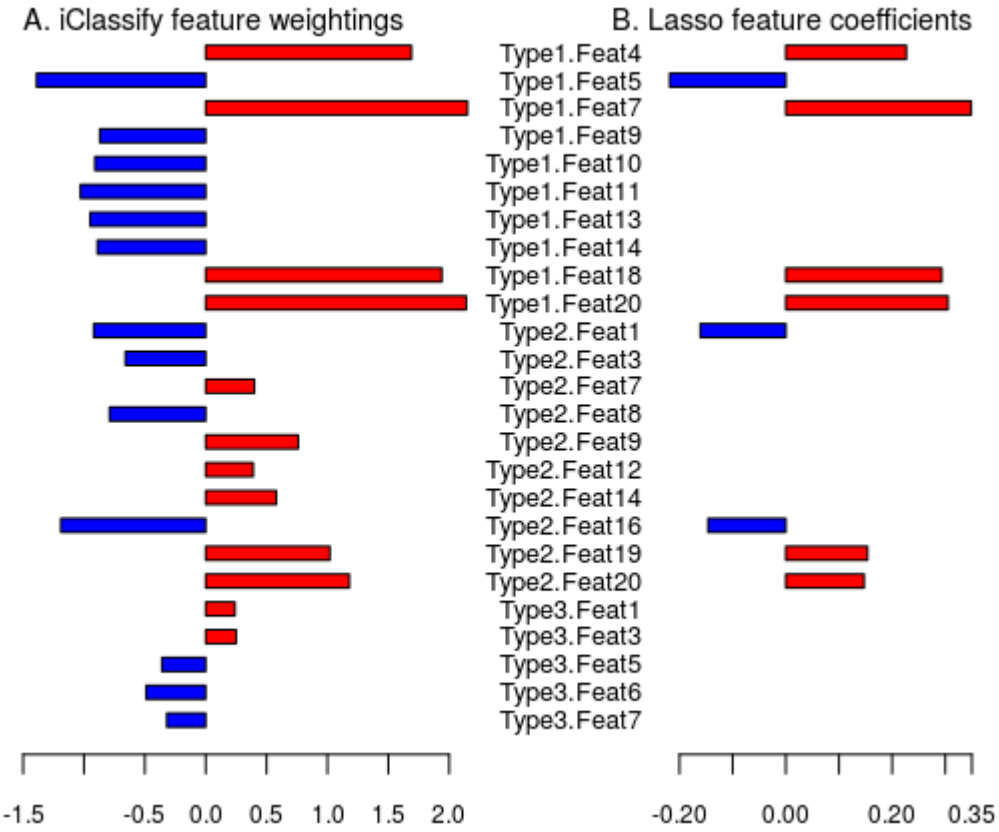


the *iClassify* accuracy at all λs is higher than the Lasso accuracy, demonstrating robust classification regardless of feature selection.

We also examined the BIC in terms of feature selection. Due to the complex nature of the model, the BIC here does not achieve a minimum (Figure 4.1). But we observe a flattening out of the change after $\lambda = 0.2$, which agrees with the cross-validation classification accuracy selection criteria.

At the hard threshold of $\lambda = 0.2$, $\boldsymbol{\beta} * \boldsymbol{\gamma}$ estimates averaged over all replicates allow iClassify to capture all 25 simulated non-noise features. Figure 4.2 shows the iClassify weightings for these 25 features, which can be interpreted as feature contributions to the model, as well as averaged coefficients for each feature that was included in the lasso model in three-fold cross-validation at least 50% of the time. While the features the Lasso did select had coefficients that matched the direction of the iClassify weightings and had the same relative magnitude, only nine of the 25 non-noise features were selected by Lasso at least

Figure 4.2: Varying γ effects simulation: feature selection



half the time—a stark illustration of the high false positive rate, and a likely explanation for the lower classification accuracy.

4.1.2 Integrative vs. single platform comparison

We were also interested in whether the model that combined the three data types improves prediction accuracy over models that only included one data type at a time, as we would expect from the reports of [Fuchs *et al.*, 2013] and [Daemen *et al.*, 2009] among others. Table 4.8 demonstrates that in the simulation setting the prediction accuracy of a combined model is better than the prediction accuracy of its best-performing single data type. Additional, we see that the prediction accuracy of a one-type model increases as γ increases.

Table 4.8: Varying $\boldsymbol{\gamma}$ effects simulation: Combined data types vs. single data type

Data Type	True γ	Estimated γ	iClass
Combined			0.892
$\boldsymbol{\gamma}_{1,1}$	2.5	2.45	
$\boldsymbol{\gamma}_{1,2}$	1.25	1.29	
$\boldsymbol{\gamma}_{1,3}$	0.75	0.69	
$\boldsymbol{\gamma}_{strong}$	2.5	2.53	0.867
$\boldsymbol{\gamma}_{medium}$	1.25	1.25	0.708
$\boldsymbol{\gamma}_{weak}$	0.75	0.80	0.611

4.2 Simulations with covariate $\times$ genomic interaction

An important feature of our method is its ability to incorporate covariates and model genomic $\times$ covariate interaction.

As illustrated in Section 3.5, the inclusion of genomic $\times$ covariate interaction affects the distributions of the latent variables Z_{it}:

$$Z_{it}|D_i = d, E_i = e \sim N(\mu_{d\mathbf{e}t}, \sigma^2), \mu_{d\mathbf{e}t} = \gamma_{0t} + \gamma_{1t}d + \gamma_{2t}e + \boldsymbol{\gamma}_3^T e^{\times} d, d = 0, 1.$$

To assess how our method performs under this type of model, we simulated datasets incorporating interactions and different strengths and prevalences. Building from Scenarios A and D from Table 4.4, we included a covariate risk factor with 80% prevalence in the dataset (OR$\approx$3.86) and a true γ_{3t} interaction coefficient of magnitude 1 or -1 in both scenarios. Tables 4.9 and 4.11 show parameter estimation for these datasets both accounting for and not accounting for the simulated covariate x genomic interaction. We also simulated datasets with covariate risk factors with 50% and 20% prevalence, and tables 4.10 and 4.12 summarize prediction accuracy for all cases.

Across all scenarios and models, prediction accuracy is strongly affected by the direction

Table 4.9: Scenario A with covariate/genomic interaction: parameter estimation

	Positive effect							Negative effect						
	Correct modeling				Naive modeling			Correct modeling				Naive modeling		
Para	TRUE	Est	Bias	MSE	Est	Bias	MSE	TRUE	Est	Bias	MSE	Est	Bias	MSE
$\boldsymbol{\beta}_{1,1}$	1.5	1.48	-0.02	0.01	1.61	0.11	0.01	1.5	1.42	-0.08	0.01	1.5	0	0.13
$\boldsymbol{\beta}_{2,1}$	1	0.99	-0.01	0.01	1.08	0.08	0.01	1	0.95	-0.05	0.01	1.01	0.01	0.06
$\boldsymbol{\beta}_{3,1}$	0	0	0	0	0	0	0	0	0.01	0.01	0	0.01	0.01	0
$\boldsymbol{\beta}_{4,1}$	0	-0.01	-0.01	0	-0.01	-0.01	0	0	0	0	0	0	0	0
$\boldsymbol{\beta}_{5,1}$	1.5	1.49	-0.01	0.01	1.62	0.12	0.01	1.5	1.41	-0.09	0.01	1.49	-0.01	0.13
$\boldsymbol{\beta}_{1,2}$	1	0.97	-0.03	0.01	1	0	0.01	1	0.98	-0.02	0.01	0.98	-0.02	0.13
$\boldsymbol{\beta}_{2,2}$	0	0	0	0	0	0	0	0	0.01	0.01	0	0	0	0
$\boldsymbol{\beta}_{3,2}$	0	0	0	0	0	0	0	0	0.01	0.01	0.01	0	0	0.01
$\boldsymbol{\beta}_{4,2}$	1.5	1.38	-0.12	0.01	1.4	-0.1	0.02	1.5	1.36	-0.14	0.01	1.3	-0.2	0.23
$\boldsymbol{\beta}_{5,2}$	0	0	0	0	0	0	0	0	0.01	0.01	0	0.01	0.01	0
$\boldsymbol{\gamma}_{1,1}$	1.5	1.48	-0.02	0.19	2.54	1.04	0.03	1.5	1.65	0.15	0.17	0.27	-1.23	0.01
$\boldsymbol{\gamma}_{2,1}$	1.5	1.48	-0.02	0.07				-1.5	-1.62	-0.12	0.08			
$\boldsymbol{\gamma}_{3,1}$	1	1.08	0.08	0.23				-1	-1.12	-0.12	0.23			
$\boldsymbol{\gamma}_{1,2}$	1.5	1.68	0.18	0.25	2.85	1.35	0.1	1.5	1.59	0.09	0.24	0.26	-1.24	0.02
$\boldsymbol{\gamma}_{2,2}$	1.5	1.63	0.13	0.09				-1.5	-1.66	-0.16	0.09			
$\boldsymbol{\gamma}_{3,2}$	1	0.98	-0.02	0.23				-1	-1.08	-0.08	0.28			

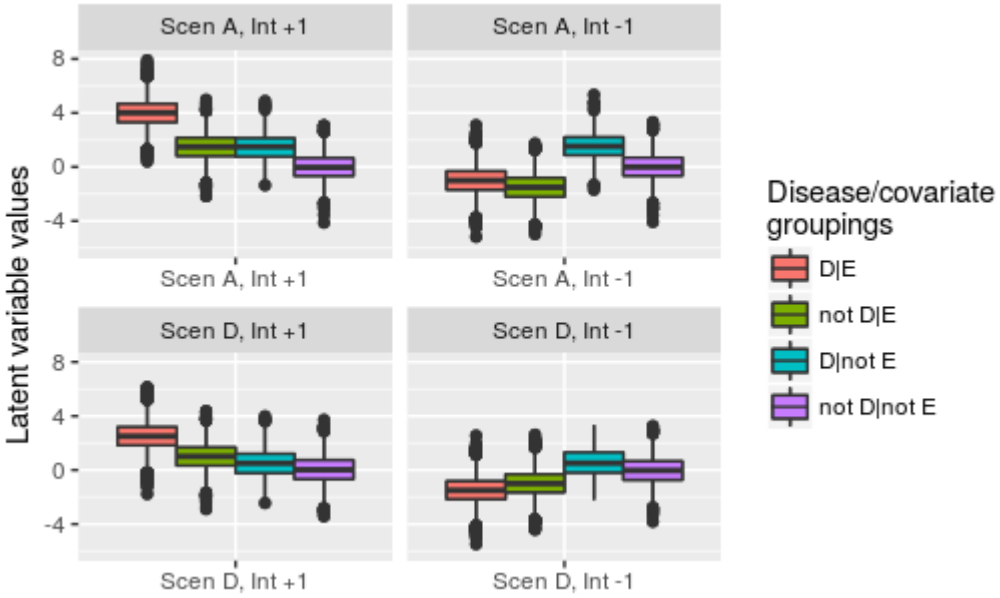
of the interaction term. When the interaction coefficients are positive (in the same direction as the genomic coefficients), we see a significant increase in prediction accuracy over simpler Scenarios A and D (Table 4.3), while negative interaction coefficients lead to prediction accuracies that are similar to (in the case of the weak interaction effects Scenario D) or worse than (in the case of strong interaction effects Scenario A) the simpler scenarios. LDA prediction accuracy is lower but follows the same pattern.

Figure 4.3 illustrates why scenarios with negative interaction coefficients (or more generally scenarios that have interaction effects in the opposite direction of genomic effects) have lower prediction accuracies. It shows distributions of the first simulated latent variable $\mathbf{Z}_{i1}$ for Scenarios A and D with interaction effects +1 and -1.

Table 4.10: Scenario A with covariate/genomic interaction: prediction accuracy

Covariate prevalence	Interaction effect	$D\|\boldsymbol{E}$ only	Full analysis		Naive analysis	
		iClassify	iClasssify	LDA	iClasssify	LDA
0.8	1	0.94	0.92	0.87	0.9	0.89
	-1	0.67	0.66	0.61	0.52	0.51
0.5	1	0.94	0.88	0.86	0.82	0.81
	-1	0.74	0.73	0.73	0.51	0.51
0.2	1	0.94	0.85	0.82	0.81	0.79
	-1	0.78	0.79	0.78	0.67	0.63

Figure 4.3: Latent variables by scenario and interaction effect sign



We recall from Section 3.5 that prediction accuracy in the presence of a covariate risk factor is dependent on the conditional distribution of latent variables, $f(\mathbf{Z}_i^{(r)}|D^* = 1, \boldsymbol{E}_i)$ and $f(\mathbf{Z}_i^{(r)}|D^* = 0, \boldsymbol{E}_i)$. The greater the separation between these two distributions of $\mathbf{Z}_i$, the more accurate $P(D^* = 1|\mathbf{X}, \boldsymbol{E})$ will be. In Scenario A, with an interaction coefficient of $+1$ (upper left-hand panel of Figure 4.3), we see the largest separation between the distribution of $D = 1|\boldsymbol{E}$ and $D = 0|\boldsymbol{E}$ with distribution medians differing by $\approx$2.5. However, when the interaction coefficient is -1 (upper right hand panel), we see a tightening of

Table 4.11: Scenario D with covariate/genomic interaction: parameter estimation

	Positive effect							Negative effect						
	Correct modeling				Naive modeling			Correct modeling				Naive modeling		
Para	TRUE	Est	Bias	MSE	Est	Bias	MSE	TRUE	Est	Bias	MSE	Est	Bias	MSE
$\beta_{1,1}$	-1	-0.96	0.04	0.01	-1.05	-0.05	0.01	-1	-0.95	0.05	0.01	-1.07	-0.07	0.01
$\beta_{2,1}$	0.5	0.48	-0.02	0.01	0.53	0.03	0.01	0.5	0.48	-0.02	0	0.54	0.04	0.01
$\beta_{3,1}$	0	-0.01	-0.01	0	-0.01	-0.01	0	0	0	0	0	0	0	0
$\beta_{4,1}$	0	-0.01	-0.01	0	-0.01	-0.01	0	0	0.01	0.01	0	0.01	0.01	0.01
$\beta_{5,1}$	-1	-0.98	0.02	0.01	-1.07	-0.07	0.01	-1	-0.95	0.05	0.01	-1.08	-0.08	0.01
$\beta_{1,2}$	-0.5	-0.5	0	0.01	-0.54	-0.04	0.01	-0.5	-0.53	-0.03	0.01	-0.62	-0.12	0.01
$\beta_{2,2}$	0	0	0	0.01	0	0	0.01	0	0.01	0.01	0.01	0.01	0.01	0.01
$\beta_{3,2}$	0	0	0	0.01	0	0	0.01	0	0	0	0.01	0	0	0.01
$\beta_{4,2}$	1	0.91	-0.09	0.01	0.92	-0.08	0.01	1	0.84	-0.16	0.01	0.86	-0.14	0.01
$\beta_{5,2}$	0	0	0	0	0	0	0.01	0	0	0	0.01	0	0	0.01
$\gamma_{1,1}$	0.5	0.53	0.03	0.26	1.54	1.04	0.04	0.5	0.56	0.06	0.24	-0.56	-1.06	0.03
$\gamma_{2,1}$	1	1.04	0.04	0.09				-1	-1.04	-0.04	0.07			
$\gamma_{3,1}$	1	1.03	0.03	0.26				-1	-1.09	-0.09	0.29			
$\gamma_{1,2}$	0.5	0.48	-0.02	0.31	1.64	1.14	0.05	0.5	0.49	-0.01	0.33	-0.64	-1.14	0.04
$\gamma_{2,2}$	1	1.07	0.07	0.1				-1	-1.1	-0.1	0.12			
$\gamma_{3,2}$	1	1.1	0.1	0.36				-1	-1.05	-0.05	0.45			

the distance between the distributions, whose medians now only differ by 0.5. The same dynamic is at work in Scenario D, which has less of a separation in the distributions in the presence of a positive interaction effect (lower left panel) than Scenario A, where distribution medians differ by ≈1.5, and an actual change in direction in the median difference of the distributions in the presence of a negative interaction effect (lower right panel).

Prevalence of the covariate risk factor plays an important role in prediction accuracy as well. With a positive interaction effect, prediction accuracy increases with prevalence in all cases. In the strong effects Scenario A with a negative interaction effect, prediction accuracy decreases linearly with prevalence. In the weaker effects Scenario D with a negative interaction effect, prediction accuracy is uniformly similar to the simpler scenario.

Table 4.12: Scenario D with covariate/genomic interaction: prediction accuracy

Covariate prevalence	Interaction effect	$D\|\boldsymbol{E}$ only iClassify	Full analysis iClasssify	Full analysis LDA	Naive analysis iClasssify	Naive analysis LDA
0.8	1	0.82	0.77	0.75	0.78	0.77
	-1	0.70	0.62	0.62	0.62	0.58
0.5	1	0.84	0.72	0.7	0.73	0.71
	-1	0.81	0.64	0.65	0.58	0.55
0.2	1	0.87	0.64	0.6	0.65	0.63
	-1	0.88	0.62	0.6	0.49	0.49

Prevalence does not have an effect on parameter estimation.

The inclusion of a covariate and interaction term in the model seems to have little effect on the bias but a noticeable effect on the MSE of $\boldsymbol{\gamma}$ estimation. Notably, even if the prediction rate is relatively low due to opposite signs in genomic and covariate interaction effects, as in Scenario A (Table 4.9) or D (4.11) with covariate prevalence 0.8, we can see from Tables 4.10 and 4.12 that our method still performs reasonable estimation of $\boldsymbol{\gamma}$, thus allowing for a way to assess whether interaction may be present.

When the analysis does not take interaction into account, prediction accuracy stays similar to the analysis that accounts for interaction, but only when the interaction effect is in the same direction as the genomic effect (i.e. positive/positive in this case). Presumably this is because the positive shift in latent variables $\mathbf{Z}_{it}$ results in larger estimations for γ_t. Conversely, when the analysis does not account for interaction, and the interaction effects are in the opposite direction from the genomic effect, prediction accuracies are significantly decreased.

We also confirmed that iClassify was able to properly perform parameter estimation in the case where there was no underlying interaction effect. Table 4.13 shows that iClassify estimates the interaction coefficients $\boldsymbol{\gamma}$ to be close to 0 both where there is an underyling

Table 4.13: Null interaction scenarios: parameter estimation

	Scenario A								Scenario D							
	W/o covariate effect				With covariate effect				W/o covariate effect				With covariate effect			
Para	TRUE	Est	Bias	MSE	TRUE	Est	Bias	MSE	TRUE	Est	Bias	MSE	TRUE	Est	Bias	MSE
$\beta_{1,1}$	1.5	1.47	-0.03	0.01	1.5	1.45	-0.05	0.01	-1	-0.84	0.16	0.09	-1	-0.97	0.03	0.01
$\beta_{2,1}$	1	0.99	-0.01	0.01	1	0.96	-0.04	0.01	0.5	0.42	-0.08	0.03	0.5	0.48	-0.02	0.01
$\beta_{3,1}$	0	0	0	0	0	-0.01	-0.01	0	0	0.01	0.01	0	0	-0.01	-0.01	0.01
$\beta_{4,1}$	0	0.01	0.01	0	0	-0.01	-0.01	0	0	0	0	0.01	0	0	0	0.01
$\beta_{5,1}$	1.5	1.47	-0.03	0.01	1.5	1.45	-0.05	0.01	-1	-0.82	0.18	0.09	-1	-0.96	0.04	0.01
$\beta_{1,2}$	1	0.99	-0.01	0.01	1	0.98	-0.02	0.01	-0.5	-0.43	0.07	0.05	-0.5	-0.51	-0.01	0.01
$\beta_{2,2}$	0	0	0	0	0	0.01	0.01	0	0	0	0	0.01	0	-0.02	-0.02	0.01
$\beta_{3,2}$	0	0.01	0.01	0	0	0	0	0	0	-0.01	-0.01	0	0	0.01	0.01	0.01
$\beta_{4,2}$	1.5	1.42	-0.08	0.01	1.5	1.41	-0.09	0.01	1	0.65	-0.35	0.1	1	0.89	-0.11	0.01
$\beta_{5,2}$	0	-0.01	-0.01	0.01	0	0.02	0.02	0	0	0.01	0.01	0.01	0	0	0	0.01
$\gamma_{1,1}$	1.5	1.56	0.06	0.17	1.5	1.52	0.02	0.17	0.5	0.54	0.04	0.21	0.5	0.57	0.07	0.17
$\gamma_{2,1}$	0	-0.01	-0.01	0.06	1.5	1.55	0.05	0.07	0	0.04	0.04	0.06	1	1.06	0.06	0.1
$\gamma_{3,1}$	0	0	0	0.2	0	0.03	0.03	0.18	0	-0.03	-0.03	0.2	0	-0.07	-0.07	0.19
$\gamma_{1,2}$	1.5	1.53	0.03	0.19	1.5	1.58	0.08	0.2	0.5	0.51	0.01	0.35	0.5	0.58	0.08	0.3
$\gamma_{2,2}$	0	0.02	0.02	0.06	1.5	1.54	0.04	0.07	0	0.11	0.11	0.09	1	1.15	0.15	0.11
$\gamma_{3,2}$	0	0.01	0.01	0.2	0	-0.04	-0.04	0.26	0	-0.1	-0.1	0.42	0	-0.08	-0.08	0.37

Table 4.14: Null interaction scenarios: prediction accuracy

Covariate	Scenario A			Covariate	Scenario D		
effect	iClasssify	LDA	LR	effect	iClasssify	LDA	LR
0	0.82	0.80	0.76	0	0.62	0.61	0.61
1.5	0.82	0.75	0.75	1	0.62	0.61	0.60

covariate effect and where there is neither a covariate nor an interaction effect. Table 4.14 shows that, as expected, prediction accuracy does not change relative to the simple scenarios A and D.

4.3 Sensitivity Analysis: Simulations with covariate risk factor only

Though the association between covariate risk factor and disease status is not of primary interest, we also performed a set of simulations to examine if including a covariate risk factor alone had any effect on estimation and prediction.

As illustrated in Section 3.5, the inclusion of one covariate risk factor, e, affects the

Table 4.15: Scenario A with covariate risk factor only: parameter estimation

	Positive effect							Negative effect						
	Correct modeling				Naive modeling			Correct modeling				Naive modeling		
Para	TRUE	Est	Bias	MSE	Est	Bias	MSE	TRUE	Est	Bias	MSE	Est	Bias	MSE
$\boldsymbol{\beta}_{1,1}$	1.5	1.38	-0.12	0.02	1.54	0.04	0.01	1.5	1.42	-0.08	0.01	1.57	0.07	0.01
$\boldsymbol{\beta}_{2,1}$	1	0.93	-0.07	0.01	1.04	0.04	0	1	0.95	-0.05	0	1.06	0.06	0
$\boldsymbol{\beta}_{3,1}$	0	-0.01	-0.01	0.01	0	0	0	0	0.01	0.01	0	0.01	0.01	0
$\boldsymbol{\beta}_{4,1}$	0	0	0	0.01	0.01	0.01	0	0	0	0	0	0.01	0.01	0
$\boldsymbol{\beta}_{5,1}$	1.5	1.4	-0.1	0.02	1.54	0.04	0.01	1.5	1.43	-0.07	0.01	1.58	0.08	0.01
$\boldsymbol{\beta}_{1,2}$	1	0.95	-0.05	0.01	1.04	0.04	0.01	1	0.99	-0.01	0.01	1.1	0.1	0.01
$\boldsymbol{\beta}_{2,2}$	0	0	0	0.01	0	0	0	0	0.02	0.02	0	0.02	0.02	0.01
$\boldsymbol{\beta}_{3,2}$	0	-0.01	-0.01	0.01	0	0	0.01	0	0	0	0	0	0	0
$\boldsymbol{\beta}_{4,2}$	1.5	1.33	-0.17	0.02	1.41	-0.09	0.01	1.5	1.38	-0.12	0.01	1.5	0	0.01
$\boldsymbol{\beta}_{5,2}$	0	0.01	0.01	0.01	0	0	0	0	-0.01	-0.01	0	-0.01	-0.01	0
$\boldsymbol{\gamma}_{1,1}$	1.5	1.63	0.13	0.07	1.76	0.26	0.03	1.5	1.61	0.11	0.04	1.17	-0.33	0.03
$\boldsymbol{\gamma}_{2,1}$	1.5	1.58	0.08	0.12				-1.5	-1.61	-0.11	0.05			
$\boldsymbol{\gamma}_{1,2}$	1.5	1.64	0.14	0.11	1.89	0.39	0.04	1.5	1.62	0.12	0.05	1.19	-0.31	0.04
$\boldsymbol{\gamma}_{2,2}$	1.5	1.64	0.14	0.15				-1.5	-1.64	-0.14	0.06			

Table 4.16: Scenario A with covariate risk factor only: prediction accuracy

Covariate	Full analysis		Covariate-naive analysis	
effect	iClasssify	LDA	iClasssify	LDA
Positive	0.82	0.82	0.82	0.80
Negative	0.83	0.82	0.73	0.71

distributions of the latent variables Z_{it}. Without an interaction term:

$$Z_{it}|D_i = d, E_i = e \sim N(\mu_{d\boldsymbol{e}t}, \sigma^2), \mu_{d\boldsymbol{e}t} = \gamma_{0t} + \gamma_{1t}d + \gamma_{2t}e, d = 0, 1.$$

We used Scenarios A and D from Table 4.4 and added in a covariate risk factor with 80% prevalence in the dataset (OR≈3.86) and a true γ_{2t} covariate coefficient of magnitude 1.5 in Scenario A and 0.5 in Scenario D. We looked at both a positive and negative covariate effect.

Table 4.15 shows the estimation for scenario A modeled correctly (including a covariate

Table 4.17: Scenario D with covariate risk factor only: parameter estimation

	Positive effect							Negative effect						
	Correct modeling				Naive modeling			Correct modeling				Naive modeling		
Para	TRUE	Est	Bias	MSE	Est	Bias	MSE	TRUE	Est	Bias	MSE	Est	Bias	MSE
$\boldsymbol{\beta}_{1,1}$	-1	-0.98	0.02	0.01	-1.05	-0.05	0.01	-1	-0.93	0.07	0.03	-0.87	0.13	0.13
$\boldsymbol{\beta}_{2,1}$	0.5	0.48	-0.02	0.01	0.52	0.02	0.01	0.5	0.47	-0.03	0.01	0.44	-0.06	0.04
$\boldsymbol{\beta}_{3,1}$	0	0	0	0.01	0	0	0.01	0	0	0	0.01	0.01	0.01	0.01
$\boldsymbol{\beta}_{4,1}$	0	-0.01	-0.01	0.01	-0.01	-0.01	0.01	0	-0.01	-0.01	0	0	0	0
$\boldsymbol{\beta}_{5,1}$	-1	-0.97	0.03	0.01	-1.04	-0.04	0.01	-1	-0.95	0.05	0.03	-0.89	0.11	0.13
$\boldsymbol{\beta}_{1,2}$	-0.5	-0.53	-0.03	0.01	-0.58	-0.08	0.01	-0.5	-0.5	0	0.02	-0.42	0.08	0.09
$\boldsymbol{\beta}_{2,2}$	0	0.01	0.01	0.01	0.01	0.01	0.01	0	0.01	0.01	0.01	0	0	0.01
$\boldsymbol{\beta}_{3,2}$	0	0.01	0.01	0.01	0.01	0.01	0.01	0	0	0	0.01	0.01	0.01	0.01
$\boldsymbol{\beta}_{4,2}$	1	0.83	-0.17	0.01	0.84	-0.16	0.01	1	0.77	-0.23	0.05	0.58	-0.42	0.16
$\boldsymbol{\beta}_{5,2}$	0	-0.01	-0.01	0.01	-0.02	-0.02	0.01	0	0.02	0.02	0.01	0.02	0.02	0.01
$\boldsymbol{\gamma}_{1,1}$	0.5	0.52	0.02	0.03	0.68	0.18	0.03	0.5	0.53	0.03	0.03	0.32	-0.18	0.02
$\boldsymbol{\gamma}_{2,1}$	1	1.05	0.05	0.06				-1	-1	0	0.07			
$\boldsymbol{\gamma}_{1,2}$	0.5	0.52	0.02	0.04	0.71	0.21	0.04	0.5	0.49	-0.01	0.04	0.31	-0.19	0.02
$\boldsymbol{\gamma}_{2,2}$	1	1.1	0.1	0.09				-1	-0.98	0.02	0.16			

Table 4.18: Scenario D with covariate risk factor only: prediction accuracy

Covariate	Full analysis		Covariate-naive analysis	
effect	iClasssify	LDA	iClasssify	LDA
Positive	0.62	0.61	0.63	0.6
Negative	0.62	0.6	0.54	0.52

risk factor) and naively (without accounting for covariate risk factor). The correct modeling produces reasonable estimation for both the scenarios with positive and negative covariate effects. The naive modeling of the scenario with the positive covariate effect produces positive bias in the genomic coefficient $\boldsymbol{\gamma}$ estimation, whereas the naive modeling of the scenario with the negative covariate effect produces significantly underestimated $\boldsymbol{\gamma}$ estimates.

These $\boldsymbol{\gamma}$ estimates help explain the prediction accuracies in 4.16. Prediction accuracy does not decrease in the model that does not take the positive covariate effect into account

because the increase in μ_{det} is picked up by the overestimation of γ_{1t}. Conversely, prediction accuracy of the model that does not take the negative covariate effect into account drops significantly as the genomic effects γ_{1t} are significantly underestimated.

Tables 4.17 and 4.18 show the analogous sets of results for scenario D, with similar conclusions.

Chapter 5

iClassify: Application

5.1 TCGA Ovarian Cancer data set

The data to which we applied our method comes from The Cancer Genome Atlas Research Network [The Cancer Genome Atlas Research Network, 2011]. It is comprised of high grade serous ovarian cancer tumors that were surgically resected before treatment with platinum chemotherapy. The paper reported that 31% of these patients were resistant to chemotherapy and experienced disease progression within 6 months of completing treatment.

The number of features and samples available in the TCGA set are detailed in Table 5.1, as well as the number of samples used in our integrative analyses. We included in our basic analysis set those samples that had mRNA, methylation and miRNA profiling, as well as platinum resistance outcome data. We also performed two interaction analyses: one on a subset of samples that also had residual disease information, and one on a subset of samples that also had BRCA germline variant data.

Clinical characteristics available from TCGA included age, and tumor stage and grade. Table 5.2 shows the distribution of those clinical characteristics in the full clinical set, the dataset used for genomic-only analysis and the data set used for the BRCA interaction

Table 5.1: Platforms, features and datasets

Data type	Platform	Features	Samples
mRNA expression profiling	RNA-seq	11864	489
CpG DNA methylation	Illumina 27K	23665	489
miRNA expression profiling	miRNA-seq	781	589
Residual disease			432
BRCA germline mutation	Whole-exome		314
Platinum status			287

Analysis datasets	Samples	# Platinum resistant
Integrative genomic analysis	285	90 (0.32)
Interaction analysis w/residual disease	259	85 (0.33)
Interaction analysis w/BRCA germline	187	59 (0.32)

anaysis. In all cases, the tumors are primarily Stage III and IV, and Grade 3. The median age is stable at 59-60.

5.2 Imbalanced data considerations

Table 5.1 details the prevalence of the outcome in our analysis datasets, with 32-33% of the tumors showing platinum resistance. This imbalance creates a challenge for analysis.

In Chapter 4, we simulated equal numbers of cases and controls for all of our scenarios, and used overall classification accuracy as our performance measure. When faced with the TCGA platinum resistance data, however, it became clear that this measure was not appropriate, as the minority class (platinum resistance) makes only a minor contribution to the overall accuracy relative to the majority class (platinum sensitive). To wit, if a method classified all tumors as sensitive, i.e. misclassified every resistant tumor, that would still

Table 5.2: Clinical characteristics

Characteristic	Full clinical set n=488	Genomic-only analysis n=285	BRCA interaction analysis n=187
Age (Median, range)	59 (27, 87)	59 (30.5,87)	60 (36, 87)
Missing (n)	11 (2%)	2 (1%)	1 (1%)
Stage (n, %)			
II	24 (5%)	13 (5%)	5 (3%)
III	381 (78%)	231 (81%)	155 (83%)
IV	79 (16%)	41 (14%)	27 (14%)
NA	4 (1%)	0 (0%)	0 (0%)
Grade (n, %)			
G2	57 (12%)	38 (13%)	14 (8%)
G3	420 (86%)	241 (85%)	168 (90%)
NA	11 (2%)	6 (2%)	5 (3%)

give an overall classification accuracy of 68%. In fact, this is exactly how the comparative method of penalized logistic regression performed, thus prompting us to modify the comparative method to weighted penalized logistic regression, with weights determined by inverse prevalence in data set, e.g. 1/.32 for platinum resistant and 1/.68 for platinum sensitive.

Further, if our interest is in identifying those tumors which have a higher probability of becoming platinum resistant, there is actually a higher cost to a false negative in the resistant class. So in the tradeoff that always exists between sensitivity and specificity, we prioritize sensitivity though are still interested in reasonable specificity as well.

5.3 Pre-screening features

We pre-processed genomic features by centering and normalizing. Methylation missing data was imputed and then batch correction was performed.

It has been shown that a pre-screening strategy based on feature correlation with outcome to reduce the dimension of the data set to a moderate scale can enhance finite sample model performance and reduce computational cost [Fan and Lv, 2008].

We used an independent ovarian cancer gene expression data set with platinum resistance outcomes [Dressman *et al.*, 2007] to select mRNA gene expression features most strongly associated with platinum resistance. miRNA features were selected from a literature review of miRNAs associated with platinum resistance outcomes [Mahdian-shakib *et al.*, 2016]. For methylation, we found CpG sites located in the selected mRNA features and included those that showed nominal correlation. This led to our analysis feature set of 1039 mRNA features, 416 methylation features, and 28 miRNA features.

5.4 Genomic-only Analysis

Table 5.3: Genomic-only analysis: iClassify and Lasso classification accuracy

	λ	# RNA	# Meth	# miRNA	Total #	Sensitivity	Specificity	Class acc
iClassify	0	1039	416	28	1483	0.541	0.591	0.575
	0.05	700	288	17	1005	**0.543**	**0.593**	**0.577**
	0.1	386	95	8	489	0.523	0.627	0.594
	0.15	172	0	3	175	0.473	0.657	0.599
	0.2	65	0	0	65	0.454	0.657	0.593
Lasso					57 (avg)	0.405	0.629	0.558

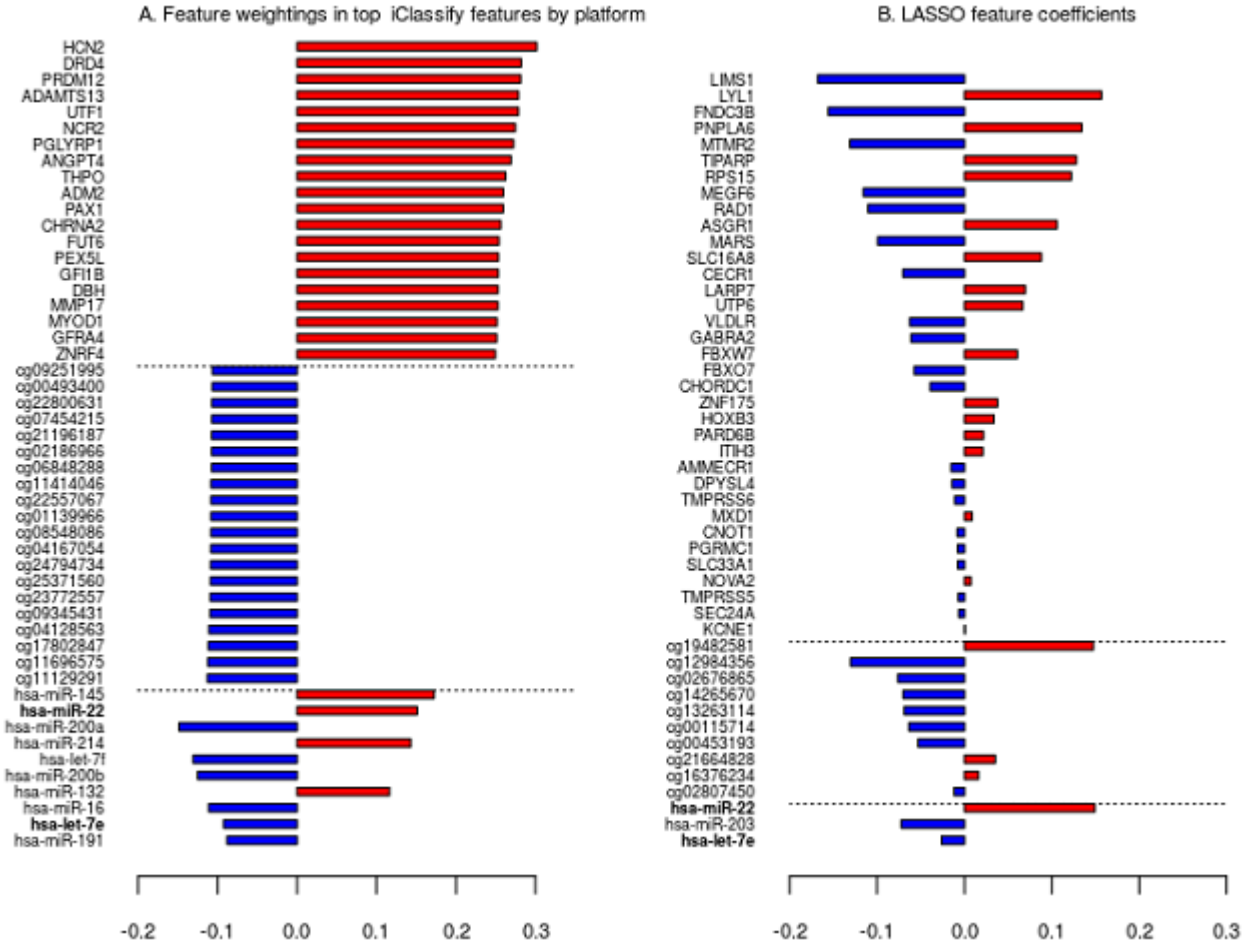
We analyzed this data set with iClassify and used the hard thresholding algorithm

with three-fold cross validation over 100 replicates to find the threshold with the highest classification accuracy. We show results at various thresholds in Table 5.3 along with results from weighted cross-validated Lasso, using classification accuracy as the loss measure. The optimal iClassify solution is at the λ=0.05 threshold, where it achieves the best performance in both sensitivity and specificity with an overall classification accuracy of 57.7%. The sensitivity at all iClassify thresholds is higher than that of the Lasso, with the optimal iClassify solution improving on Lasso's sensitivity by almost 14%. We chose to present the results in terms of sensitivity and specificity at the optimal λ for each method because of their clinical significance. Another way to compare the two methods would have been to look at other measures like AUC over the range of each method's λ values.

The difference in feature selection between iClassify and the Lasso is striking, and recapitulates what we saw in the simulations on a larger scale. The optimal iClassify solution chooses 1005 features overall while the Lasso solution chooses 57 on average. The relative insensitivity and lower prediction accuracy of the penalized logistic regression solution suggest that the Lasso solution may have a high false negative rate, and that there are likely many genomic features with small effects contributing to the optimal model, contrasting the typical sparsity assumption that a small subset of features are relevant for prediction, which Lasso and other penalized methods rely on. Further, we suspect that many of these features are correlated, an idea supported by the fact that performing an analysis with L2-penalized logistic regression ("ridge regression") yields increases in overall classification accuracy (0.56) but still underperforms *iClassify*.

Figure 5.1 shows the top features in each platform for iClassify with iClassify weightings ($\boldsymbol{\beta} * \boldsymbol{\gamma}$ estimates) and the selected features for a Lasso solution. All 48 features chosen by the Lasso are also selected by iClassify, leaving a pool of 950 iClassify-selected features,

Figure 5.1: Genomic-only analysis: comparative feature selection



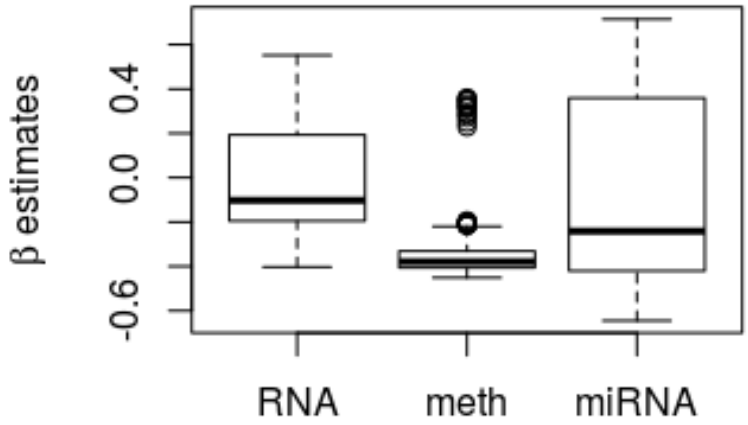
many of which could be false negatives for the Lasso. None of the Lasso-selected RNA and methylation features appear in the iClassify top-weighted features list. iClassify feature weightings for RNA features are larger than for features in other data types, reflecting the larger γ coefficient estimate for RNA relative to methylation and miRNA (see Table 5.4). In fact, RNA features comprise the top 88 features ranked by iClassify feature weightings overall. Notably, hsa-miR-22, one of the three Lasso-selected miRNA features, is the 89th top-ranked feature for iClassify. Top iClassify features in both the RNA and methylation platforms have very similar feature weightings in both magnitude and direction, illustrating

the strongly correlated nature of the features that contribute to the model.

Table 5.4: Genomic-only analysis: parameter estimation

	# features	$\hat{\gamma}$	γ bootstrap 95% CI	γ bootstrap SD	$\boldsymbol{\beta}$ boostrap SD
RNA	700	0.55	(0.19, 0.98)	0.23	0.04
meth	288	0.23	(-0.26, 0.77)	0.29	0.05
miRNA	17	0.23	(-0.04, 0.49)	0.14	0.13

Figure 5.2: Genomic-only analysis: $\boldsymbol{\beta}$ estimation



The parameter estimates for $\boldsymbol{\gamma}$ are in Table 5.4 along with 95% bootstrap confidence intervals. These estimates quantify the degree of association of each individual data type with outcome. We constrain the overall γ estimate to be positive, and see that RNA appears to have the largest effect on the model and methylation and miRNA have smaller effects. As we saw in simulations that $\boldsymbol{\beta}$ MSE values were significantly smaller than $\boldsymbol{\gamma}$ MSE values (see tables 4.1, 4.3, and 4.5), here we see that bootstrap standard deviations are similarly smaller for $\boldsymbol{\beta}$ than for $\boldsymbol{\gamma}$. Only $\gamma_{1,1}$, associated with RNA, has a bootstrap confidence interval that does not cross 0.

The parameter estimates for $\boldsymbol{\beta}$ are plotted in Figure 5.2. Notable here is that most of the 288 methylation features have negative $\beta_{2,p}$ values, with lower average values in

the platinum-resistance group than the platinum-sensitive group. This finding of relatively hypomethylated features in the platinum-resistance group dovetails with the report in [Yu *et al.*, 2011] that platinum-resistant ovarian cancer cells show a global decrease in methylation of CpG islands. Other reports in the literature, including [Zeller *et al.*, 2012] find increased methylation in platinum-resistant cell lines, but importantly the assays were performed after exposure to platinum whereas our cohort was assayed after surgery but before treatment. Also we see that the miRNA β values have a wider spread than the other platforms.

5.4.1 Single platform vs Integrative Analysis

Table 5.5: Genomic-only analysis: Single platform vs Integrative analysis

Data type	$\hat{\gamma}$	Sensitivity	Specificity	Class acc
Combined		0.543	0.593	0.577
RNA	0.55			
meth	0.23			
miRNA	0.23			
RNA	0.70	0.476	0.662	0.604
meth	0.36	0.535	0.493	0.506
miRNA	0.25	0.159	0.875	0.649

Table 5.5 shows results of integrative analysis vs. single platform. The combined integrative analysis is more effective than any of the single analyses alone in terms of balancing sensitivity and specificity.

Interestingly, the single platform analysis reveals that it is the methylation platform that seems to be driving the higher levels of sensitivity. However, methylation alone results in the lowest specificity. The integrated analysis both gives the highest sensitivity and a 10% increase in specificity over methylation's specificity level. miRNA turns out to be the

least effective single platform by far, with extremely low sensitivity.

5.5 Interaction analysis

One of the main advantages of our method is its ability to effectively include an interaction term in a high dimensional data setting whereas in most high dimensional settings, testing for interaction is unwieldy and impractical. Here, we investigate potential interactions with two covariates, BRCA germline variants and residual disease.

5.5.1 BRCA germline mutation

As discussed earlier in Section 1.2.2.2, ovarian cancer cells deficient for BRCA1 and BRCA2 have been found to be more sensitive to cisplatin, and restoration of BRCA1 and BRCA2 expression has been found to increase resistance. Given this known interaction, our interest was to to investigate whether there might be other BRCA-interacting genomic partners.

Unfortunately, only 316 tumors in the TCGA ovarian cancer cohort had BRCA germline variant data available, and only 187 of those had platinum resistance information. Thus, our sample size was limited for this interaction analysis. Nevertheless, the results are of interest.

In order to be able to make a more meaningful comparison between theresults of the genomic-only model and the results of the model with the genomic x BRCA interaction term, we performed an *iClassify*genomic-only analysis on the subset of 187 tumors with BRCA variant data. Results are in Table 5.6.

In our dataset, BRCA germline variant alone shows a moderate but not statistically significant protective effect of BRCA mutation against resistance with RR=0.53, CI (0.23, 1.20) and classification accuracy of 0.462 (Table 5.7).

Table 5.6: Genomic-only subset analysis: n=187

	λ	# RNA	# Meth	# miRNA	Total #	Sensitivity	Specificity	Class acc
iClassify	0	1039	416	28	1483	0.478	0.560	0.534
	0.05	502	233	16	751	**0.495**	**0.563**	**0.542**
	0.1	142	0	9	151	0.437	0.618	0.561
	0.15	40	0	1	41	0.388	0.642	0.562
hline Lasso					40 (avg)	0.379	0.617	0.542

Table 5.7: Classification accuracy: BRCA only and Genomic x BRCA interaction

	Threshold	RNA	meth	miRNA	Sensitivity	specificity	class acc
BRCA only					0.556	0.418	0.462
Genomic x BRCA interaction							
Combined genomic	0.00	1039	416	28	0.478	0.564	0.537
RNA only	0.00	1039			0.449	0.608	0.558
meth only	0.05		312		0.444	0.523	0.498
miRNA only	0.00			28	0.167	0.848	0.633

Table 5.7 also shows the results for the combined and invidividual genomic platform models that include a genomic $\times$ BRCA germline variant interaction term. They all show better overall classification accuracies than BRCA alone. There is not much that separates the prediction accuracies from the genomic-only subset analysis and the Genomic x BRCA interaction analysis. The genomic-only model has 2% higher sensitivity whiel the specificity and overall classification accuracies are almost equivalent. In Chapter 3 simulations, we did not see a scenario where a model with true interaction performed worse in classification accuracy than a model that did not account for the true interaction (see Tables 4.10 and 4.12). This suggests that significant interaction may not be present in the underlying true model.

However, we did see some scenarios of weak interaction where the overall classification accuracies were equivalent in the model with true interaction and the model that did not take interaction in account (the naive model). So it remains possible that interaction is present but that our sample size limits our power to detect it.

Table 5.8: BRCA interaction analysis: $\boldsymbol{\gamma}$ estimates and 95% bootstrap confidence intervals

	# feat	Genomic effect (γ_1)	Covariate effect (γ_2)	Interaction effect (γ_3)
RNA	1039	0.35 (-0.09, 1.25)	-0.41 (-1.31, 0.6)	-0.55 (-1.95, 0.57)
meth	416	0.17 (-0.47, 0.74)	-0.48 (-1.52, 0.7)	-0.77 (-5.75, 2.18)
miRNA	28	0.29 (-0.13, 0.57)	0.26 (-0.3, 0.81)	-0.05 (-0.79, 1.13)

The 95% bootstrap confidence intervals in Table 5.8 support the conclusion that interaction is not significant in this model, as the confidence intervals for interaction effects $\boldsymbol{\gamma}$ for all three platforms include 0. Again, the sample size is a limitation, and it is possible that with a larger dataset, we would be able to detect an interaction effect.

5.5.2 Residual disease

Residual disease after resection has been shown to be associated with worse prognosis in ovarian cancer, and also with platinum resistance. And, indeed, this is borne out in the TCGA dataset as mentioned in Section 1.2.2.2. Following the report in [Tucker *et al.*, 2014] that survival was significantly better for patients in this cohort with no residual disease compared to any residual disease at all, we used an any/no residual disease dichotomy when analyzing potential association between residual disease and platinum resistance.

The TCGA dataset had 259 samples with platinum resistance information, three genomic platform profiles, and residual disease dichotomized by no vs. any residual disease. Any residual disease increases the risk of platinum resistance with RR=2.33, 95% CI (1.25,

4.35). This strong effect yields surprisingly low overall classification accuracy of 0.478 and specificity of 0.316 but a notably high sensitivity of 0.811 (Table 5.9). We can thus infer that for sensitivity, clinical factors remain most influential.

Table 5.9: Classification accuracy: Residual disease only and Genomic x residual interaction

	Threshold	RNA	meth	miRNA	Sensitivity	specificity	class acc
Residual disease only					0.811	0.316	0.478
Genomic x Residual disease interaction							
Combined genomic	0.05	791	162	11	0.505	0.613	0.577
RNA	0.05	784			0.488	0.671	0.611
meth	0.0		416		0.490	0.493	0.492
miRNA	0.2			8	0.197	0.783	0.591

We did not find any reports in the literature that the effect of residual disease on drug response could be modified by genomic features. Thus, we would expect to see a null interaction effect. Results here were similar to the BRCA interaction results. Again, the interaction models all show better overall classification accuracies than residual disease alone. However, the highest sensitivity at the threshold of 0.05 was 51%, a 3% decline from the genomic-only model. Here too it is possible that sample size may play a role in this decrease.

Table 5.10: Residual interaction analysis: $\boldsymbol{\gamma}$ estimates and 95% bootstrap confidence intervals

	# feat	Genomic effect (γ_1)	Covariate effect (γ_2)	Interaction effect (γ_3)
RNA	791	1.3 (0.23, 2.89)	0.05 (-0.48, 0.62)	-0.82 (-2.35, 0.23)
meth	162	0.32 (-0.83, 1.59)	0.39 (-0.25, 1.06)	-0.13 (-1.48, 0.96)
miRNA	11	0.43 (-0.53, 1.55)	0.21 (-0.04, 0.56)	-0.22 (-1.3, 0.88)

95% bootstrap confidence intervals in Table 5.10 show, as we might expect, that the

confidence intervals for interaction effects $\boldsymbol{\gamma}$ for all three platforms include 0. Notably, in this case, the estimate for the RNA covariate effect $\gamma_{1,2}$ is close to 0 while the interaction effect $\gamma_{1,3}$ estimate is relatively high in the opposite direction.

5.6 Summary of analysis

We used our new method *iClassify* to perform classification on response to platinum therapy in the TCGA HGSOC cohort using genomic features from mRNA, methylation and miRNA assays. Overall we saw a crossvalidated classification accuracy of 58% with 54% sensitivity and 59% specificity. This is compared to the Lasso's overall classification accuracy of 56% with 41% sensitivity and 63%. Poorer performing methods are likely to have lower sensitivity and higher specificity in the context of unbalanced data, so Lasso's comparatively higher specificity is not surprising.

This analysis of TCGA data suggests that current knowledge of clinical and molecular factors are not sufficient to yield high sensitivity and specificity. As more comprehensive molecular studies are generated, larger sample sizes may allow us to make better conclusions. Nevertheless, our general framework is applicable to future datasets.

Additionally, we were able to gain understanding about different genomic data types' contributions to platinum resistance, and in particular our results suggest that methylation patterns may be particularly important in determining a tumor's potential for platinum resistance.

We also demonstrated the ability of *iClassify* to perform tests for genomic-covariate interaction, avoiding the challenge of multiple testing comparisons that are often faced in the high throughput genomic setting. Though the samples sizes in our dataset were a limitation, this capability has potential for future genomic analyses.

Chapter 6

Discussion and Future Research

In this dissertation, we investigated integrated genomic datasets with joint latent variable approaches. In these approaches, we introduce latent variables that we hypothesize to represent underlying factors that drive disease and that explain phenotypes manifested by thousands of genomic features.

We used an already existing method, *iCluster+*, to integratively cluster datasets from The Cancer Genome Atlas, with 4 or more genomic data types. *iCluster+* models latent variables across all data types simultaneously, and jointly models heterogenous data types through a diverse range of generalized linear models, thus accommodating the different scale and variance structures of the different data types. Through data-type-specific sparsity tuning parameters, it also allows for feature selection that takes into account the contribution of each data type to the model.

Through our analyes, we have demonstrated that *iCluster+* can detect clinically and prognostically meaningful subtypes of cancer. For example, prognostically differential mesothelioma subtypes discovered using *iCluster+* on a TCGA cohort were validated in two external datasets, suggesting potential for clinically relevant improved risk stratification. Integrative

clustering with *iCluster+* across 33 types of cancers yielded two mixed-cancer-type clusters that were enriched for immune-related signaling features, suggesting specific immune pathways as potential targets for new therapies.

For classification, we proposed a new model using a joint latent variable model, *iClassify*, for predicting binary treatment response outcome by integrating multiple data sets. In this model, latent variables represent underlying driving factors for each data type. There are several advantages to the proposed approach. First, multiple types of genomic features are integrated through a latent variable approach, which allows effective dimension reduction and can handle heterogenous data types of different scale and diverse variance structure. As our approach is not reliant on pre-existing genomic knowledge, it has the advantage of allowing for the discovery of previously unknown mechanisms.

Because the latent variables in our model are associated with the binary outcome though a linear model, we also have the flexibility to incorporate covariates and test for risk factor by genomic feature interaction, which is not straightforward in traditional methods but is of great interest in genomic investigations.

Our feature selection methodology allows for a systematic way of ranking the importance or contribution of a different data type on the overall model, and of an individual genomic feature, and is achieved by examining feature-specific effects through β_{jt} and data-type specific effects through γ_t.

We compared the proposed method for genomic-only data with penalized regression and LDA using simulated data sets. In simulations, *iClassify* outperformed classification accuracy of the other methods and minimized both the false positive rate and notably the high false negative rate of the Lasso. The improvement in particular of the false negative rate could have important clinical/biological implications. The comparatively high false

negative rate of Lasso feature selection may stem from the need to make decisions about which correlated features to exclude, when in fact the correlated features may each have clinical or biological relevance. In simulations and in the data analysis, penalized regression chose significantly fewer features for its optimized model. Additionally, our simulations confirm the improvement in prediction accuracy gained by combining multiple genomic datasets compared to genomic datasets of just one modality.

In our data analysis, we were able to measure performance through a combination of sensitivity and specificity and showed better results than weighted penalized logistic regression. In all cases, classification on the combination of data platforms performs better than on single platforms alone. Consistent with simulations, our feature selection methodology chose notably more features than penalized logistic regression, which may imply fewer false negatives, but would need to be confirmed in replication studies.

While in general the classification performance of platinum resistant/sensitive tumors in our data set did not reach high levels of sensitivity and specificity, the ability to estimate latent genomic effects provides a framework that can offer new perspectives and increased understanding of the data. For example, through our analyses, we were able to understand that the methylation platform was driving most of the sensitivity in our model in contrast to the miRNA platform, which had very low sensitivity, effects we would not have anticipated. Further, the capacity to estimate genomic x covariate interaction effects offers a potentially valuable approach to the multiple testing problem that has classically made testing for gene-environment interaction so impractical.

There are a few potential extensions to our method. First, *iClassify* can be extended to accommodate non-normal genomic factors, which will require changing equation (3.3) depending on the distribution of the features included. For example, to accommodate

mutation status in our method, equation (3.3) takes the following form for binary genomic variables: $\text{logit}\{\Pr(X_{ijt} = 1|\mathbf{Z}_{it})\} = \alpha_{jt} + \boldsymbol{\beta}_{jt}^T \mathbf{Z}_{it}$. Furthermore, it will be of interest to extend this method for time-to-event outcomes under cox models with random effects.

Another extension is to develop a systematic method of ranking the importance or contribution of a different data type on an individual subject. Currently our model provides the contribution of a data type to the overall model by the estimates of $\boldsymbol{\gamma}_t$, but it is also of interest to formalize a ranking of the importance of each genomic variable for an individual subject through the posterior distributions of Z_{it}, which could allow for a more personalized approach to treatment.

While the association between covariates and disease status was not of our primary interest in Equation (3.6), we understand the usefulness of including it in our joint likelihood model. On the prediction side, this would allow classification accuracy to increase when there is a direct covariate effect on disease, so that we could see the "additive" effect of covariate on prediction accuracy over genomic effect alone.

In practice, it is possible that not all genomic platforms are available on all subjects. For example, some subjects may only have mRNA and methylation. Our methods can analyze unbalanced data and include all available genomic data collected on a subject under the likelihood framework in a similar spirit to the mixed effects models for longitudinal data analysis.

On a practical level, we will make an R package of *iClassify* available for use by the research community. While the parallel computing algorithm we employ goes a long way towards making computation feasible, the computational burden of joint maximization and Monte-Carlo resampling remains a concern. We may remedy this by implementing a C++ version of *iClassify*. Another approach to easing computational burden would be to use

variational Bayesian inference ([Blei *et al.*, 2017]) to approximate probability rather than the Monte Carlo approach we currently employ.

Bibliography

[Adourian *et al.*, 2008] Aram Adourian, Ezra Jennings, Raji Balasubramanian, Wade M Hines, Doris Damian, Thomas N Plasterer, ..., and Ina Schuppe-Koistinen. Correlation network analysis for data integration and biomarker selection. *Molecular bioSystems*, 4:249–259, 2008.

[Agarwal and Kaye, 2003] Roshan Agarwal and Stan B. Kaye. Ovarian cancer: strategies for overcoming resistance to chemotherapy. *Nature Reviews Cancer*, 3(7):502–516, 2003.

[Alter *et al.*, 2000] O. Alter, P. O. Brown, and D. Botstein. Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Sciences*, 97(18):10101–10106, 2000.

[Anttila *et al.*, 2010] Verneri Anttila, Hreinn Stefansson, Mikko Kallela, Unda Todt, Gisela M. Terwindt, M. Stella Calafato, ..., Aarno Palotie, and International Headache Genetics Consortium. Genome-wide association study of migraine implicates a common susceptibility variant on 8q22.1. *Nature Genetics*, 42(10):869–873, 2010.

[Barretina *et al.*, 2012] Jordi Barretina, Giordano Caponigro, Nicolas Stransky, Kavitha Venkatesan, Adam A. Margolin, Sungjoon Kim, ..., and Levi A. Garraway. The cancer

cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, 483(7391):603–607, 2012.

[Berger *et al.*, 2018] Ashton C. Berger, Anil Korkut, Rupa S. Kanchi, Apurva M. Hegde, Walter Lenoir, Wenbin Liu, ..., and Rehan Akbani. A Comprehensive Pan-Cancer Molecular Study of Gynecologic and Breast Cancers. *Cancer Cell*, 33(4):690–705.e9, 2018.

[Bindea *et al.*, 2013] Gabriela Bindea, Bernhard Mlecnik, Marie Tosolini, Amos Kirilovsky, Maximilian Waldner, Anna C. Obenauf, ..., and Jér Ôme Galon. Spatiotemporal dynamics of intratumoral immune cells reveal the immune landscape in human cancer. *Immunity*, 39(4):782–795, 2013.

[Blei *et al.*, 2017] David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

[Bowtell *et al.*, 2015] David D. Bowtell, Steffen Bohm, Ahmed A. Ahmed, Paul-Joseph Aspuria, Robert C. Bast Jr, Valerie Beral, ..., and Frances R. Balkwill. Rethinking ovarian cancer ii: reducing mortality from high-grade serous ovarian cancer. *Nat Rev Cancer*, 15(1):668–679, 2015.

[Bueno *et al.*, 2016] Raphael Bueno, Eric W. Stawiski, Leonard D. Goldstein, Steffen Durinck, Assunta De Rienzo, Zora Modrusan, ..., and Somasekar Seshagiri. Comprehensive genomic analysis of malignant pleural mesothelioma identifies recurrent mutations, gene fusions and splicing alterations. *Nature Genetics*, 48(4):407–416, 2016.

[Campbell *et al.*, 2018] Joshua D. Campbell, Christina Yau, Reanne Bowlby, Yuexin Liu, Kevin Brennan, Huihui Fan, ..., Esther Drill, Ronglai Shen, The Cancer Genome Atlas

Research Network, ..., and Carter Van Waes. Genomic, Pathway Network, and Immunologic Features Distinguishing Squamous Carcinomas. *Cell Reports*, pages 194–212, 2018.

[Cancer Genome Atlas Network, 2017] Cancer Genome Atlas Network. Comprehensive and Integrated Genomic Characterization of Adult Soft Tissue Sarcomas. *Cell*, 171(4):950–965.e28, 2017.

[Cheng *et al.*, 2015] Chao Cheng, George Tseng, Debashis Ghosh, and Xianghong Jasmine Zhou. *From Transcription Factor Binding and Histone Modification to Gene Expression: Integrative Quantitative Models*, page 380402. Cambridge University Press, 2015.

[Chien *et al.*, 2013] Jeremy Chien, Rui Kuang, Charles Landen, and Viji Shridhar. Platinum-Sensitive Recurrence in Ovarian Cancer: The Role of Tumor Microenvironment. *Frontiers in Oncology*, 3(September):1–6, 2013.

[Cookson *et al.*, 2009] William Cookson, Liming Liang, Goncalo Abecasis, Miriam Moffatt, and Mark Lathrop. Mapping complex disease traits with global gene expression. *Nature Reviews Genetics*, 10(3):184–194, 2009.

[Dacic *et al.*, 2008] Sanja Dacic, Hannelore Kothmaier, Stephanie Land, Yongli Shuai, Iris Halbwedl, Patrizia Morbini, ..., and Helmut Popper. Prognostic significance of p16/cdkn2a loss in pleural malignant mesotheliomas. *Virchows Archiv*, 453(6):627–635, 2008.

[Daemen *et al.*, 2009] Anneleen Daemen, Olivier Gevaert, Fabian Ojeda, Annelies Debucquoy, Johan Suykens, Christine Sempoux, ..., and Bart De Moor. A kernel-based integration of genome-wide data for clinical decision support. *Genome Medicine*, 1(4):1, 2009.

[De Reynies *et al.*, 2014] Aurelien De Reynies, Marie Claude Jaurand, Annie Renier, Gabrielle Couchy, Ilir Hysi, Nabila Elarouci, ..., and Didier Jean. Molecular classification of malignant pleural mesothelioma: Identification of a poor prognosis subgroup linked to the epithelial-to-mesenchymal transition. *Clinical Cancer Research*, 20(5):1323–1334, 2014.

[Dressman *et al.*, 2007] Holly K. Dressman, Andrew Berchuck, Gina Chan, Jun Zhai, Andrea Bild, Robyn Sayer, ..., and Johnathan M. Lancaster. An integrated genomic-based approach to individualized treatment of patients with advanced-stage ovarian cancer. *Journal of Clinical Oncology*, 25(5):517–525, 2007. PMID: 17290060.

[Dubois *et al.*, 2010] Patrick C A Dubois, Gosia Trynka, Lude Franke, Karen A Hunt, Jihane Romanos, Alessandra Curtotti, ..., and David A van Heel. Multiple common variants for celiac disease influencing immune gene expression. *Nature Genetics*, 42(4):295–302, 2010.

[Fan and Lv, 2008] Jianqing Fan and Jinchi Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 70(5):849–911, 2008.

[Friedman *et al.*, 2010] Jerome H. Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *journal of Statistical Software*, 33(1):1–22, 2010.

[Fuchs *et al.*, 2013] Mathias Fuchs, Tim Beissbarth, Edgar Wingender, and Klaus Jung. Connecting high-dimensional mrna and mirna expression data for binary medical clas-

sification problems. *Computer methods and programs in biomedicine*, 111(3):592–601, 2013.

[Galluzzi *et al.*, 2012] L. Galluzzi, L. Senovilla, I. Vitale, J. Michels, I. Martins, O. Kepp, M. Castedo, and G. Kroemer. Molecular mechanisms of cisplatin resistance. *Oncogene*, 31(15):1869–1883, 2012.

[Galluzzi *et al.*, 2014] L Galluzzi, I Vitale, J Michels, C Brenner, G Szabadkai, M Castedo, and G Kroemer. Systems biology of cisplatin resistance: past, present and future. *Cell Death and Disease*, 428, 2014.

[Gamazon *et al.*, 2013] Eric Gamazon, R. Stephanie Huang, Eileen Dolan, Nancy Cox, and Hae Kyung Im. Integrative genomics: Quantifying significance of phenotype-genotype relationships from multiple sources of high-throughput data. *Frontiers in Genetics*, 3(202):1–7, 2013.

[Gill *et al.*, 2012] Ritu R. Gill, William G. Richards, Beow Y. Yeap, Shin Matsuoka, Andrea S. Wolf, Victor H. Gerbaudo, ..., and Hiroto Hatabu. Epithelial malignant pleural mesothelioma after extrapleural pneumonectomy: Stratification of survival with CT-derived tumor volume. *American Journal of Roentgenology*, 198(2):359–363, 2012.

[Glueck *et al.*, 2013] Glueck, Alberto Monterro, and Muaiad Kittane. Molecular Profiling for Breast Cancer: A Comprehensive Review. *Biomarkers in Cancer*, page 61, 2013.

[Gonzalez Bosquet *et al.*, 2016] Jesus Gonzalez Bosquet, Andreea M. Newtson, Rebecca K. Chung, Kristina W. Thiel, Timothy Ginader, Michael J. Goodheart, Kimberly K. Leslie, and Brian J. Smith. Prediction of chemo-response in serous ovarian cancer. *Molecular Cancer*, 2016.

[Guo *et al.*, 2015] Xiangqian Guo, Vickie Y. Jo, Anne M. Mills, Shirley X. Zhu, Cheng Han Lee, Inigo Espinosa, ..., and Matt Van De Rijn. Clinically relevant molecular subtypes in leiomyosarcoma. *Clinical Cancer Research*, 21(15):3501–3511, 2015.

[Hastie *et al.*, 1995] B Y Trevor Hastie, Andreas Buja, Robert Tibshirani, and T Bell Laboratories. Penalized discriminant analysis. *The Annals of Statistics*, 23(1):73–102, February 1995.

[Hastie *et al.*, 2009] T Hastie, R Tibshirani, and J Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer, New York, second edition, February 2009.

[Hoadley *et al.*, 2014] Katherine A. Hoadley, Christina Yau, Denise M. Wolf, Andrew D. Cherniack, David Tamborero, Sam Ng, The Cancer Genome Atlas Research Network, ..., and Joshua M. Stuart. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell*, 158(4):929–944, 2014.

[Hoadley *et al.*, 2018] Katherine A. Hoadley, Christina Yau, Toshinori Hinoue, Denise M. Wolf, Alexander J. Lazar, Esther Drill, Ronglai Shen, ..., The Cancer Genome Atlas Network, ..., and Peter W. Laird. Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. *Cell*, 173(2):291–304.e6, 2018.

[Holter *et al.*, 2000] N. S. Holter, M. Mitra, A. Maritan, M. Cieplak, J. R. Banavar, and N. V. Fedoroff. Fundamental patterns underlying gene expression profiles: Simplicity from complexity. *Proceedings of the National Academy of Sciences*, 97(15):8409–8414, 2000.

[Hood and Friend, 2011] Leroy Hood and Stephen H. Friend. Predictive, personalized, preventive, participatory (P4) cancer medicine. *Nature Reviews Clinical Oncology*, 8(3):184–187, 2011.

[Jennings *et al.*, 2013] Elizabeth Jennings, Jeffrey Morris, Raymond Carroll, Ganiraju Manyam, and Veerabhadran Baladandayuthapani. Bayesian methods for expression-based integration of various types of genomics data. *EURASIP Journal on Bioinformatics and Systems Biology*, 2013(1):1–13, 2013.

[Ji *et al.*, 2015] Hongkai Ji, Yingying Wei, George Tseng, Debashis Ghosh, and Xianghong Jasmine Zhou. *Integrative Analysis of Multiple ChIP-X Data Sets Using Correlation Motifs*, page 110132. Cambridge University Press, 2015.

[Kamieniak *et al.*, 2015] Marta M. Kamieniak, Daniel Rico, Roger L. Milne, Ivan Muñoz-Repeto, Kristina Ibáñez, Miguel A. Grillo, ..., and María J. García. Deletion at 6q24.2-26 predicts longer survival of high-grade serous epithelial ovarian cancer patients. *Molecular Oncology*, 9(2):422–436, 2015.

[Le Van *et al.*, 2016] Thanh Le Van, Matthijs Van Leeuwen, Ana Carolina Fierro, Dries De Maeyer, Jimmy Van Den Eynden, Lieven Verbeke, ..., and Siegfried Nijssen. Simultaneous discovery of cancer subtypes and subtype features by molecular data integration. *Bioinformatics*, 32(17):i445–i454, 2016.

[Leong *et al.*, 2015] Su Lyn Leong, Rizka Zainudin, Laurie Kazan-Allen, and Bruce W. Robinson. Asbestos in Asia. *Respirology*, 20(4):548–555, 2015.

[Li and Jung, 2017] Gen Li and Sungkyu Jung. Incorporating covariates into integrated factor analysis of multi-view data. *Biometrics*, 73(4):1433–1442, 2017.

[Li *et al.*, 2015a] Cong Li, Can Yang, Greg Hather, Ray Liu, Hongyu Zhao, George Tseng, ..., and Xianghong Jasmine Zhou. *Drug-Pathway Association Analysis: Integration of High-Dimensional Transcriptional and Drug Sensitivity Profile*, page 425444. Cambridge University Press, 2015.

[Li *et al.*, 2015b] Wenyuan Li, Chao Dai, Xianghong Jasmine Zhou, George Tseng, Debashis Ghosh, and Xianghong Jasmine Zhou. *Integrative Analysis of Many Biological Networks to Study Gene Regulation*, page 6887. Cambridge University Press, 2015.

[Li, 2013] Hongzhe Li. Systems biology approaches to epidemiological studies of complex diseases. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 5(6):677–686, 2013.

[Liu *et al.*, 2015] Jin Liu, Xingjie Shi, Jian Huang, Shuangge Ma, George Tseng, Debashis Ghosh, and Xianghong Jasmine Zhou. *Penalized Integrative Analysis of High-Dimensional Omics Data*, page 174204. Cambridge University Press, 2015.

[Liu *et al.*, 2018] Yang Liu, Nilay S. Sethi, Toshinori Hinoue, Barbara G. Schneider, Andrew D. Cherniack, Francisco Sanchez-Vega, ..., The Cancer Genome Atlas Research Network, and Peter W. Laird. Comparative Molecular Analysis of Gastrointestinal Adenocarcinomas. *Cancer Cell*, 33(4):721–735.e8, 2018.

[Lloyd *et al.*, 2015] Katherine L. Lloyd, Ian A. Cree, and Richard S. Savage. Prediction of resistance to chemotherapy in ovarian cancer: a systematic review. *BMC Cancer*, 15(1):117, 2015.

[Lock *et al.*, 2013] Eric F. Lock, Katherine A. Hoadley, J. S. Marron, and Andrew B. Nobel. Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. *Annals of Applied Statistics*, 7(1):523–542, 2013.

[López-Ríos *et al.*, 2006] Fernando López-Ríos, Shannon Chuai, Raja Flores, Shigeki Shimizu, Takatoshi Ohno, Kazuhiko Wakahara, ..., and Marc Ladanyi. Global gene expression profiling of pleural mesotheliomas: Overexpression of aurora kinases and P16/CDKN2A deletion as prognostic factors and critical evaluation of microarray-based prognostic prediction. *Cancer Research*, 66(6):2970–2979, 2006.

[Mahdian-shakib *et al.*, 2016] Ahmad Mahdian-shakib, Ruhollah Dorostkar, Mahdi Tat, Mohammad Sadegh Hashemzadeh, and Navid Saidi. Differential role of micrornas in prognosis, diagnosis, and therapy of ovarian cancer. *Biomedicine & Pharmacotherapy*, 84:592–600, 2016.

[Mak *et al.*, 2016] Milena P. Mak, Pan Tong, Lixia Diao, Robert J. Cardnell, Don L. Gibbons, William N. William, ..., and Lauren Averett Byers. A Patient-Derived, Pan-Cancer EMT Signature Identifies Global Molecular Alterations and Immune Target Enrichment Following Epithelial-to-Mesenchymal Transition. 22(3):609–620, 2016.

[Mankoo *et al.*, 2011] PK Mankoo, R Shen, N Schultz, DA Levine, and C Sander. Time to recurrence and survival in serous ovarian tumors predicted from integrated genomic profiles. *PLoS ONE*, 6(11):e24709, 2011.

[Matsuo *et al.*, 2010] Koji Matsuo, Yvonne G Lin, Lynda D Roman, and Anil K Sood. Overcoming platinum resistance in ovarian carcinoma. *Expert Opinion on Investigational Drugs*, 19(11):1339–1354, 2010.

[Mo *et al.*, 2013] Qianxing Mo, Sijian Wang, Venkatraman E. Seshan, Adam B. Olshen, Nikolaus Schultz, Chris Sander, ..., and Ronglai Shen. Pattern discovery and cancer gene identification in integrated cancer genomic data. *Proceedings of the National Academy of Sciences*, 110(11):4245–4250, 2013.

[Montgomery and Dermitzakis, 2011] Stephen B. Montgomery and Emmanouil T. Dermitzakis. From expression qtls to personalized transcriptomics. *Nature Reviews Genetics*, 12(4):277–282, 2011.

[Monti *et al.*, 2005] Stefano Monti, Kerry J Savage, Jeffery L Kutok, Friedrich Feuerhake, Paul Kurtin, Martin Mihm, ..., and Margaret a Shipp. Molecular profiling of diffuse large B-cell lymphoma identifies robust subtypes including one characterized by host inflammatory response. *Response*, 105(5):1851–1861, 2005.

[Mylavarapu *et al.*, 2018] Sanghamitra Mylavarapu, Asmita Das, and Monideepa Roy. Role of BRCA Mutations in the Modulation of Response to Platinum Therapy. *Frontiers in Oncology*, 8(February):1–11, 2018.

[National Cancer Institute, 2018] National Cancer Institute. Nci dictionary of cancer terms. `https://www.cancer.gov/publications/dictionaries/cancer-terms`, 2018.

[Newton *et al.*, 2017] Yulia Newton, Adam M. Novak, Teresa Swatloski, Duncan C. McColl, Sahil Chopra, Kiley Graim, Alana S. Weinstein, Robert Baertsch, Sofie R. Salama, Kyle Ellrott, Manu Chopra, Theodore C. Goldstein, David Haussler, Olena Morozova, and Joshua M. Stuart. TumorMap: Exploring the molecular similarities of cancer samples in an interactive portal. *Cancer Research*, 77(21):e111–e114, 2017.

[Nicolae *et al.*, 2010] DL Nicolae, E Gamazon, W Zhang, S Duan, ME Dolan, and N Cox. Trait-associated snps are more likely to be eqtls: Annotation to enhance discovery from gwas. *PLoS Genetics*, 6(4):e1000888, 2010.

[Norquist *et al.*, 2011] Barbara Norquist, Kaitlyn A. Wurz, Christopher C. Pennil, Rochelle Garcia, Jenny Gross, Wataru Sakai, Beth Y. Karlan, Toshiyasu Taniguchi, and Elizabeth M. Swisher. Secondary somatic mutations restoring BRCA1/2 predict chemotherapy resistance in hereditary ovarian carcinomas. *Journal of Clinical Oncology*, 29(22):3008–3015, 2011.

[Pavlidis *et al.*, 2002] Paul Pavlidis, Jason Weston, Jinsong Cai, and William Stafford Noble. Learning gene functional classfications from multiple data types. *Journal of Computational biology*, 9(2):401–411, 2002.

[Pidsley *et al.*, 2013] R Pidsley, C C Y Wong, M Volta, K Lunnon, J Mill, and L C Schalkwyk. A data-driven approach to preprocessing Illumina 450 K methylation array data. *BMC Genomics*, 14:293, 2013.

[Ricketts *et al.*, 2018] Christopher J. Ricketts, Aguirre A. De Cubas, Huihui Fan, Christof C. Smith, Martin Lang, Ed Reznik, ..., The Cancer Genome Atlas Research Network, and W. Marston Linehan. The Cancer Genome Atlas Comprehensive Molecular Characterization of Renal Cell Carcinoma. *Cell Reports*, 23(1):313–326.e5, 2018.

[Sekido, 2013] Yoshitaka Sekido. Molecular pathogenesis of malignant mesothelioma. *Carcinogenesis*, 34(7):1413–1419, 2013.

[Shen *et al.*, 2009] 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, 2009.

[Shen *et al.*, 2013] Ronglai Shen, Sijian Wang, and Qianxing Mo. Sparse integrative clustering of multiple omics data sets. *Annals of Applied Statistics*, 7(1):269–294, 2013.

[Storey and Tibshirani, 2003] John D. Storey and Robert Tibshirani. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, 100(16):9440–9445, 2003.

[Svejstrup, 2013] Jesper Q. Svejstrup. Synovial sarcoma mechanisms: A series of unfortunate events. *Cell*, 153(1):11–12, 2013.

[Swisher *et al.*, 2008] Elizabeth M. Swisher, Wataru Sakai, Beth Y. Karlan, Kaitlyn Wurz, Nicole Urban, and Toshiyasu Taniguchi. Secondary BRCA1 mutations in BRCA1-mutated ovarian carcinomas with platinum resistance. *Cancer Research*, 68(8):2581–2586, 2008.

[Tan *et al.*, 2010] Ke Tan, Kazunori Kajino, Shuji Momose, Akiko Masaoka, Keiichi Sasahara, Kazu Shiomi, ..., and Hiroaki Fujii. Mesothelin (MSLN) promoter is hypomethylated in malignant mesothelioma, but its expression is not associated with methylation status of the promoter. *Human Pathology*, 41(9):1330–1338, 2010.

[Teo *et al.*, 2015] Guoshou Teo, Christine Vogel, Debashis Ghosh, Sinae Kim, Hyungwon Choi, George Tseng, Debashis Ghosh, and Xianghong Jasmine Zhou. *A Mass-Action-Based Model for Gene Expression Regulation in Dynamic Systems*, page 362379. Cambridge University Press, 2015.

[The Cancer Genome Atlas Research Network, 2011] The Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. *Nature*, 474:609–615, Jun 2011.

[Tseng *et al.*, 2015] George Tseng, Debashis Ghosh, and Xianghong Jasmine Zhou. *Integrating Omics Data*. Cambridge University Press, 2015.

[Tucker *et al.*, 2014] Susan L. Tucker, Kshipra Gharpure, Shelley M. Herbrich, Anna K. Unruh, Alpa M. Nick, Erin K. Crane, Robert L. Coleman, Jamie Guenthoer, Heather J. Dalton, Sherry Y. Wu, Rajesha Rupaimoole, Gabriel Lopez-Berestein, Bulent Ozpolat, Cristina Ivan, Wei Hu, Keith A. Baggerly, ..., and Anil K. Sood. Molecular biomarkers of residual disease after surgical debulking of high-grade serous ovarian cancer. *Clinical Cancer Research*, 20(12):3280–3288, 2014.

[Van De Wiel and Van Wieringen, 2007] Mark A. Van De Wiel and Wessel N. Van Wieringen. CGHregions: Dimension reduction for array CGH data with minimal information loss. *Cancer Informatics*, 3(0):55–63, 2007.

[VanderKraats *et al.*, 2013] Nathan D. VanderKraats, Jeffrey F. Hiken, Keith F. Decker, and John R. Edwards. Discovering high-resolution patterns of differential dna methylation that correlate with gene expression changes. *Nucleic Acids Research*, 41(14):6816–6827, 2013.

[Vaske *et al.*, 2010] Charles J. Vaske, Stephen C. Benz, J. Zachary Sanborn, Dent Earl, Christopher Szeto, Jingchun Zhu, ..., and Joshua M. Stuart. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics*, 26(12):237–245, 2010.

[Verhaak *et al.*, 2010] Roel G.W. Verhaak, Katherine A. Hoadley, Elizabeth Purdom, Victoria Wang, Yuan Qi, Matthew D. Wilkerson, ..., 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.

[Williams *et al.*, 2015] Marissa Williams, Michaela B. Kirschner, Yuen Yee Cheng, Jacky Hanh, Jocelyn Weiss, Nancy Mugridge, ..., and Glen Reid. miR-193a-3p is a potential tumor suppressor in malignant pleural mesothelioma. *Oncotarget*, 6(27):23480–23495, 2015.

[Witten *et al.*, 2009] Daniela M. Witten, Robert Tibshirani, and Trevor Hastie. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10(3):515–534, 2009.

[Yap *et al.*, 2017] Timothy A. Yap, Joachim G. Aerts, Sanjay Popat, and Dean A. Fennell. Novel insights into mesothelioma biology and implications for therapy. *Nature Reviews Cancer*, 17(8):475–488, 2017.

[Yu *et al.*, 2011] Wei Yu, Chengmeng Jin, Xiaoyan Lou, Xu Han, Lisha Li, Yinghua He, ..., and Biaoyang Lin. Global analysis of DNA methylation by methyl-capture sequencing reveals epigenetic control of cisplatin resistance in Ovarian cancer cell. *PLoS ONE*, 6(12), 2011.

[Zeller *et al.*, 2012] C. Zeller, W. Dai, N. L. Steele, A. Siddiq, A. J. Walley, C. S.M. Wilhelm-Benartzi, ..., and R. Brown. Candidate DNA methylation drivers of acquired cisplatin resistance in ovarian cancer identified by methylome and expression profiling. *Oncogene*, 2012.

[Zou and Hastie, 2005] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.