UNIVERSITY OF CALIFORNIA
SANTA CRUZ

LEARNING FROM NEW PERSPECTIVES: USING SPARSE DATA AND MULTIPLE VIEWS TO PREDICT CANCER PROGRESSION AND TREATMENT

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

DOCTOR OF PHILOSOPHY

in

BIOMOLECULAR ENGINEERING & BIOINFORMATICS

by

Kiley Graim

December 2016

The Dissertation of Kiley Graim
is approved:

Professor Joshua M. Stuart, Chair

Professor David Haussler

Sofie Salama, Ph.D.

Dean Tyrus Miller
Vice Provost and Dean of Graduate Studies

Table of Contents

List of Figures

List of Tables

Abstract

Learning from new perspectives: Using sparse data and multiple views to predict cancer progression and treatment

by

Kiley Graim

Advancements in sequencing technology have led to an influx of cancer genomics data, transforming cancer research into a field limited by data interpretation rather than acquisition. Machine learning methods that can make use of this wealth of data are desperately needed. Similarly, patient stratification is a critical task in cancer diagnosis and treatment. While stratification approaches using various biomarkers for patient-to-patient comparisons have been successful in elucidating previously unseen subtypes, the potential of many other sparse but rich genotype and phenotype data (*e.g.* tumor images) remains untapped.

To this end, I present two methods. The first uses social network analysis techniques to extract subtypes from sparse data. The second is a semi-supervised multiview learning framework that integrates both prior knowledge and a variety of genomic data to predict outcomes in cancer. Crucially, this method accommodates samples for which we have different data types, paving the way for integration of data from past studies.

I apply these methods to several cancer datasets. Of note, I show that TCGA–defined molecular subtypes of glioblastoma are independent of both tumor location and volume, and that both the imaging and genomic data provide important perspectives of

the disease. Analysis of a large drug sensitivity database identifies an epigenetic effect from chromatin modifiers that lends sensitivity to Panobinostat. Multiview learning, the second method I developed, also outperforms other methods in predicting sensitivity in all of the study drugs. In this dissertation I begin with unsupervised single–platform analysis, then combine multiple platforms, and finally analyze many data platforms using semi–supervised analysis.

To my amazing husband,

James Cahill,

for keeping me happy and healthy.

To my parents,

Jane Schmidt and Tom Graim,

for teaching me to love knowledge,

and for their unconditional love and support.

Acknowledgments

Thank you to all of my lab members for being so wonderful to work with, and to Josh for setting up a fantastic lab for all of us. This has been the most incredibly positive PhD experience. To all of my committee members, thank you for all the things you have taught me over the years and for having been so supportive of my career.

Mark Diekhans and Karen Miga, thank you for being such amazingly supportive mentors. I would like to thank the MVL team, Katie and Verena, for their helpful discussions and collaboration. Thank you Christina Yau and Chris Benz for teaching me PARADIGM analysis. Also many thanks to the CBSE staff and the UCSC browser. To the TCGA program office and all of the people involved, it has been a wonderful experience to be a part of.

And lastly, thank you so much to Audrey Kim. I looked forward to your group every week.

Chapter 1

Introduction

Cancer is a disease of information in DNA and its 'digital age' has dawned; The plummeting cost of -omics technologies is transforming cancer research from a field limited by data acquisition to one limited by data interpretation. We desperately need biomarkers and machine-learning methods to predict outcomes, especially those that make use of a battery of multiple different measurement platforms to provide an integrated view. Unfortunately the large number of variables compared to the few samples available leads to many biologically irrelevant solutions [350].

Furthermore, I have found that we still lack flexible methods that can integrate data from multiple studies; Most require complete data for each sample. In the case of RNA-Seq and microarray data, investigators often subset down to the genes that have measured expression in all samples. Tools such as ComBat [167] can combine data from multiple expression platforms and remove batch effects, but can also inadvertently remove key biological differences. Similarly, samples that are missing one or more data

types used in an analysis often are excluded from that analysis, further shrinking the sample size. Missing values can be imputed, which runs the risk of poor imputation and cannot be done when the majority of samples are missing values, or be ignored, which introduces bias based on the pattern of missing data.

In the past years many new data modalities have become available, for example imaging data. Researchers can now analyze imaging, genomic, and phenotypic data together. As more imaging data becomes available it behooves researchers to incorporate it into analysis, as another perspective of the disease. The method presented in Chapter 3 provides a technique for analyzing such imaging data, as well as other sparse data platforms.

In Chapter 4 I discuss my integrative clustering analysis on several collaborative projects. The Cancer Genome Atlas (TCGA, *cancergenome.nih.gov*) began in 2005 and has been ongoing for more than 10 years now. Both the Haussler and Stuart labs have worked as part of the TCGA core to uncover novel cancer biology in each of the cancer cohorts. I participated in several TCGA working groups and present several findings from that work.

The nascent sub-field of pathway-informed learning currently is in need of methodologies that use pathway information for predicting outcomes using a principled formulation, allowing models to be tuned to training data in a unified and consistent manner. In Chapter 5 I present a multiview learning (MVL) framework that optimizes learning based on not only data with known outcome but those that are missing outcome information, and without ignoring data from any platform. My MVL framework

incorporates all data from all samples, and will be able to do so without losing the unique information within each data view. Through MVL learning, each data view can be treated independently or in combination– allowing use of multiple feature transformations and selection for each without losing vital gene and pathway information.

In Chapter 3, I motivate integration of MRI with clinical and genomic data. I then move into my contributions to several projects within The Cancer Genome Atlas (TCGA) and in ongoing clinical trials, in Chapter 4. In Chapter 5 I combine multiple biologically–driven views in a multiple view learning framework, and use these views to predict drug sensitivity in the Cancer Cell Line Encyclopedia (CCLE) [17]. This method is also applied to a combination of two ongoing projects studying treatment-resistant prostate cancer, in which MVL is used to predict a rare histologic subtype within the combined datasets.

Chapter 2

Previous Work

2.1 A Brief Survey of the Literature

2.1.1 The Cancer Genome Atlas

The Cancer Genome Atlas (TCGA, *cancergenome.nih.gov*) is a collaboration between the National Cancer Institute (NCI) and National Human Genome Research Institute (NHGRI). TCGA has made cancer informatics accessible to a number of fields by making cancer data readily available to researchers not affiliated with cancer biology labs. It standardized the quality, format, and types of genomic data available, while releasing patient cohorts of previously unheard sizes. Today over 11,000 patients and 35 cancers have been characterized by TCGA. Its goal has been to improve diagnosis and treatment of cancer.

2.1.2 Joint Analysis of Magnetic Resonance Images and Genomic Data

Brain tumors are initially diagnosed through Magnetic Resonance Imaging (MRI). Annual competitions such as the Multimodal Brain Tumor Segmentation Challenge (BraTS, *www.braintumorsegmentation.org*) [237] promote development of methods to automatically segment tumor images. Similarly, TCGA efforts have been manually curating images into known survival-linked features. *Visually AcceSAble Rembrandt Images* (VASARI) [131] features were selected by a cohort of domain experts (neuro-radiologists) and from a review of current imaging literature. The largest current database, TCGA VASARI, contains 130 curated tumor images [131].

Cancer imaging research is in its early stages. There is limited available data and small sample sizes, leading to discordance in discoveries. For example, brain tumor imaging studies have alternately found no mutations associated with imaging features [132], or for example found TP53 mutations to be associated with either the frontal [400] or temporal lobe [365]. While several imaging features are associated with survival [66, 88, 132], combination with genomic data is difficult. All of these studies quote a need for more patients, since the high dimensionality of imaging data makes it difficult to find meaningful associations between imaging and genomic features. It is my hope that the data from these studies will be combined in a joint analysis in the future, to help minimize spurious results due to small sample size.

2.1.3 Pairwise & Subspace Clustering

"The aim of clustering is to find structure in data and is therefore exploratory in nature [159]." While clustering finds cliques of similar samples within a large group, subspace clustering specifically aims to identify cliques of samples along a certain subspace within the data features, rather than the entire data [192]. The key issue is taking into account the definition of similarity so that only certain subspaces are considered. By looking at correlation within local structures one can identify long–range connections, critical for finding network paths [192]. For example the 'small world phenomenon' claims that we are all connected via short chains of links [186].

Similarly, pairwise clustering is based on sample–sample similarities. Pairwise clustering techniques consider the n nearest neighbors with edge weights based on sample–sample affinities. Pairwise clustering is widely applicable due to its ability to find clusters with arbitrary shapes [299]. There are many instances of pairwise clustering in bioinformatics. For example clustering pre-miRNA sequences based on pairwise, sequence, and secondary structure alignment [170]; FOLDALIGNM performs well on samples with low sequence similarity [334]. Another example, HyperPrior, correlates genomic features in a graph–based learning framework, to cluster gene experession and arrayCGH data using a biological prior [332]. WGCNA is similar in that it uses weighted correlation networks to cluster gene expression data [199]. SNN-Cliq clusters single cell transcriptome data [377]. Pairwise clustering is also widely used outside of bioinformatics, for example a recent paper proposes HD–MSL, which uses high–order distance

learning from the hypergraph (rather than pairwise distances) to classify images [387].

Correlation clustering is a graph–based clustering approach. One can use a linear discriminant function so that the support vector machine approach can be used as in Finley *et. al.* [99] and Taskar *et. al.* [328]. This can then be extended to 'higher-order' correlations, where features and groups of features are considered in tandem [182]. Such approaches are popular because of their speed and success in clustering problems with latent similarities.

2.1.4 Integrated Genomics Analysis

Integrating together different genomics data requires expertise from many fields [193]. As data becomes more available it is easier to create multidisciplinary teams and more integrated analysis tools are available. Some tools create user interfaces that negate the need for programming, for example the cBio Cancer Genomics Portal (cbioportal.org) [48], Cytoscape [294, 305], and StratomeX [211]; Other tools help with subtype identification. Jiang *et. al.* [164] integrate protein-protein interaction network with gene expression and histone marks to predict gene essentiality. Another approach uses linked 2D sample similarities to combine different genomic data [30]. Hoadley *et. al.* [37, 148] use consensus clustering of 6 genomic platforms to identify cancer subtypes. iCluster [242, 298] uses a joint latent model to cluster data based on many genomics platforms, whereas others use multiple kernel learning so that each genomic data platform comprises a kernel for combined clustering [121, 312]. PARADIGM [347] integrates multiple platforms and outputs 'inferred pathway levels' (IPLs) which can then be clus-

tered. All of these approaches attempt to identify global signals of dysregulation in cancer using a combination of a variety of genomic data. Integrative analysis enables us to see multiple perspectives of patient disease. I will discuss the importance of this in more detail in Chapter 5.

2.1.5 Bioinformatics Competitions

Bioinformatics challenges are becoming more popular. They both increase awareness of bioinformatics questions and reduce evaluation bias [31]. Challenges provide benchmarking data and methods. Similar to the UCI machine learning repository (*archive.ics.uci.edu/ml*), they provide easy-to-use data and gold standards for method evaluation. Many challenges run every year or few years. For example, there is the (1) Critical Assessment of protein Structure Prediction (CASP, *predictioncenter.org*), (2) Dialogue for Reverse Engineering Assessments and Methods (DREAM, *dreamchallenges.org*), (3) Critical Assessment of protein Function Annotation algorithms (CAFA, *biofunctionprediction.org/cafa*), and (4) Assemblathon (*assemblathon.org*). Methods that performed the best in the DREAM7 Drug Sensitivity Challenge used a combination of clinical and genomic features with a biological prior [28]. Many other challenges have similar results; methods that used biological priors and integrate multiple data have the best performance [147].

2.1.6 Multiview Learning

Fully labeled data is not always available; Partially labeled data, despite being tricky to analyze, is [290]. Multiview methods have been developed for just this situation.

Multiview learning initially was introduced through co-training [29], and later through the use of multiple kernels [15]. The International Conference on Machine Learning (ICML) had a workshop in 2005 entitled 'Learning with Multiple Views' [282] which covered a multitude of both supervised and unsupervised MVL methods. Neural Information Processing Systems (NIPS) also had a session on multiview learning in 2008 [49]. It has since intermittently been a topic at various machine learning conferences and is starting be to introduced in bioinformatics problems [69, 118, 406, 312].

The canonical MVL scenario is website classification, where one has a small subset of manually curated web pages that have been labeled 'interesting' or 'not interesting' in relation to some person. A plethora of information is available on other websites, but curation is both expensive and time-consuming. Website data innately has two 'views:' the text in the document and the hyperlinks pointing to the site from other locations on the Internet. For example, a link called 'my advisor' is an indication that we will be directed to a faculty website and are on a student website, whereas 'my research' in the text is an indication that it is an academic site. By using the independently sufficient views (meaning that each view is capable of accurately predicting whether or not a website belongs to a faculty member), one can co-train using the

plethora of unlabeled data to cheaply and accurately find labels.

There are 3 general types of multiview learning [321, 376]:

1. co-training [29, 73, 76, 128, 361, 362, 391, 409]

2. multiple kernel learning [14, 115, 173, 295]

3. subspace learning [94, 171, 197, 313, 354, 360]

Cited are some examples of each types of multiview learning. The 3 styles can be thought of as early (subspace learning), intermediate (multiple kernel learning), or late (co-training) integration approaches, each with different strengths.

Co-training relies on 3 principles: (1) sufficiency of each view (aka high accuracy) (2) compatibility and (3) conditional independence [321, 376]. It can be successful with two views or fewer by using different algorithms for each view [361, 363], although the benefits may be directly related to sample size [21]. Recently, method have incorporated 'weak' views [246, 274]. Co-training is vulnerable to mislabeled samples [321], which can be mitigated by using canonical correlation analysis to inspect new labels [322]. It is ideal in scenarios with missing data and views with divergent information content [376].

Multiple kernel methods construct kernels from subspaces within the full data so as to limit the feature spaces. By using multiple kernels instead of a more stringent feature selection method on the full data set, users can enforce domain-specific knowledge [124]. Furthermore, by using multiple kernels one can find latent subspaces within the different kernels that would not surface in the full feature set. Multiple KLM are

most suited to scenarios where there is little missing data [376].

Subspace learning projects views into correlated subspaces, usually using canonical correlation analysis (CCA), and benefits from highly correlated views [376]. For example predicting drug sensitivity [69], and to predict Alzheimer's disease using imaging and genomic data [405]. It is best suited for use in problems with highly correlated views [321]. Multiview learning relies on views being accurate and complementary; as with ensemble learning, it benefits from have many independent representations of the samples.

2.2 Looking Forward

This dissertation presents a method for single platform cluster analysis that applies social networks techniques to genomic data. Chapter 3 discusses how this new view of a single dataset helps with interpretation. The following chapter, Chapter 4, introduces integrative clustering analysis from several collaborative projects. Lastly, in Chapter 5, I present a semi-supervised method that integrates predictions from multiple data platforms and prior knowledge databases. Thus I progress from a unsupervised single platform approach, to unsupervised multiple platforms, and lastly to a semi-supervised multiple platform approach.

Chapter 3

HOCUS: Higher-Order Correlations to Uncover Subtypes

3.1 Introduction

The establishment of expression-based subtypes, have shown to be of tremendous use in predicting patient outcomes (e.g. PAM50 and MammaPrint subtypes for breast cancer prognosis). Most recently, transcriptome-wide RNA sequencing data or other high-throughput measurements have been used to segregate patient samples, which in turn has led to changes in treatment of many cancers. A personalized approach to medicine

Both the sparsity of mutations and mutual exclusivity common in mutation profiles (within the same molecular pathways), complicates the task of subtyping because similarities computed from the original mutation events lack specificity and ro-

bustness, due to the small number of overlapping events between any two samples. One encounters a similar situation when subtyping patients based on imaging data such as from magnetic resonance (MR), as has been recently proposed [218]. In this case, the sparseness of anatomical/spatial MR image data is due to the fact that tumors occupy only a fraction of the affected tissue (e.g. local area in the brain). Manual steps have been used to aid the clustering and therefore may be viewed as subjective.

In this work, we test the use of "higher-order" similarity measures between the samples to identify biologically relevant subsets. Intuitively, we derive a metric to compare two samples based on how similar their sample "neighborhoods" are to one another. To illustrate (Fig. 3.1), one can picture a first-order network created by linking any two samples that have high first-order similarities. Then, a second order network could be formed by linking samples with highly overlapping neighborhoods in the first-order network. Repetition of this procedure generates higher-order networks from a lower-order version, that could reveal community structure.

We use such a similarity transformation, here referred to as *Higher-Order Correlations to Uncover Subtypes* (HOCUS), and show that HOCUS enhances the detection of biologically-relevant subtypes for several Cancer Genome Atlas (TCGA) cohorts. Examples in which highly relevant subtypes are identified from cancer mutational and copy number data are given, demonstrating the method's usefulness applied to both categorical and continuous data modalities. In addition, we apply HOCUS to magnetic resonance imaging data (MRI/MR images) and establish links between MR images and patient survival.

We looked for inspiration from fields with similar data, such as social network analysis. Detecting community structure is an important problem in the study of many different types of networks including social (e.g. connected friends), online (linked web pages), and molecular (regulatory gene signaling). In these applications, communities represent sets of densely connected nodes within a larger set of nodes in a network. Cliques of friends with shared interests or a gene module representing the function of genes in a biological pathway are examples of such communities.

Community detection techniques have so far been under-utilized for the purpose of subtyping patients based on shared genomic- and image-based events. Yet the application is straightforward – the data can be converted readily into a network of patient samples using sample-sample similarities. We hypothesize that using these methods will boost performance of community detection especially when the data are sparse. Mutations and MR images are sparse, since few mutation events are shared between patients and the relative ratio of tumor to normal tissue in the brain means that most regions are tumor-free. We demonstrate several cases in which the HOCUS community detection approach identifies communities missed by standard clustering.

3.2 Results

3.2.1 Overview of HOCUS Clustering

HOCUS uses a technique from network analysis in which samples are compared based on their neighborhood similarity [1, 111, 186] and can be pictured as the construc-

tion of progressively higher-order networks (Fig. 3.1). The original data are provided as features for each of the patient samples, and represent somatic mutations, copy number events, or 3D images of tumor specimens. Next, sample-sample similarities are calculated using an appropriately chosen similarity metric (Fig. A.1, Supp Section A.1.2.3) that can be viewed as a sample-by-sample network. Higher-order similarities are derived from lower-order similarities by treating the lower-order similarities computed at step $k-1$ as the features used to compute new similarities at step k, similar to Yu *et al* [387] and Yu *et al* [386], and to exponentiating a network's adjacency matrix to reveal connected components linked by reachable paths. The samples can be clustered using either the original features (i.e. use first-order similarities) or those derived from higher-order similarities, identifying groups of patients having a higher proportion of transitive relations.

We applied the approach to the problem of detecting cancer subtypes using two very different data modalities - 3D tumor imaging data and somatic mutations. Clustering patient samples by their shared genomic events or related imaging features may reveal common disease etiology important for outcome assessment. Yet mutation and imaging data are sparse – sample pairs have few overlapping events. It is therefore problematic to use these data as features directly for clustering since similarities calculated from sparse spaces suffer in sensitivity and specificity [100]. Similarities based on the network neighborhood can be more sensitive because the likelihood that two samples have an indirect coincidence through other samples is higher than having directly coinciding events. We show that the use of HOCUS for both mutations and imag-

ing data also adds specificity as it produces inferred subtypes that are biologically- and treatment-relevant that were undetected by the equivalent approaches using lower-order metrics.

3.2.2 Community Detection Reveals Cancer Subtypes Using Somatic Mutation Data

The particular ways in which a tumor genome is damaged and altered leaves a signature that reflects the type of cell and mutagen involved. Driving events involving specific genes are associated with certain cancer types and not others. For instance, BCR-ABL fusions are characteristic of chronic myeloid leukemias. The question is whether the pattern of mutations within these cells of origin can further subdivide the patient samples into meaningful categories that inform treatment.

We applied HOCUS to mutation data for 3 TCGA cancers: ovarian cysadenocarcinomas (OV), glioblastoma multiforme (GBM), and bladder urothelial carcinoma (BLCA). We computed Hamming similarity (Eq. 3.1, sum of matched voxels) for all sample pairs, resulting in an adjacency matrix of $m \times m$ samples. For higher-order clustering, we raised the precomputed similarity matrix S to the $d-1$ power, where d is the order of clustering. We then supplied this similarity matrix as the feature matrix for input to ConsensusClusterPlus (see Methods). Figure 3.1 shows a conceptual example of this principle– as the order of clustering increases, cliques in the network emerge and form clusters.

We retained for clustering all metrics that provided a non-redundant set of

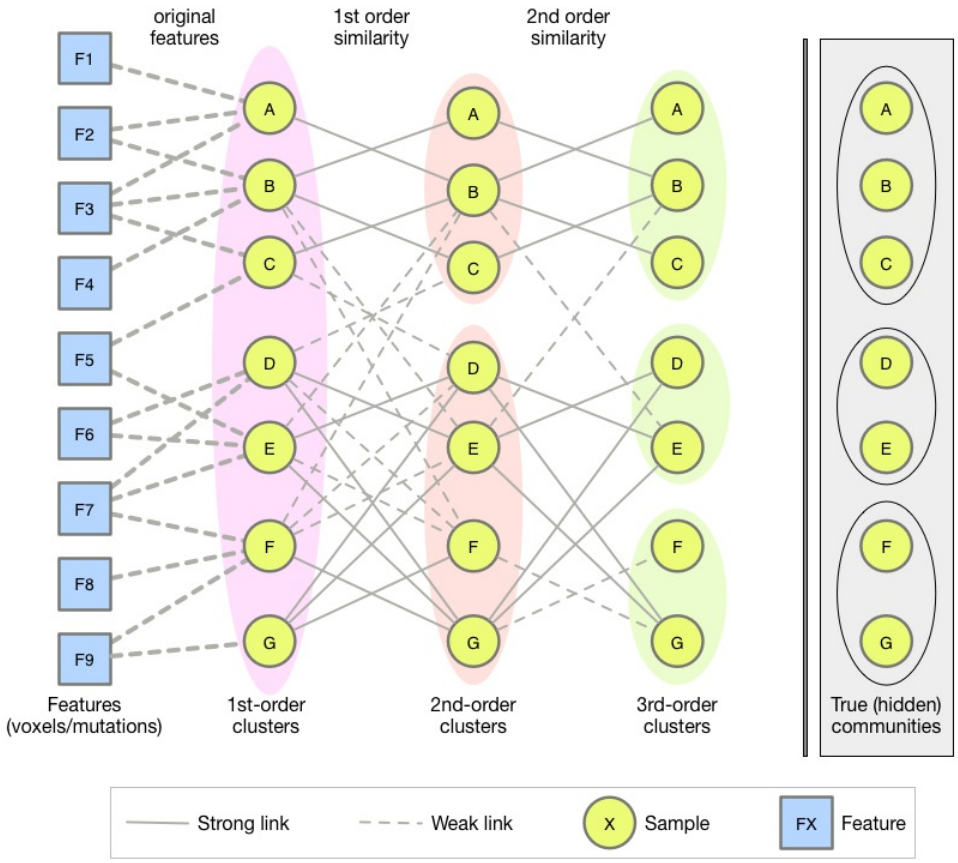

Figure 3.1: Social network approach to clustering patient samples. First we transform/encode the mutation/voxel data, then compute all patient–patient similarities. At each order of similarities, clustering is based on similarities in that order, resulting in different clustering solutions. Shown here from left to right: features, 1st-order, 2nd-order, 3rd-order, 'true' communities.

relations between samples not captured by lower-order metrics. To do this, we identified all k^{th}-order metrics and lower such that the $(k+1)^{th}$ metric produced highly similar relative similarities to the k^{th} metric as measured by a kernel alignment test [89] (see Fig. A.2). We sought to determine if higher-order feature-based similarity measures (i.e. those based on mutations, images etc) had an enrichment for connecting patients with similar survival outcomes compared to using first-order feature-based measures.

For each tumor type, we clustered the patient samples based on either Pearson correlation, the first-order Hamming similarities, or non-redundant higher-order similarities. We used K-means ConsensusClustering [372], varying the choice of the number of clusters (K), and calculated the degree to which the solutions separated patients with different outcomes as a measure of biological relevance. A Kaplan-Meier test was performed on each clustering solution and the significance (-log P-value) was recorded (Fig. 3.2).

We applied HOCUS to the TCGA GBM dataset containing 283 patients for which 14,910 mutations were found across 7,874 distinct genes, and found 3 distinct clusters. Survival differentiation has proven difficult to achieve in previous analyses of GBM datasets [32, 351], however the HOCUS results show some difference in survival between clusters. In the best surviving, cluster 1, the patients have low EGFR and TTN mutation occurrence compared to patients in other clusters but few mutations in either; TTN mutations are mostly in cluster 3 and EGFR distributed between clusters 2 and 3. All 14 of the IDH1 mutants are in cluster 1, as are nearly all (11 of 16) of the ATRX mutants. Cluster 1 corresponds well with mRNA cluster 3 (LGr3) from the

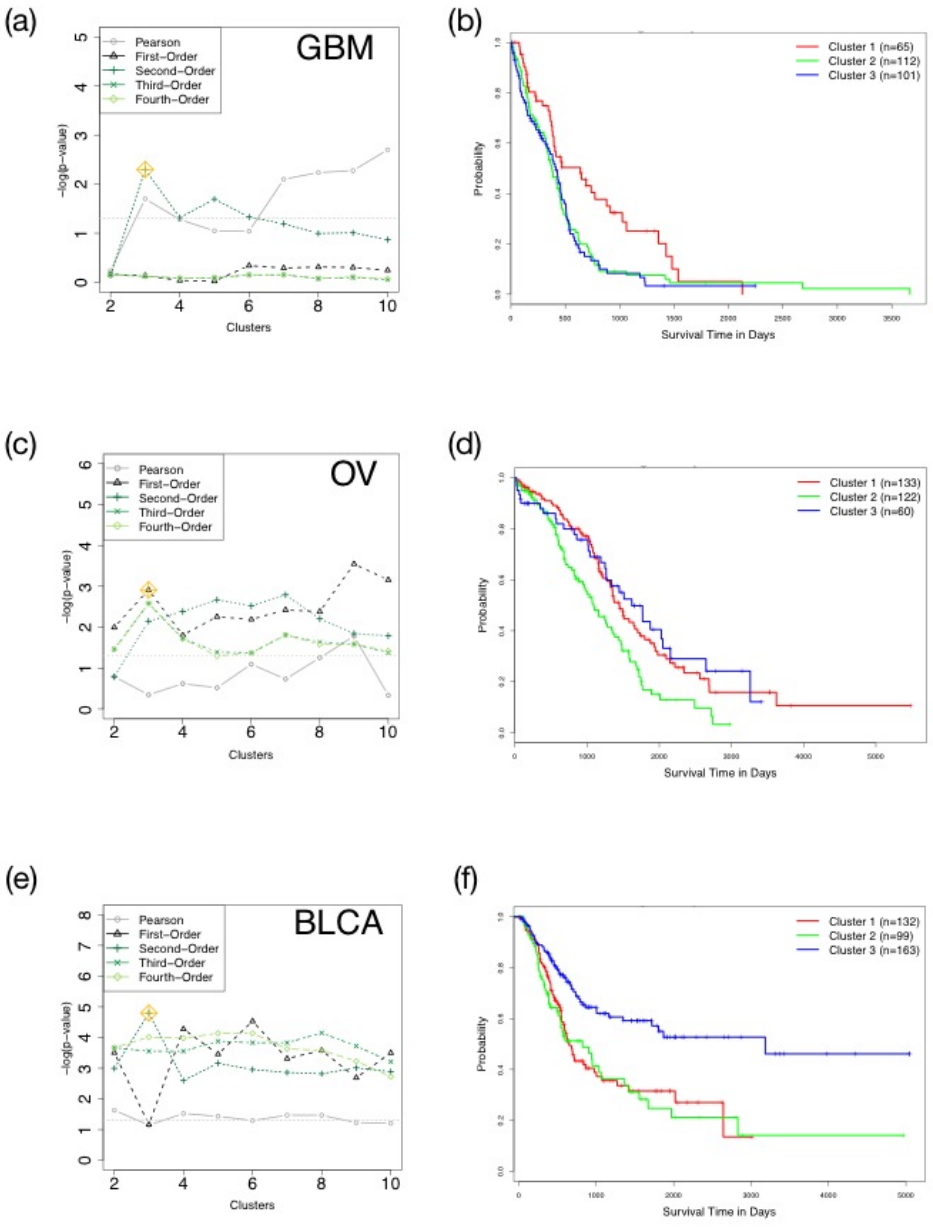

Figure 3.2: HOCUS in first- through fourth-orders, and pearson clustering of (a) GBM (c) OV and (e) BLCA survival p-values vs number of clusters. (b) GBM, (d) OV, and (f) BLCA Kaplan-Meier plots for selected HOCUS clustering solutions. Clusters with fewer than 5 samples are excluded from the KM analyses.

recent TCGA paper [42]. Thus, the HOCUS clustering on mutations seems to have been able to tease out a low grade diffuse subtype, defined by IDH1 mutation status often seen in younger individuals, characterized by the absence of a 1p/19q codeletion and a lack of TERT expression and an overall better prognosis. Furthermore, every cluster 1 sample has a TP53 mutation (Fig. A.3), whereas there are none in cluster 2 and only 14 (of 105 samples) in cluster 3.

We next applied HOCUS to the TCGA OV dataset containing 316 patients for which 14,810 mutations in 8,258 distinct genes were reported by the TCGA analysis working group. For the OV dataset, the first order solution found the greatest separation in survival between clustered groups. Higher-order metrics gave different solutions but were comparable with the first order solution in separating out patient groups with differences in outcome. One of the main divisions of the samples shows a significant difference in overall mutation rate. In addition to TP53 mutations, several genes that are characteristic of passenger mutations are also predominant in the highly mutated cluster including TTN, MUC16, and RYR2. Other mutations are significantly associated with these clusters, highlighted in Figure A.4. HOCUS OV clusters correlate with platinum resistance, which is a survival marker.

These findings were surprising given that the TCGA OV dataset has posed a significant challenge for analysts to identify meaningful genome-based distinctions between the patients [38, 351] One of the most successful attempts to date was reported by Hofree *et. al.* (2013) [149] in which patient samples were clustered based on a network diffusion transformation of the mutation data. To compare the two approaches

we ran HOCUS using the TCGA OV data as filtered by Hofree *et. al.*, whereas in the main text we do not filter the mutation datasets. Our results indicate that comparable survival differences to the Hofree approach can be obtained by using a different metric (e.g. Hamming distance used here) and higher order HOCUS, eliminating the need to introduce prior knowledge (Fig. A.5). A similar result was obtained when applying HOCUS to a TCGA breast cohort (see Supplemental Information) in which both the first order and second order results revealed similar survival separation while producing different solutions. Thus, since the first and second order solutions for both OV and BRCA gave different clustering solutions but comparable outcome separation, it is possible that a solution combining first and second order solutions could produce a better outcome predictor for the patients. Furthermore, since HOCUS performed better on the OV dataset when hypermutated samples and hypomutated genes are excluded from analysis, it would be beneficial to experiment with more extensive data preprocessing.

As a final test for clustering patients using mutation data, we applied HOCUS to the TCGA BLCA cohort of 394 patients for which 84,048 mutations were called based on exome sequencing, covering 15,553 distinct genes. We inspected the BLCA clusters for novel groupings uncovered by HOCUS clustering. BLCA 2nd-order HOCUS has the largest separation in survival of the clustered patients. We note that, like the case for OV, the clusters are associated with the number of mutations per sample. Indeed, clustering by mutation rate alone yields comparable separation in patient outcomes ($P < 4e10^{-5}$; Fig. A.6) as the HOCUS solution However, the HOCUS solution also correlates with the papillary subtype and so combines the influence of mutation rate

and histology into its solution. Since mutation data was the only data used, we searched for genes with mutations that discriminate the patient clusters to understand which may underlie their different etiologies. Figure 3.3 shows the top 15 genes associated with each cluster via a χ^2 test of independence (due to overlap in the 'top' genes, only 20 genes are shown). Many of these genes are associated with several cancer types, for example LRP1B has been associated with thyroid, ovarian, renal, and brain cancers [70, 205, 324]. Other known oncogenes such as PIK3CA (p-value $3.6e^{-4}$) and TP53 (p-value $3.9e^{-8}$) are also significantly associated with the clusters. Interestingly, the highly mutated BLCA cluster has the best survival prognosis; approximately $\frac{2}{3}$ of patients surviving the entire study time period. The cluster matches well with the papillary-enriched cluster from the TCGA study. In both cases, a higher rate of TP53 mutations was found (80% compared to the background rate of 50%), and a slightly higher rate of smokers was in the category. Indeed, we compared our clusters to the TCGA BLCA clusters, which were generated using mutation and copy number data with an integrated NMF approach, and found only a weak correspondence (Fig. A.6(b), p-value 0.128). Thus, using only mutation data, HOCUS is able to automatically reproduce a solution with similar separation in survival but with a somewhat different division of the patients.

3.2.3 Community Detection Subtypes Using (Continuous-Valued) Copy Number Data

To test the applicability of HOCUS to continuous-valued data, we also applied the technique to the clustering of patients based on copy number data. We applied HO-

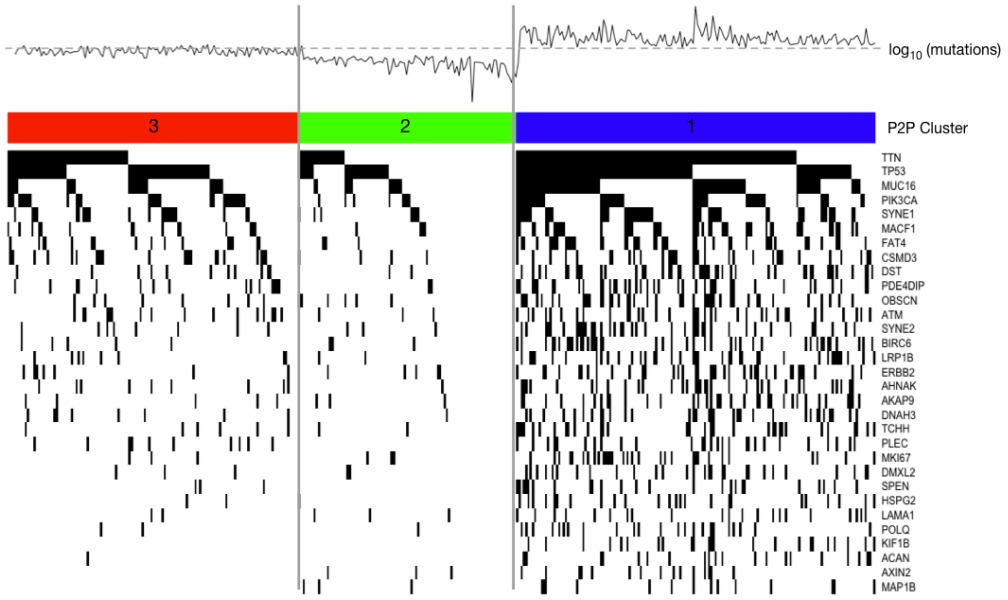

Figure 3.3: Oncoprint showing a subset of mutations in BLCA. Line plots above the oncoprint shows the total number of mutations per sample. The grey dotted lines indicate median mutational load across the cohort. This BLCA oncoprint includes genes with the smallest p-values in a χ^2 test of independence when compared to mutation rates outside the cluster. We compared each cluster to all others combined.

CUS to TCGA prostate adenocarcinoma (PRAD) copy number data because prostate cancers are known to harbor significant copy number events over the evolution of the tumor including AR amplifications, TMPRSS2-ERG fusions, and even whole genome level events such as chromoplexy. We used the output of the Broad's GISTIC2 pipeline [238] that provides gene-level associated copy number estimates as continuous-valued data with limited range of unique values. GISTIC2 scores indicate copy number aberrations, where 1 indicates low-level and 2 indicates high level amplifications, negative scores indicate the same but deletions rather than amplifications, and a score of 0 indicates no copy number alterations. For the TCGA PRAD cohort, survival rates are sufficiently high making patient survival time an inappropriate measure of disease subtype. Clusters show significant association (Fig. A.14) with Gleason score and PSA (prostate-specific antigen), both of which are associated with disease aggression.

3.2.4 Community Detection from Magnetic Resonance Imaging Data

We next applied the HOCUS method to the task of grouping patients with GBM based on the imaging of their tumors. Current practice uses MR images to localize tumors and to characterize their appearance. MR images have not been used extensively to subtype patients because it is not clear how to use the information. MR images are large and human brains have variable size and shape, making it incredibly difficult to compare between patients. By mapping to the MNI brain atlas (Montreal Neurological Institute 152), we are able to compare between patients in the cohort, and using HOCUS we are able to find clinically relevant imaging subtypes.

We applied HOCUS clustering using the GBM voxel data from the TCGA collection of 184 patients with first- and higher-order metrics to find community structures. MRI data are part of the TCGA GBM cohort, downloaded from the Cancer Imaging Archive (*www.cancerimagingarchive.net*) and processed by Stanford University as described in Liu *et. al* [218]. To reduce noise and the size of the MR images, we first preprocessed the data by filtering to a set of informative voxels containing tumor in some, but not all, of the patients (Fig. A.7). We removed all noninformative voxels mutated in fewer than 15 of the individuals from analysis. We computed sample-to-sample similarities using the remaining voxels and performed higher order calculations and clustering as described above for the mutation data (*e.g.* Hamming distance and ConsensusClusterPlus were used). Cluster solutions revealed that the metrics converged by the fourth-order (Fig. 3.5).

We sought to determine which metric based on the imaging data best matched up with the observed differences in patient outcomes. We defined the outcome-based similarity metric by computing all pairwise absolute differences between the survival time of every pair of patients, $d_{ij} = |T(i) - T(j)|$, where $T(i)$ is the survival time in days of patient i. These distances were converted to similarities via the linear transform $s_{ij} = \frac{1-d_{ij}}{m}$, where $m = max_{ij}(d_{ij})$ is the maximum absolute differences between any two patients. We then quantified the correlation between imaging-based and outcome-based similarity measures using a normalized version of the kernel alignment method [72] that calculates a centered correlation between two full sample-by-sample similarity matrices. We repeated the kernel alignment comparison to survival for 1st order and higher order

HOCUS metrics (Table A.1).

To visualize the results of the kernel alignment comparisons, we used a conditional density visualization. Through visual inspection, and reflected in the kernel alignment correlation score, we found that the third- and fourth-order had the highest scores (Fig. 3.4). Interestingly, second-order had a lower association than first-order for this dataset, illustrating the benefit of attempting higher order metrics that look beyond the immediate network neighborhood. This could indicate that, while no association may be present at a lower order, the higher order may detect associations among combinations of lower order features that could be the critical factors. We note that second-order HOCUS stratifies MR images into tumor groups by anatomic location (Fig. A.8(a)). On the other hand, third-order clusters were driven by a combination of location and volume. In addition, third-order produced a larger separation of survival in groups than location or volume alone.

Each clustering solution based on different metric orders identifies unique characteristics in the MR images that are associated with survival prognosis. While first-order clusters (Hamming similarity) align with tumor volume, and second-order with anatomic location, third-order clustering captures aspects of both tumor volume and location (Fig. 3.5(c-d)). Each solution has statistically significant separation in survival (Fig. 3.5(a)), with third-order having the greatest separation in survival of image cluster groups. Patients with tumors in the frontal lobe and which are smaller in volume have significantly better survival than larger tumors in the lower rear portions of the brain.

Interestingly, the third-order solution pulled together patients that made up

two separate poor surviving clusters in the second-order solution. To better understand the third-order subtypes revealed by the imaging data, we inspected the genetic pathways that distinguish the poorer surviving subtype from the others using RNA-Seq gene expression data. We computed a differential expression score for each gene to indicate whether a gene's expression level was higher or lower on average in the poorer surviving cluster (cluster 3) relative to the others using the Statistical Analysis of Microarrays technique [343]. We then connected any gene with an absolute differential expression higher than one standard deviation above the average of all genes. Finally, we retained pathway interactions connecting only those genes that were both in this set and plotted them with the Cytoscape viewer [305]. Several pathways involved in major growth and proliferation signaling were implicated from these networks (Fig. 3.6). ERK (MAPK1) was found to be significantly overexpressed in cluster 3 tumors along with JUN-kinase (MAPK8). In addition, AKT1 and PLK1 were also found to be higher in cluster 3, both known to drive cell cycle progression.

3.3 Discussion

As demonstrated here, community detection approaches may have merits for subtyping patients when using sparse data (few events in any single patient sample). To explore how patient-to-patient similarity transformations influence subtyping, We used a method called *Higher-Order Correlations to Uncover Subtypes* (HOCUS) that iteratively calculates higher order metrics using each similarity space to define patient

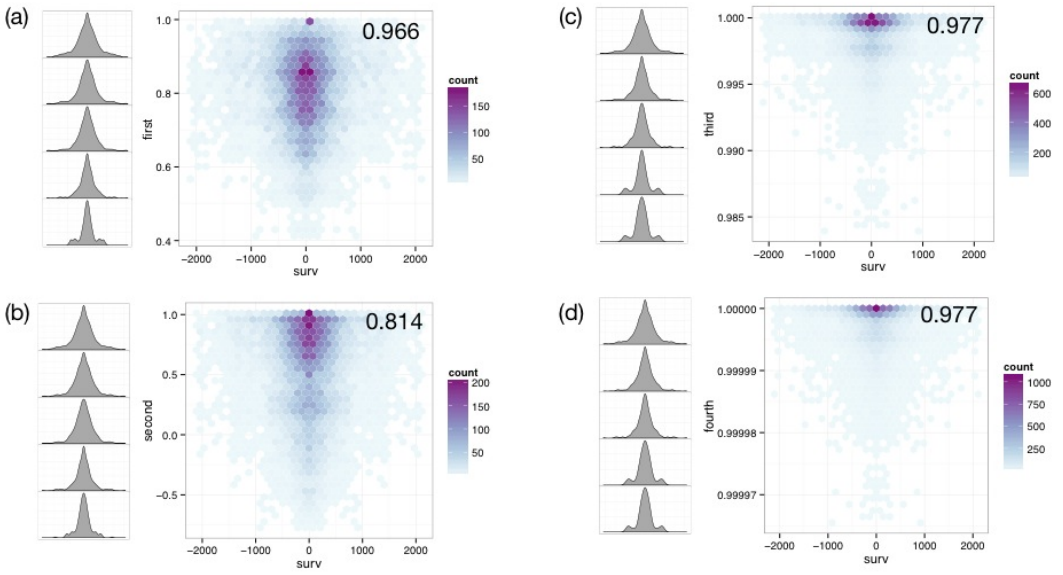

Figure 3.4: Visualization of joint and conditional densities of image-based metrics compared to survival outcome metric; results on the (a)first-order, (b)second-order, (c) third-order, and (d) fourth-order HOCUS.

clusters. HOCUS uses network connectivity to define groups or 'communities' of patients, related by both direct and indirect connections, reinforced by transitive relations in a local subnetwork. The higher-order metrics incorporate information from local neighborhoods to assess if two patient samples are related. In several cases we find that HOCUS provides an improvement over methods that use the molecular features directly to compare samples (Fig. 3.2). We find that higher order metrics yield better clusters for BLCA and GBM patients based on mutations, as well as GBM patients based on their tumor images.

We introduced a visualization method to augment the quantitative kernel alignment for identifying when a similarity measure is associated with an outcome measure of interest. The visualization inspects the conditional distribution of the outcome sim-

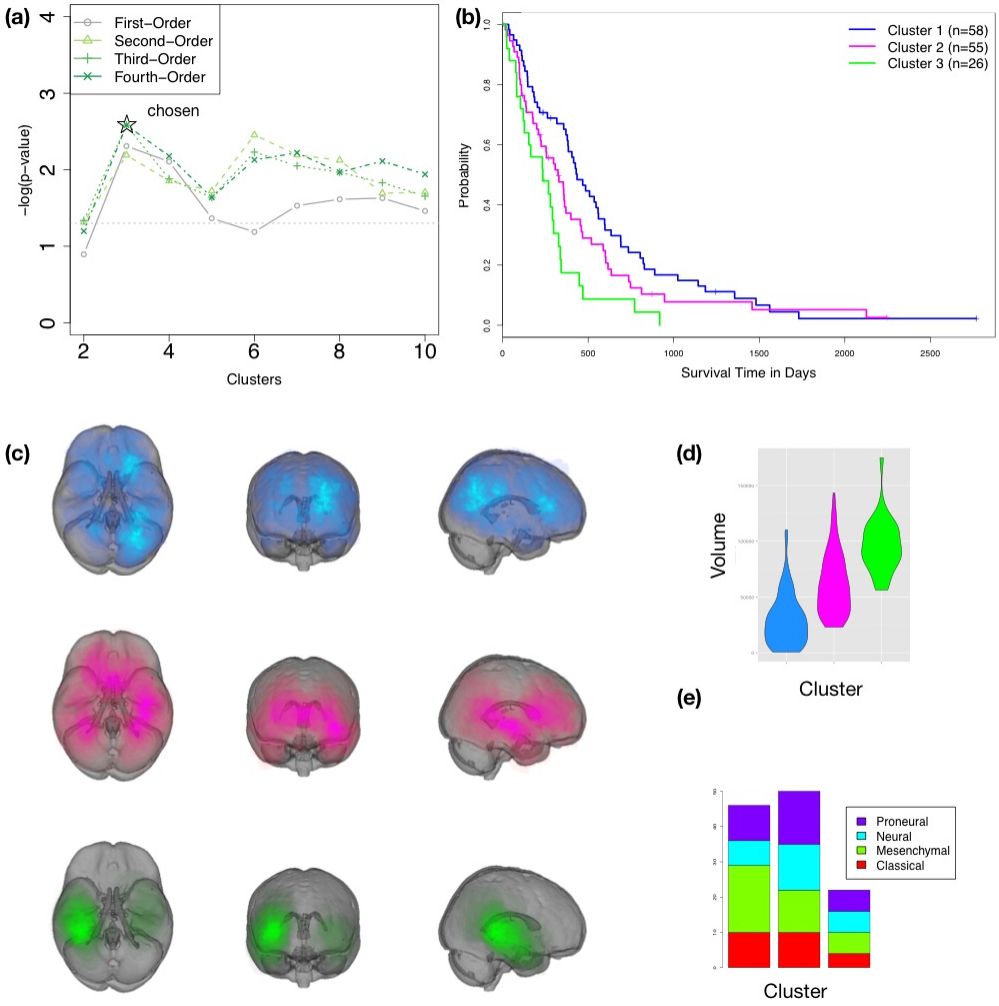

Figure 3.5: HOCUS of GBM MR Images. (a) P-values of survival separation for each of the orders of clustering across a range of k clusters. (b) Kaplan-Meier plot of the third-order HOCUS clusters. (c) Images of tumors within each cluster projected onto the MNI brain atlas. Showing saggital, coronal, axial views. Brightness of color indicates the number of patients with tumor at a given location. Generated using Slicer [97]. (d) Violin plot showing tumor volumes within each third-order cluster. (e) Molecular (gene expression based) subtypes within the clusters.

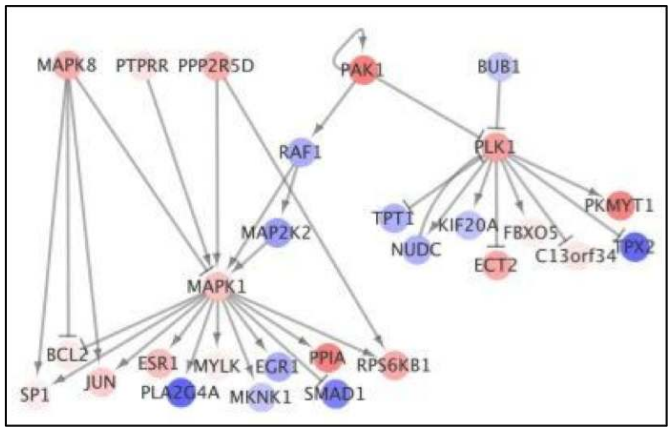

Figure 3.6: PathMark analysis of the poor surviving third-order cluster vs others. Node size and color indicates differential expression levels.

ilarities as a function of the feature-based similarities. In several cases tested, the procedure revealed that a higher-order metric was more associated with survival than non community-informed metrics. This supports the notion of using community detection techniques for the analysis of genomics and imaging data, especially given sparse feature spaces.

In the case of BLCA cancer, the second order metrics revealed groupings of the patients where tumors with higher mutation rates are separated from the other tumors and these patients have an overall better survival outcome. Most notably, the solutions for BLCA and OV separate tumors with higher mutation rates from the others and those patients with higher mutated tumors have a better survival outlook relative to the other patients. This result may reflect that highly mutated tumors are more sensitive to DNA damaging agents (*e.g.* cisplatin treatment for OV patients). Alternatively, a higher mutation rate could increase the number of neo-antigens present on tumor cell surfaces, helping a patient's innate immune system to identify and eliminate tumor cells

that lack immunosuppressive protection such as through the expression of PD-L1 and/or CTLA4. Consistent with this idea, recent clinical trials have found that combining DNA damaging agents with immunotherapies can have synergistic effects [326]. Alternatively, tumors with higher mutation rates could reflect a different subtype with an intrinsically distinct progression pattern. In support of this, we do find a somewhat higher proportion of papillary BLCA tumors in the higher mutated cluster (44% of papillary BLCA tumors are in cluster 3), but this association is not significant based on a χ^2 test.

Medical images are an underused resource that have vital information [6, 43, 46, 60, 79, 82, 134, 163, 176, 229]. However, important information for comparing tumors is clearly present in the imaging data. A key piece of data conveyed by imaging is the location of a tumor in the brain, which can influence when the tumor is detected due to the tumor affecting certain location-associated brain functions. For example, some tumors may be detected early because they reside in regions that induce extreme nausea in patients. Furthermore, some tumor locations may be more resectable than others, such as the frontal lobe and surface [304]. Thus, imaging data carries important aspects for subtyping patients according to disease outcome and treatment selection. In the TCGA data, molecular subtype is independent of both tumor location and volume (Fig. A.9, Fig. A.8(c)).

HOCUS clustering using GBM imaging data automatically clustered the patients into groups by anatomic tumor location, citing no need for expensive and time-intensive expert manual curation. Our work corroborates that of others in finding regions associated with poor survival in GBM patients as the third-order solution incor-

porated both location and size [218]. By mapping MR images to a reference brain atlas we are able to compare physical tumor characteristics between patients and combine this with more common genomic analyses. Here and in previous works [66, 218, 380] we highlight the benefit of combining genomic and image data to better predict cancer progression. We also show how location influences genomics in GBM independently of molecular subtypes. Both image and genomic data are key to understanding GBM. Of note, IGF1 was found to be the most differentially expressed gene in cluster 3. Higher levels of the insulin growth factor receptor could point to an alternate metabolic requirement for these tumors. It would be interesting to follow up on this observation by testing if the protein is present on tumor cell surfaces to support the possible role of this growth pathway. If tumor growth is dependent on this pathway than blocking IGF receptor activity may show benefit in these patients.

HOCUS is applicable to both binary (ie mutation and voxel) data and continuous data (CNV). HOCUS is simple and flexible enought to be used wherever a suitable similarity metric between individuals can be generated, even for non-sparse data such as expression or methylation data. As we have shown, its application has the potential to reveal groupings missed when using standard metrics.

3.4 Methods

3.4.1 Data Preprepocessing

3.4.1.1 MR Images

Tumor location was extracted as previously described [218]. Patient tumor was identified in the MR images by having two experts delineate tumors' regions of interest, then feeding through the image processing pipeline developed in an earlier paper submitted by the collaboration group [218]. This results in a per-patient 3-dimensional binary matrix of tumor-containing and tumor-free voxels (3-dimensional pixel) in the brain. Each 1 millimeter MR image slice was rotated and fitted to a brain atlas (Montreal Neurological Institute [93]), to make voxels comparable between patients.

3.4.1.2 Mutations

Mutation data was downloaded from firehose (*firebrowse.org*) and separated into silent/nonsilent mutations. It is translated to a patients by genes matrix of mutated vs not mutated binary information. We consider genes that have at least one nonsilent mutation within the cohort and patients with at least one nonsilent mutation.

3.4.1.3 Copy Number

GISTIC2 [24] copy number variation data was downloaded from firehose (*firebrowse.org*). Patient-patient networks were calculated based on Hamming distance us-

ing the GISTIC scores ($[-2, -1, 0, 1, 2]$). Thus any similarities between patients are considered a match. For example, -1 would not be considered a match to a score of -2.

3.4.2 Visualization of Joint Densities

To visualize the association between feature- and survival-based measures, we plotted the the proportion of sample pairs with similarities in both metric spaces. If the distribution of survival similarities for sample pairs changes as a function of the feature-derived similarities, it suggests that the feature-based metric carries outcome-relevant information. For example, if we restrict the pairs to those with high similarity in mutation space and we observe that there are more pairs with similar survival compared to the background (or to pairs with low mutation-based similarity) it would indicate mutation-based similarity carried information about survival outcome. To view such a dependency, we group sample pairs into bins of approximately equal feature-based similarity. Then, for each bin, we plot the distribution of outcome similarities, shown along the left-hand side of each joint density plot. A distribution that changes significantly across the bins reflects an association between the feature- and outcome-based similarities. In the case of patient survival, we are interested in whether higher similarities computed from the feature data reveal a higher proportion of patient pairs with similar survival times.

3.4.3 Community Detection Using Higher-Order Sample Similarities

Our analysis is similar to the common inference-by-transitivity technique used in social networks, summarized by the statement 'a friend of my friend is also my friend.' This technique finds cliques of similar patients in a network by connecting patients that are similar in the original network and then clustering based on those similarities. Given samples j and k, and feature vectors X, we calculate the similarity matrix $S^{(1)}$ (using Hamming similarity ($S^{(1)}_{(H)}$) when the features are binary such as for mutations and imaging voxels).

$$S^{(1)} : s^{(1)}(j,k) = \frac{1}{n}\sum_{i=1}^{n} I(x_{i,j}, x_{i,k}), \tag{3.1}$$

where n is the number of features (e.g. voxels), $I(a,b)$ is the indicator function that returns 1 if its first argument equals its second and returns 0 otherwise. Using this similarity metric, we compute the 2nd-order similarities from the 1st-order matrix. Let m be the number of samples in the cohort. The second order metric is calculated as:

$$S^{(2)} : s^{(2)}(j,k) = \frac{\frac{1}{m}\sum_{l=1}^{m} S^{(1)}(j,l) \times S^{(1)}(l,k)}{\sqrt{\sum_{l=1}^{m} S^{(1)}(j,l) \times \sum_{l=1}^{m} S^{(1)}(l,k)}}$$

$$= corr(S^{(1)}(j,*), S^{(1)}(*,k)). \tag{3.2}$$

For higher-order clustering, the precomputed similarity matrix is raised to the d power, where d is the order of clustering. Because the ConsensusClusterPlus R package [372] that we used computes an internal metric prior to clustering and only takes as input a feature matrix, we would raise the matrix to the $(d-1)^{th}$ power and supply this

matrix as the feature matrix to ConsensusClusterPlus as input. Using centered Pearson Correlation as the metric is then equivalent to squaring the feature matrix. In this way, we tested all even powers of d when using ConsensusClusterPlus. For example, our "third order" solution effectively uses a fourth order metric since $S^{(2)}$ is squared and our "fourth order" solution is actually a sixth order metric since $S^{(3)}$ is effectively squared inside the ConsensusClusterPlus package.

3.5 Integrated Analysis using HOCUS

Our next step was to use HOCUS in an integrated data setting. Section A.1 outlines experiments where we used HOCUS on non-binary data such as CNV, and a similar and highly cited approach called WCGNA (Weighted correlation network analysis) is designed for co-expression data [200]. Thus the approach can be applied to many types of data individually.

There is a ongoing project at UCSC called Tumor Map (Newton 2016, in review) [250], where correlation networks of different types of cancer data are being used to visualize 'maps' of cancer genomes. Tumor Map uses the Google Maps framework to visualize patient–patient similarities in a 2D space. Correlation networks can be combined using this method. Map layout is determined by the data layers, and HOCUS is used for the mutation layers.

The mutation layer is built from 313 high–confidence mutation calls within 2 TCGA studies [174, 325], both of which analyze the PanCancer12 data [39]. Applying

HOCUS to this data transformed the layer from one large 'island' to many islands (Figure 3.7(a)). This indicates HOCUS was able to find distinct mutation–derived subgroups within the data. Other approaches were unable to differentiate between patient groups because of the sparsity of the mutation data. Since HOCUS also uses indirect shared mutations, it is able to find distinct patient groups. Of note, HOCUS identifies a key difference in the COAD&READ tumors (Figure 3.7(b)). The most frequent mutation differentiating these groups is KRAS, however that single mutation is not the only difference– were that the case, the groups would be physically closer in the Tumor Map. There is a network of indirectly shared mutations within each group that makes them distinct.

While the mutation–only maps are of interest, Tumor Map is able to combine many maps into one integrated data map. The HOCUS mutation maps are combined with somatic copy number to create what we call the 'Genome Space' map, and with several combinations of data platforms in the 'Integrated Space' maps. These maps are described in detail in Newton *et. al.*, currently in review. Tumor Map visualizes integrated data and incorporates the (sparse) mutation data by using HOCUS to compare samples. Thus, HOCUS is becoming part of an integrated analysis framework.

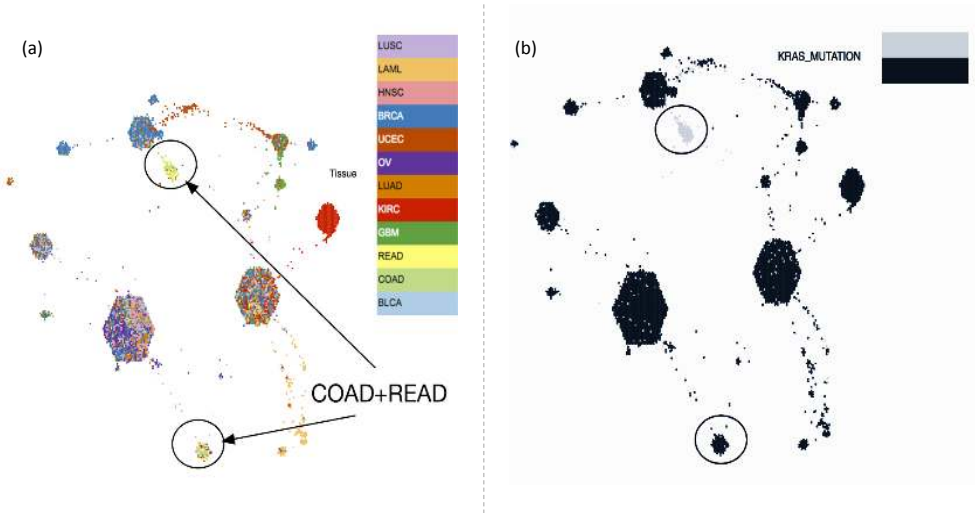

Figure 3.7: PanCan12 Mutation Map using HOCUS identifies a KRAS–dependent subtype in COAD&READ cancers. In (a) samples are colorcoded by TCGA–defined cancers by tissue type and in (b) samples are colored by presence of a KRAS mutation.

Chapter 4

Integrative Clustering Analysis

4.1 Introduction

In this chapter I transition from the single data platform clustering in Chapter 3 to integrated analysis. TCGA network and others recently have shown the benefits of integrated clustering methods, for example iCluster [298] and PARADIGM [347]. I participated in several of these projects, and this chapter covers a few of them. First is an ongoing project about an asbestos–related cancer, mesothelioma. Second is another TCGA project, hepatocellular carcinoma, whose manuscript is in review. Last is a meta–analysis of prostate cancer datasets, initiated by my work in the TCGA prostate adenocarcinoma group but which includes 8 prostate cancer studies. The last project was done together with Yulia Newton.

This chapter contains text from the paper 'Integrative Molecular Characterization of Malignant Pleural Mesothelioma' (TCGA Network 2016, in prep) in Section 4.2,

and the paper 'Comprehensive and integrative genomic characterization of Hepatocellular Carcinoma' (TCGA Network 2016, in review) in Section 4.3. I present my integrated analysis contributions to these cancer working groups as well as highlights from the manuscripts. I focus on the UCSC contributions, and encourage readers to access the complete published papers.

4.2 Integrative Molecular Characterization of Malignant Pleural Mesothelioma

4.2.1 Introduction

We report a comprehensive molecular analysis of 74 primary, non-pretreated Malignant pleural mesothelioma (MPM) samples. The sex ($\frac{62}{74}$, [83%] male), age (median 64 years) and tumor histological type ($\frac{50}{74}$, [67%] epithelioid) distributions in our cohort are typical of MPM [13, 258]. Recurrent somatic mutations were detected in BAP1, NF2, TP53, LATS2 and SETD2, all known drivers of MPM. Moreover, a significant number of cases were found to have extensive ($> 50\%$) loss of heterozygosity (LOH). Among these, 3 were found to have $> 80\%$ LOH.

4.2.2 Methods

We used median centered, log scaled mRNA expression and SCNA GISTIC2 [238] data to calculate inferred pathway activity levels using PARADIGM [347]. I clustered the PARADIGM data using ConsensusClusterPlus [372], and identified 4 distinct clus-

ters. To compare Cluster 1 (worst prognosis) to Cluster 3 (best prognosis), I ran PathMark [143] on the statistically significant differential activities obtained from SAM to extract connected components of the global PARADIGM regulatory network. Activities that fall outside 2 standard deviations of the empirical distribution of the statistically significant differentials are included the final result. A network connection is extracted if both vertices connected by that connection pass the filter. Networks are then visualized using Cytoscape [294] and CircleGraph (Figure 4.2).

4.2.3 Results

4.2.3.1 Pathway Level Expression Changes

The most up-regulated network difference between the worst and best prognosis groups is centered around AURKA (Fig. 4.2(b)). Fig. 4.2(a) shows the full PathMark network found and highlights several subnetworks of interest. Best prognosis cluster patients have upregulated androgen receptor and TP63 networks. Furthermore, the poor surviving cluster has increased expression in several subnetworks commonly associated with aggressive disease– ERBB, PLK1, VAV1, and the Alpha/Beta integrin subnetworks. The good prognosis subgroup is relatively copy-number quiet and all but one patient have BAP1 alterations (compared to 50% in other groups). This is true even when we compare only the epithelioid samples.

Across multiple tumor types, MPM possessed the second highest overall EMT score after sarcoma. EMT score correlated with histology, with lower EMT scores in epithelioid MPM. Cluster 1 patients have higher EMT scores and Cluster 3 have lower

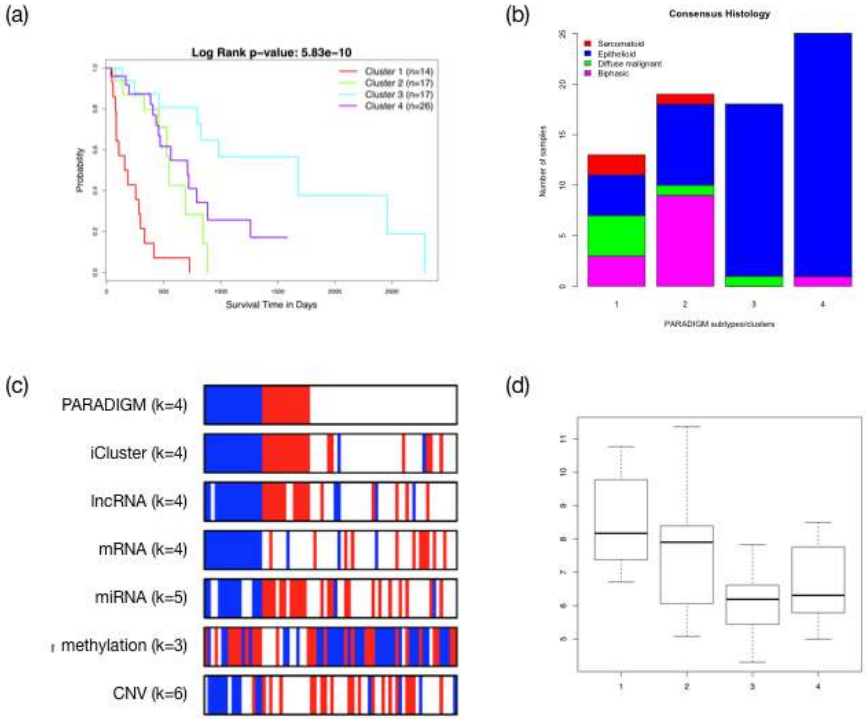

Figure 4.1: Overview of the PARADIGM results. (a) KM of PARADIGM clusters. (b) Histology enrichment. (c) 'Best' (blue) and 'worst' (red) survival groups recapitulated in the single platform clusters. (d) EMT scores by cluster.

scores than average within the cohort (Fig. 4.1(d)).

Cluster 1 patients have significantly worse survival and have higher in AURKA, E2F targets, G2M checkpoints, as well as PI3K and mTOR pathway expression. A drug currently in a phase 2 clinical trial inhibits AURKA activity [47] and may be an effective treatment. Furthermore, this cluster recapitulates platform-specific clustering of miRNA data (Figure 4.1(c)).

Cluster 3 patients have the best prognosis, and have upregulation in EGFR signaling; Kinase subnetworks are downregulated. There is an enrichment of epithelioid patients in this cluster (Fig. 4.1(b)), however after correction for this enrichment, the

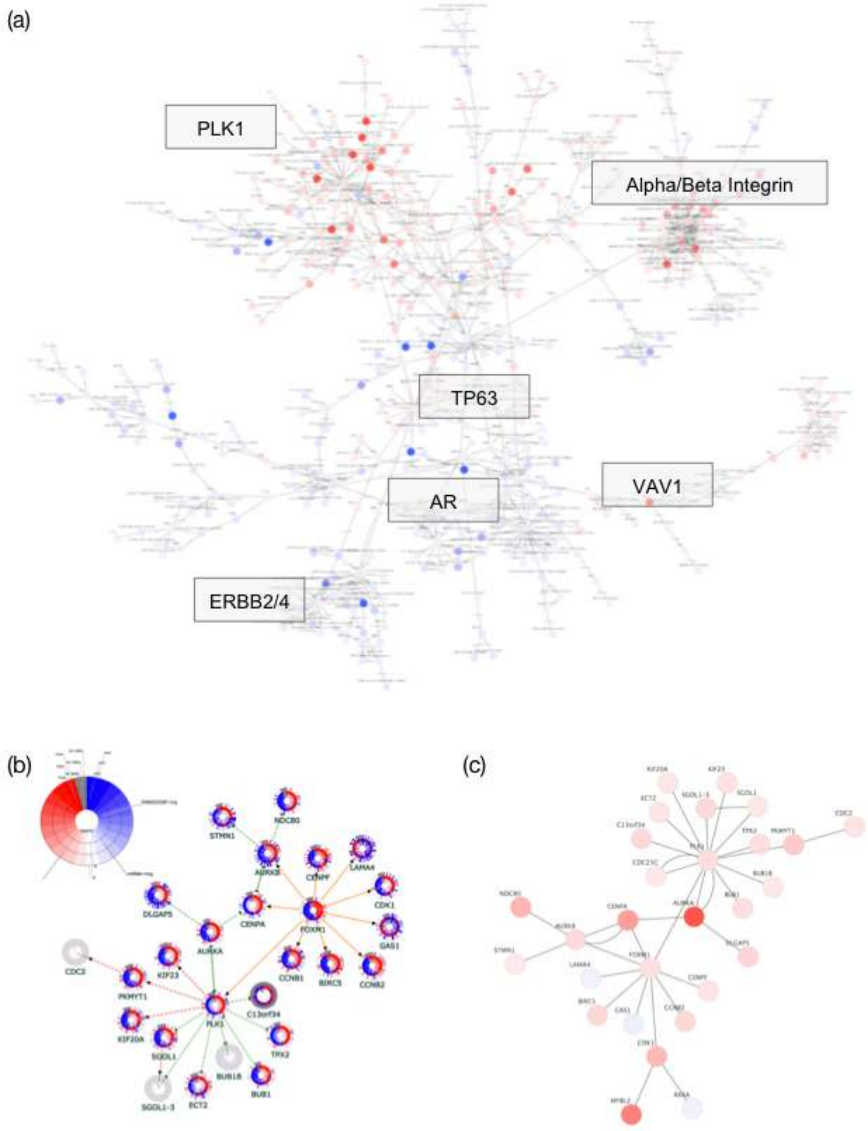

Figure 4.2: (a) Overview of the PathMark results. This shows the connected subnetwork fo genes that are greater than 2 standard deviations outside of normal expression of patients within the best and worst surviving clusters. Red means upregulated in the poor prognosis group and blue is upregulated in the best prognosis group. Brightness of color shows the degree of difference between the two groups. (b) Differential analysis finds this AURKA subnetwork upregulated in the worst prognosis cluster. Showing circleMaps with PARADIGM cluster, PARADIM IPL, mRNA expression, and CNV data for each patient. (c) PathMark subnetwork for AURKA.

Table 4.1: TCGA tumor types included in the PanCancer Tumor Map.

Name	Acronym	# Patients
Malignant pleural mesothelioma	MPM	74
Sarcoma	SARC	469
Skin cutaneous melanoma	SKCM	469
Lung adenocarcinoma	LUAD	516
Lung squamous cell carcinoma	LUSC	501
Uterine corpus endometrial carcinoma	UCEC	194
Uterine carcinoma	UCS	57
Basal breast invasive carcinoma	BRCA	143
Total		3,176

group remains distinct. Patients are also more likely to have undergone pneumonectomy. Patients in clusters 2 and 4 have similar prognosis, but are genomically distinct. Cluster 2 patients have lower stage (N1) and cluster 4 is predominately epithelioid.

4.2.3.2 Tumor Map

PARADIGM data is projected onto a Tumor Map to visualize the similarities between patients (Fig. 4.3(a)). The Tumor Map represents a dimensionality reduction and visualization method for high dimensional genomic data (see Newton *et. al.* for a detailed explanation of the method). Samples are arranged in a 2D space and then assigned to hexagons in a regular grid. Relative distances in the map approximate relative similarities between the samples, so that samples with similar genomic profiles are placed near each other in the map. Thus, clusters of samples that appear as 'islands' in the map share genomic and/or epigenomic events.

We built a multiple cancer map using mRNA expression from 3,176 patients across 8 cancer types (Table 4.1). Map layout is constructed using the 6 nearest neigh-

bors for each sample based on pairwise similarity. While the driving factor in the map was the tissue of origin, we found that MPM clustered near SARC tumors (Fig. 4.3(b)). Furthermore, some of the MPM samples co-clustered with the SARC tumors in that group. This SARC group is enriched for undifferentiated phenotypes, specifically dedifferentiated liposarcoma and undifferentiated 'pleomorphic' sarcoma. MPM tumors that directly clustered with or near these undifferentiated sarcomas are enriched for biphasic and sarcomatoid histology and belong to the poor survival subtype defined by the PARADIGM analysis. This result leads us to hypothesize that poor prognosis MPM might be associated with de-differentiation and stem-like molecular signatures.

4.2.4 Conclusions

Comprehensive molecular characterization of 74 MPM cases confirms that MPM is driven by loss/inactivation of tumor suppressors, not by aberrant activation of oncogenes. MPM in the PanCan analysis (Fig. 4.3(b)) shows an association with dedifferentiation. These cases are marked by low mutation rate and gene expression profile. They cluster close to sarcomas, have high EMT score, and high VISTA (C10orf54) expression. The PARADIGM clusters offer an update on diagnosis that may eventually augment/replace histologic subtypes.

Unsupervised clustering showed good concordance across several analysis platforms (Fig. 4.1(c)) and several potential therapeutic targets are supported by our findings, of note the AURKA pathway and VISTA. Integrated clustering was able to identify a signal that persists in part in every data platform, but is most clear in the PARADIGM

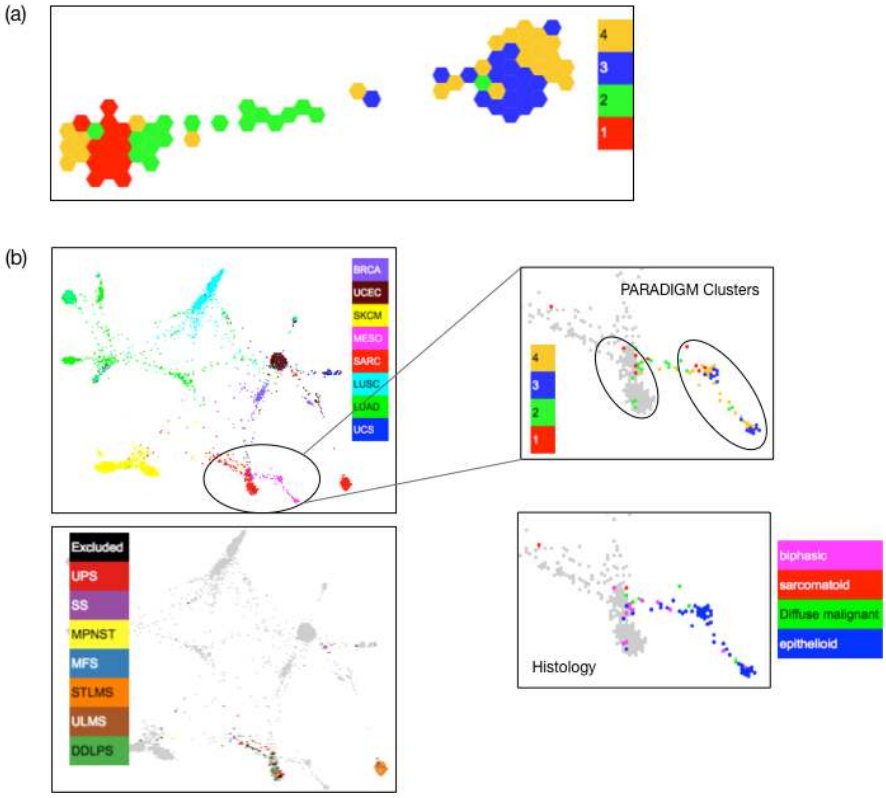

Figure 4.3: Tumor Maps for the MESO project. (a) PARADIGM map colored by PARADIGM clusters. (b) PanCancer-8 map showing the Sarcoma-like MESO tumors. Breakout windows show PARADIGM clusters, histology, and dedifferentiated SARC types.

results.

4.3 Comprehensive and Integrative Genomic Characterization of Hepatocellular Carcinoma

4.3.1 Introduction

Hepatocellular carcinoma (HCC) is the second most common cause of death from cancer worldwide. There have been 700,000 annual deaths recorded globally in recent years [381]. HCC has several known risk factors including chronic HBV and HCV infections, autoimmune hepatitis, diabetes mellitus, obesity, alcohol abuse, and several metabolic diseases [96]. There has been a worldwide rise in HCC incidence, and in developed nations this is partly attributed to its association with known risk factors such as obesity and diabetes [96, 381]. While initiation and progression of HCC is considered a multi-step process, the underlying driver mutations and molecular events remain only partially understood.

Recent HCC genomics studies have identified frequent mutations in TERT, TP53, and CTNNB1 (β-catenin)[3–8]. Response to Sorafenib, a kinase inhibitor which is the only drug approved for HCC management, can be predicted based on FGF3/4 and VEGFA amplifications [222]. Unfortunately, since its approval more than ten other drugs have failed to meet clinical end points in phase III trials. Thus there is a need for new drug discovery for HCC [221].

As part of The Cancer Genome Atlas (TCGA) project we have analyzed ge-

nomic data from 196 HCCs to understand the genomic landscape of HCCs. The recognition of new mutations and the characterization of robust subclasses with prognostic implications in this study have the potential to influence clinical management of HCC and target identification for drug discovery. To this end, UCSC provided integrated PARADIGM analysis and identified both clinical and genomic events related to the pathway-level differences between patients.

4.3.2 Methods

PARADIGM [347] was run on 188 cases with mRNA expression and copy number data. Expression data was $\log_2$ scaled and median-centered; copy number was taken from the GISTIC output. We then used consensus kmeans clustering [372] to cluster PARADIGM IPLs with greater than 0.5 standard deviation, using pearson correlation.

Clustering identified 5 distinct PARADIGM clusters. Fig. 4.4(a) shows the difference in survival between groups, and Fig. 4.4(b) shows sample–sample similarities in PARADIGM space; Several of the platform–specific clustering solutions overlap with the PARADIGM solution. While there is little difference in survival between most groups, the cluster 3 patients have the worst prognosis. To identify genomic differences between these patients and the others in the cohort, we run differential expression analysis on the PARADIGM IPLs, comparing cluster 3 patients to the entire cohort. PathMark computes SAM [61] scores then projects connected subnetworks onto the superpathway used in PARADIGM. From this we extract connected networks that are

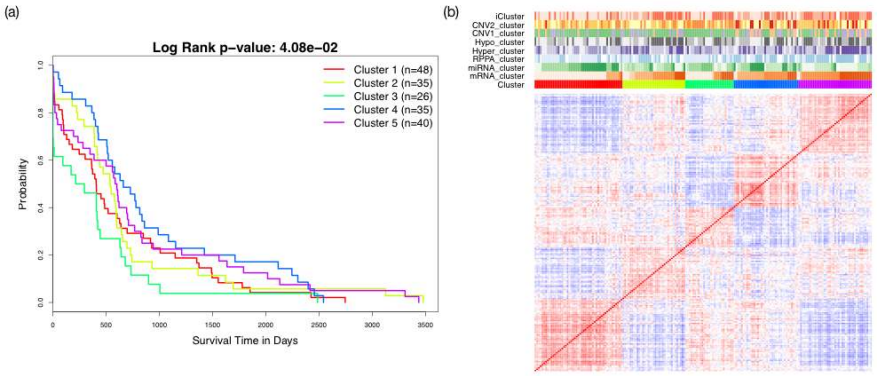

Figure 4.4: PARADIGM clusters: (a) KM plot and (b) sample-sample IPL similarity matrix ordered by PARADIGM cluster and annotated with platform–specific clusters.

significantly differently expressed in the poor surviving group.

4.3.3 Results

PARADIGM cluster 3 has the worst survival of and contains the majority of HBV+ patients. Furthermore, the HBV infected patients are much younger than the others (Fig. 4.7(c), median 53 vs 65 years). PARADIGM clusters are also enriched for obesity, BMI, and grade (Fig. 4.7(a-b,d)). Obesity and high BMI are known risk factors for HCC.

The group found that some samples appear to be HBV+ in mRNA expression but are not clinically HBV. We were unable to determine if these patients had an unknown HBV infection. Similarly, the group identified a TGF-β factor associated with poor survival, reminiscent of triple–negative breast cancer. When comparing the worst surviving PARADIGM cluster (3) to the other clusters, we identified a proliferation subnetwork expressed much higher than expected (Fig. 4.6(a)). Another analysis

(Fig. 4.5) found a prevalance of TP53 mutations in cluster 2 whereas PathMark analysis of PARADIGM IPLs show the TP53 subnetwork to be downregulated in cluster 3 (Fig. 4.2(b)).

4.3.3.1 TP53 Pathway Alterations

Mutations involving TP53 were found in 31% (n=60) of patients. Here we used an alternate methodology to determine p53 functional status by assessment of p53 target gene expression. The degree of p53 target gene upregulation is used as a surrogate for p53 functionality. Tumours were stratified based on p53 target gene expression (Fig. 4.5(a)). While virtually no HCCs with high p53 target expression had TP53 mutations, 11 out of 48 (23%) samples in the low p53 target expression were TP53 wildtype. Thus, many HCCs without TP53 mutations appear to have dysfunctional p53, consistent with the known existence of non-mutational p53 inactivating mechanisms [310]. We examined specific inhibitors of p53 function and found that MDM4 was significantly increased in copy number and expression in low signature WT TP53 HCCs relative to other HCCs ($p = 3.6 \times 10^{-4}$ and $p = 5.4 \times 10^{-4}$, respectively) (Fig. 4.5(a)). Thus, increased MDM4, a molecule that binds to p53 and inhibits its transcriptional functions [230], may provide a mechanism for low p53 signatures in non-TP53 mutated HCCs [85].

We further analyzed clinical and molecular correlations with the p53 expression signature clusters. Tumors having low p53 target expression exhibited significant associations with increased copy number instability (including high frequency chromosome

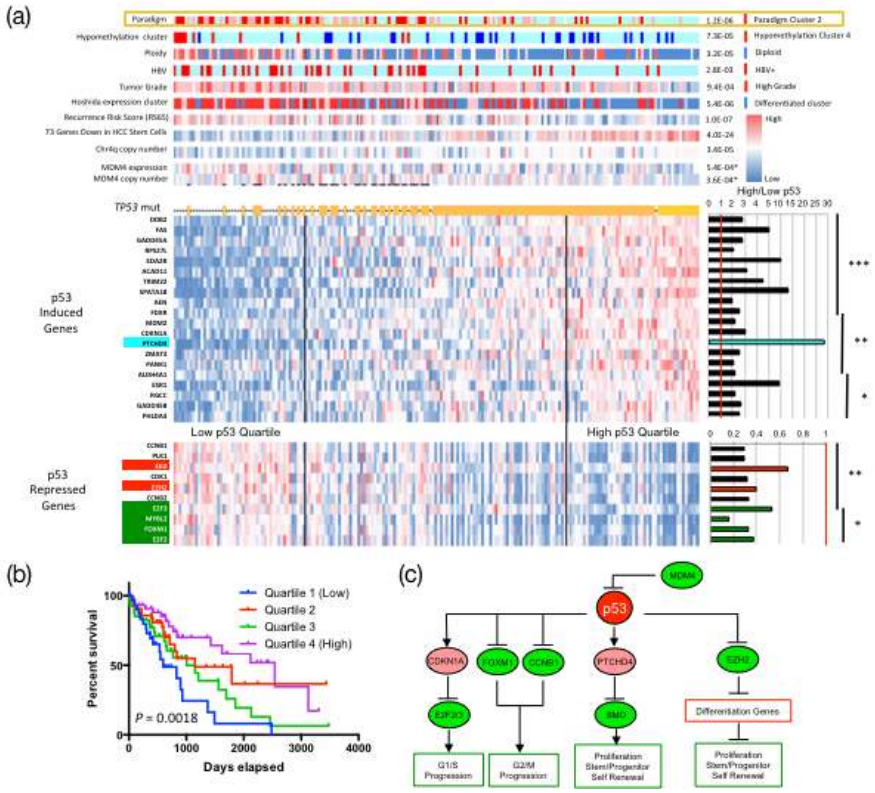

Figure 4.5: P53-induced gene target expression signature, with biological attributes and clinical outcomes. PARADIGM Cluster 2 is enriched for the TP53 signature found by the TCGA working group. (a) Clustering of 191 HCC by expression of 20 known p53-induced target genes that are frequently upregulated in HCC with wildtype TP53 relative to mutant TP53. Ranked lowest to highest by composite signature expression. We include 20 induced targets and 10 p53-repressed genes. (b) overall survival of the low, high, and intermediate quartiles of the p53 signature (c) model of key pathways likely regulated by the p53 signature, effecting clinical and molecular parameters.

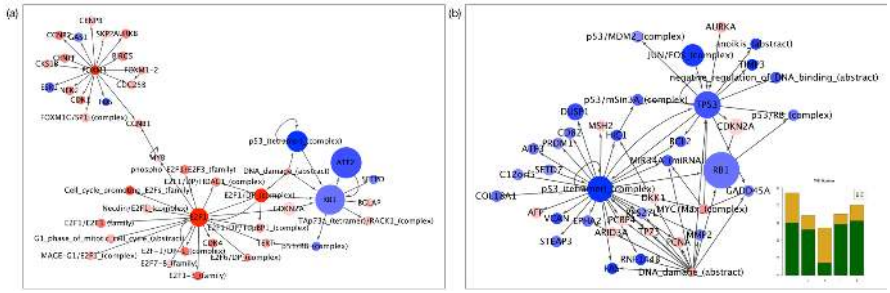

Figure 4.6: Differential analysis using PathMark identified (a) a TP53 subnetwork and (b) a proliferation subnetwork.

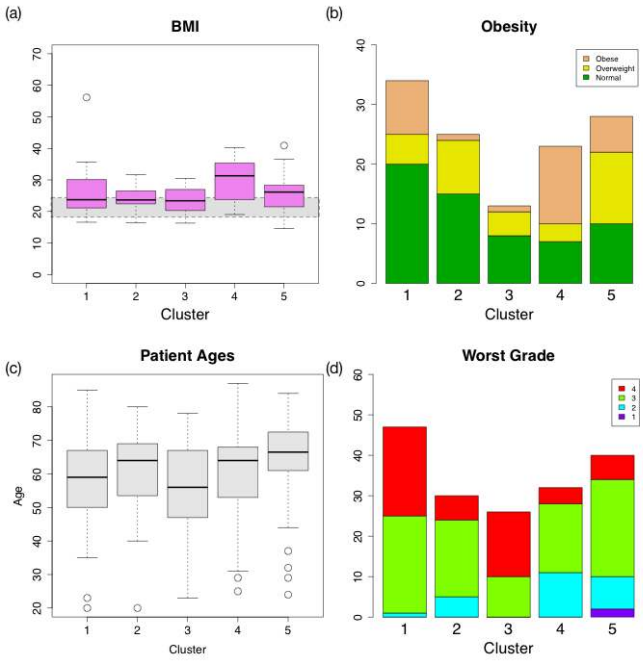

Figure 4.7: PARADIGM clusters are enriched for (a) BMI and (b) obesity. Obesity is a known risk factor for HCC.

4q loss71), higher pathological grade, reduced expression of mature hepatocyte marker genes, and increased risk of tumor recurrence (Fig. 4.5(a)). The lowest p53 signature quartile patients had a median overall survival of 596 days versus 2,542 days for the highest quartile ($p = 0.0018$) (Fig. 4.5(b)). Among the p53-regulated HCC target genes PTCHD4 showed a 28-fold increased expression in the highest p53 expression quartile relative to the lowest p53 quartile (Fig. 4.5(a)). PTCHD4 suppresses sonic hedgehog (SHH) signaling in colorectal cancers [63] and SHH signaling is important in liver regeneration and HCC [36]. Expression of SHH pathway gene expression was significantly upregulated in low p53 signature tumurs relative to high p53 signature tumors by GSEA analysis [316].

Another p53-repressed target gene, EZH2, was significantly upregulated in low p53 signature HCC (Fig. 4.5(a)). EZH2 encodes a histone methyltransferase that epigenetically regulates stem cell maintenance [357] and its enhanced expression in low p53 signature HCC coincides with increased stem/progenitor gene expression. The low p53 signature HCC also showed increased expression of the p53-repressed cell cycle regulatory genes CCNB1/2, E2F2/3, and FOXM1, consistent with enhanced stem cell gene expression phenotypes, and robust global upregulation of G1/S and G2/M promoting genes. We hypothesize that p53 regulates HCC phenotypes through at least three major signaling arms, the sonic hedgehog pathway via PTCHD4, the polycomb repressive complex 2 via EZH2, and cell cycle progression pathways via cell cycle regulatory genes (Fig. 4.5(c)).

4.3.4 Conclusions

This comprehensive integrated analysis of 196 hepatocellular carcinomas enhances our understanding of the molecular genetic events relevant to this cancer. The mutation and pathway analyses provide potential directions for future therapeutic efforts. Aside from the RTK inhibitor, sorafenib, no targeted therapies are clinically available for this disease [220, 239]. We showed that WNT or p53 signaling or the telomerase promoter are altered in 77% of HCC. WNT pathway small molecule inhibitors are currently in preclinical and clinical development [264]. Moreover, targeted approaches that restore wildtype p53 activity to tumours with TP53 mutations have been in clinical testing [57]. Because p53 can be rendered dysfunctional by alterations

in upstream regulator function (e.g. MDM2, MDM4, p14ARF), p53 signature analysis may provide a more accurate representation of p53 functional activity and may better predict clinical outcomes than previous mutation-based studies.

Interestingly, we showed that a significant fraction of HCC with WT TP53 have elevated MDM4 expression, hence currently available MDM4 small molecule inhibitors might be efficacious in these HCC [166]. The very high frequency of TERT promoter mutations suggests that upregulated TERT expression in HCC might be targeted with telomerase inhibitors currently in development and clinical testing [281]. The activated TGF$-\beta$ signature observed in a high fraction of HCC indicates that TGF$-\beta$ signaling presents an attractive target, and this is supported by preliminary studies showing that TGF$-\beta$ inhibitors have HCC anti-tumor activity in initial clinical trials [110].

Computational pathway analysis of less frequently mutated genes implicated alterations in the SHC-RAS-MAPK related pathways, consistent with sensitivity of HCC sensitivity to the RTK inhibitor Sorafenib. The high expression of immune checkpoint genes CTLA-4, PD-1, and PD-L1 in 20% of HCC make this subset of tumours particularly attractive candidates for monoclonal antibody-based therapies specifically targeted to these genes [146, 201, 269, 339]. In conclusion, established and novel analytic approaches have been applied to multiple data platforms from a large number of clinically annotated HCC to provide a better understanding of molecular targets that may lead to better therapeutic strategies.

4.4 A Signature of Metastasis in Prostate Adenocarcinoma

4.4.1 Introduction

Prostate Adenocarcinoma (PRAD) is the most common form of prostate cancer, with 180,890 diagnoses and 26,120 deaths estimated in 2016 [301]. One in every seven men will be diagnosed with the disease in his lifetime [301]. Together with Yulia Newton, I analyzed tumor progression in a combined set of prostate cancer patients. Our goals were to identify patients whose disease are likely to metastasize and to identify genomic mechanisms causing metastasis. To do this we combine the datasets, determine subtypes for primary and metastatic samples, train predictors and finally use those to link the primary and metastatic subtypes (Fig. 4.8). Tumor cells undergo very specific molecular changes during invasion and metastasis that allow cells to be detached from the tumor, travel to another location and successfully start a new colony. Some of such changes include epithelial-mesenchymal transition (EMT) in preparation for invasion or intramural growth and micro-colony formation by circulating tumor cells (CTC). As a result, metastatic tumors have distinct molecular signatures that are responsible for activating signaling pathways in the cell that are specific to metastasis. For example, NOTCH signaling is known to be activated in metastatic tumors [268, 393]. We identify early metastatic signature that can be detected in primary tumors, predicting likelihood of metastasis.

4.4.2 Methods

Dataset	Normals	Primes	Mets	Genes	Platform
Cai [35]	0	22	29	10,523	Microarray
Chandran [50]	0	10	21	14,997	Microarray
Grasso [125]	28	59	32	15,830	Microarray
GTEx [223]	42	0	0	13,256	Microarray
Monzon [345]	52	65	25	9,383	RNASeq
Taylor [329]	29	131	19	19,923	Microarray
TCGA [41]	21	246	0	20,500	RNASeq
Erho [91]	0	545	0	20,500	Affy Human Exon
Joint		172	1,078	126	4,895

Table 4.2: Datasets used in the meta-analysis.

Preprocessing the Data We combined the eight datasets in Table 4.2 into a new dataset of 1,659 samples and 4,894 genes. Dataset batch effect was removed with an Empirical Bayes approach (ComBat, [167]), which was given dataset as a batch and the type of sample (normal, primary, or metastatic) as biological covariates. Fig. 4.9 shows principle component analysis of the data before and after ComBat, showing that the batch effect has been successfully eliminated.

Subtyping Primary and Metastatic Prostate Adenocarcinoma In order to identify molecular subtypes of the primary and metastatic prostate adenocarcinoma we performed consensus k-means clustering [372] on the ComBat-transformed data. Clustering was performed using 1,313 genes after variance filtering. The 785 primary samples and 126 metastatic samples were clustered separately.

Based on silhouette score, we identified four primary and three metastatic

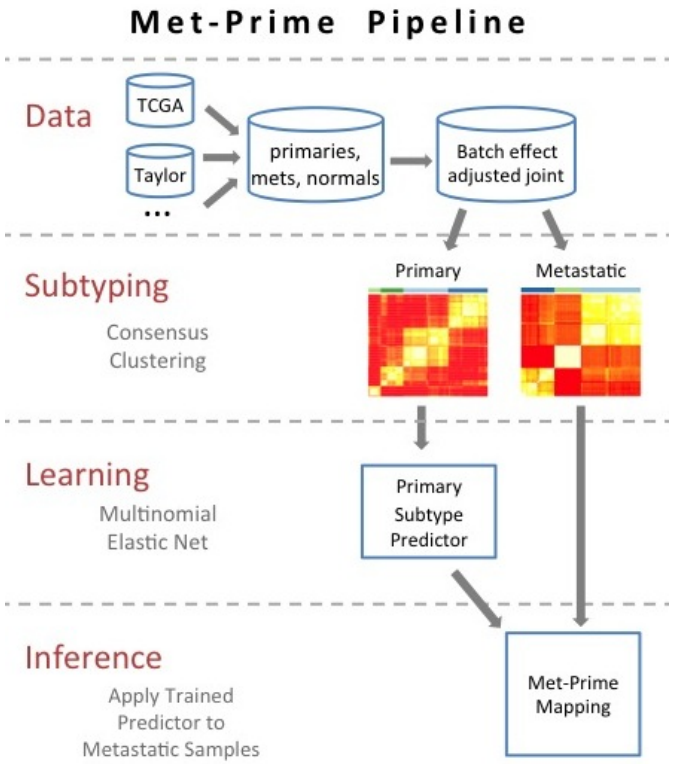

Figure 4.8: Workflow for predicting metastatic signal in primary samples.

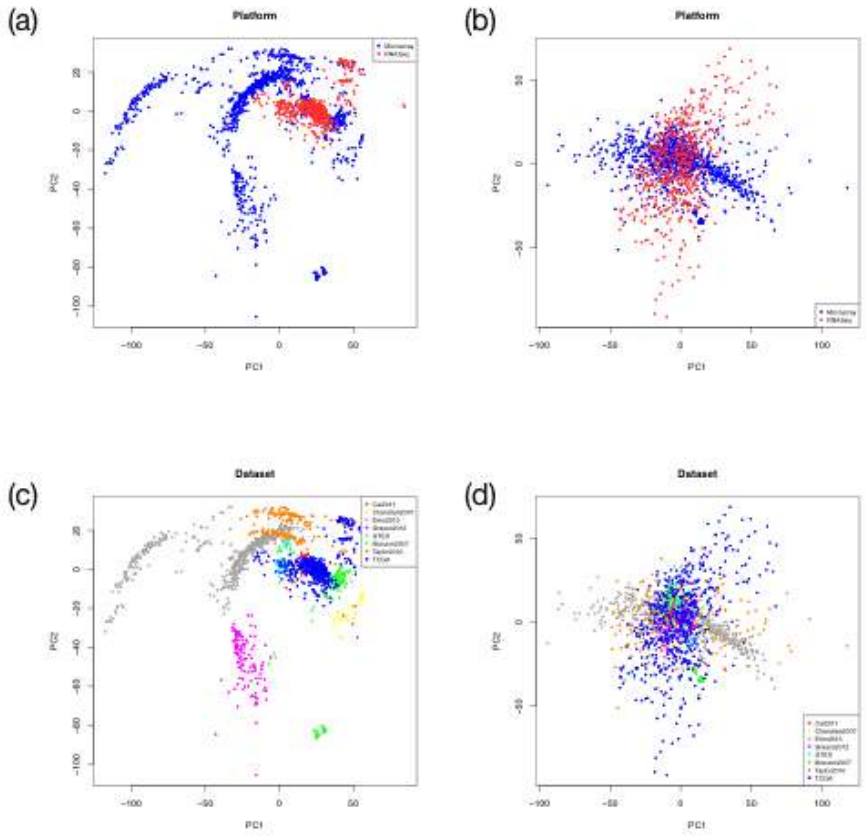

Figure 4.9: PCA plots of the mRNA expression data (a) before and (b) after ComBat application for batch effect removal with respect to the dataset source distribution in the data. Also colored by platform distribution (c) before and (d) after ComBat application.

subtypes (Fig. 4.10). In most cases, subtypes are dataset and platform independent. Grasso2012 is the only exception, however this is expected since the Grasso2012 dataset was obtained from autopsy biopsies and the rest of the metastatic samples were live biopsies. We hypothesize that this Grasso2012 cluster reflects true biology of samples from a dead tissue, exhibiting different molecular signal than sampled from live patients.

Primary Subtype Predictor Applied to Metastatic Samples We trained a multi–class elastic net model on the primary data using the glmnet package in R [102]. Models were trained to predict the primary cluster memberships. Success rates for each class is high, and the balanced success rate (BSR) is 0.991 (Eq. 4.1).

$$\sum_{i=1}^{n} \frac{\text{tpr}_i}{\text{pos}_i} \tag{4.1}$$

We next applied the trained primary subtype predictor to the metastatic samples, to find predicted primary cluster for metastatic samples. The ribbon plot in Fig. 4.10 shows the distribution of predicted primary cluster within each metastatic cluster. While the metastatic clusters do not correlate with predicted primary subtypes, the majority of metastatic samples are predicted as primary subtype 2. We call these the met-like primaries. Fig. 4.11 shows enrichment of clinical covariates with the primary clusters. The met-like primaries have higher Gleason score, which suggests that the met-like primaries have more aggressive disease

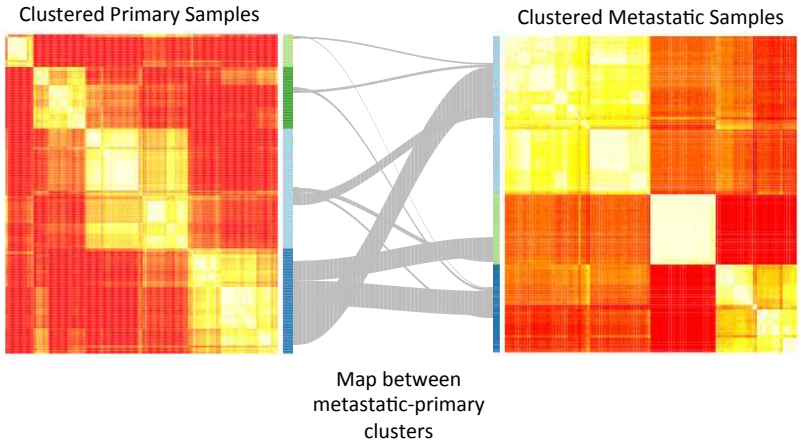

Figure 4.10: Ribbon plot showing distribution of the predicted primary subtype labels in each metastatic cluster, suggesting a stronger association between one of the primary subtypes and the majority of the metastatic samples. Enrichment analysis of the clinical phenotypes in the primary clusters suggests that this subtype is more aggressive.

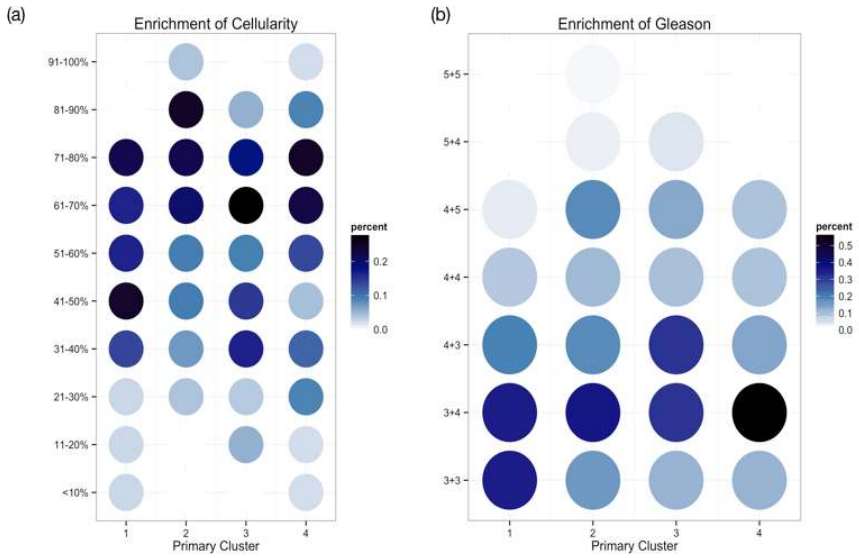

Figure 4.11: Clinical enrichment in the primary clusters of (a) celluarity (sample tumor purity) and (b) Gleason score (histologic appearance). Gleason scores of above 7 are considered high risk. The met-like primaries (cluster 2) tend to have higher cellularity and Gleason scores.

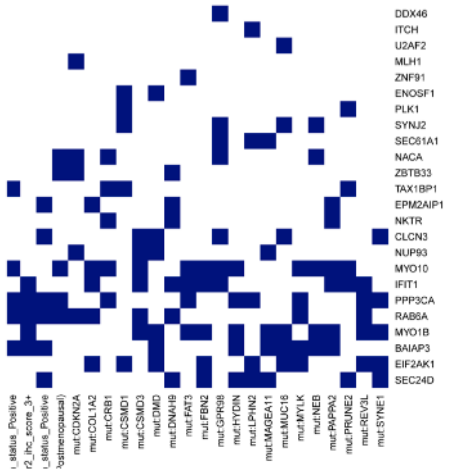

Figure 4.12: The top 20 PanCancer event signatures that overlap the most with the met-like primaries signature.

4.4.3 Results

4.4.3.1 Correlation with Existing Mutation Signatures

To determine if metastasis is driven by specific mutation(s), we compared our aggressive primary prostate signature to mutation signatures from the TCGA Pancan12 dataset [39]. Unfortunately there is no mutation data for several of the prostate studies, so we compare to expression signatures as an alternative. There was little correlation between our signature and the 5,000 PanCan mutation signatures. Fig. 4.12 shows the top 20 mutation signatures (columns) and the genes (rows) that overlap between those signatures and our derived signature. This result support the hypothesis that metastatic activity in prostate cancer is not driven by any specific mutation or set of mutations.

Next, we looked for relationships between our signature and predictors of clinical attributes in cancer. We built 50 linear predictors of several clinical labels, then computed Spearman rank correlation between those signatures and mRNA expression of every sample in the primary prostate cancer cohort. For each set of correlations between a signature and the primary prostate samples we computed, there is enrichment score for cluster 2 samples in the positive correlation tail. Three signatures were significantly enriched: increased monocyte count, CACNA1A_BRD4_NOTCH3 amplification, and IDH1 mutation in leukemia.

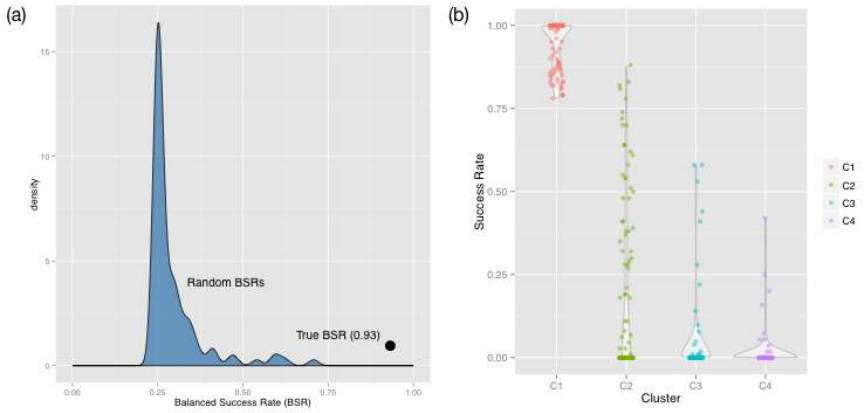

Figure 4.13: (b) Success rates per cluster and (a) balanced success rates for 100 tests of randomly assigned clusters, retaining original cluster sizes.

4.4.3.2 Cluster Stability

Cluster sizes are irregular, and the cluster sizes can contribute to the classifier performance. We randomly assign samples to the clusters 30 times, and calculate BSR for each cluster for each of these random assignments. The scores from these help put the scores from the true model in perspective.

I randomly assigned cluster labels to the samples, retaining original cluster sizes, then used 10-fold cross validation to test the predictability of each 'new' cluster. I repeated this test 100 times and calculated BSR for each. In general, the models trained found no traction and simply assigned samples to the two largest clusters. Fig. 4.13 shows violin plots of the ranges of scores for each cluster, as well as balanced success rates for each of the tests. To compare to the true clusters, I also trained a model on the true cluster assignments using 10-fold cross validation (Fig. 4.13a). The true cluster assignments has a much higher balanced success rate than the randomly assigned

clusters. Furthermore, the randomly assigned clusters tend to perform well in one or two classes at most.

4.4.3.3 Validation Using Matched Primary–Metastatic Samples

We validate our results on a held-out validation set; Erho *et. al* [91] analyzed 545 prostate cancer patients from the Mayo Clinic Registry, from 1987–2001. Median followup for these patients was 17 years, as reported by Erho, and of these patients 212 were identified as early metastasis. Metastatic patients were grouped into no recurrence and recurrence within 5 years of biochemical reccurence. Erho used a random forest to classify patients into metastatic vs. not, and randomly split the data into training and test sets. Of the control patients, 21 had clinical metastasis.

While microarray, patients in this dataset have matched primary and metastatic tumors. This makes it ideal for validation of our metastatic detection signature. We apply our trained predictor to this dataset, to determine if the metastatic samples are also classified as met-like primaries. In order to validate our model using this data, we first transform it to exist in the same space as our other datasets, then apply the trained primary subtype predictor to the primary samples. We then compare the predicted primary subtypes to the actual final sample classes– ones that did not metastasize and those that did (Fig. 4.16). Our approach improves over the original by using several types of data, introducing cross-validation into the model, and by using an independent dataset to validate the results. We also include many more samples, increasing the power of the analysis.

		Metastatic Event	
		No	Yes
Predicted Primary Cluster	1	249	156
	2	31	20
	3	0	4
	4	53	32

Table 4.3: Predicted primary clusters are enriched in samples that metastasized early.

4.4.3.4 Differential Expression Networks

Fig. 4.14 shows several subnetworks of interest from PathMark [251] analysis of the two more aggressive primary clusters versus the others. First, there is a proliferation-related subnetwork that includes master regulators PLK1 and FOXM1. Enrichment of cellularity (aka tumor purity) within the clusters (Fig. 4.11) shows that the met-like primaries have higher cellularity. In some cases higher cellularity has been linked to more aggressive cancers, however it has also been attributed to the ease of acquiring samples in larger tumors. There is also a transcription regulation subnetwork focused on MYB/MYC, which is also independently found by GSEA [316] and overlaps with our predictor signature.

As stated above, GSEA results corroborate with PathMark, indicating more aggressive disease in the met-like primaries. Several cancer pathways are up-regulated in these samples, and several metastasis-linked pathways and genes are dysregulated in both PathMark (several MMP genes are directly associated with metastasis) and GSEA results. Specifically, several MMP genes are identified as differentially regulated in the met-like primaries, and Fig. 4.15 shows GSEA-identified pathways linked to cancer

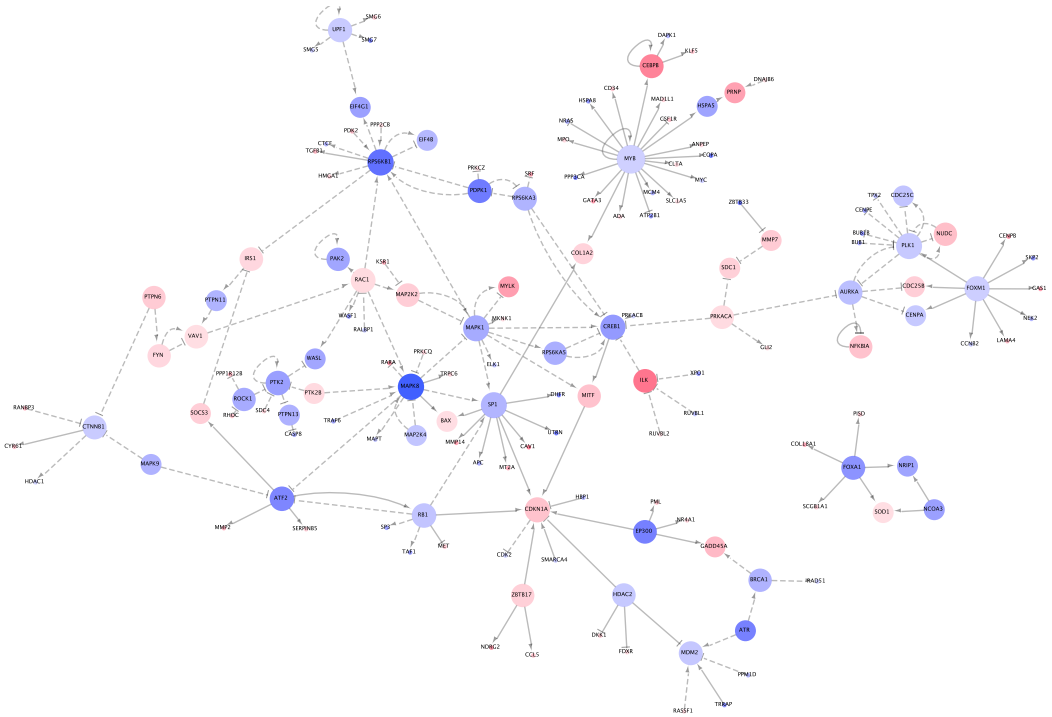

Figure 4.14: PathMark–derived differential subnetworks, based on mRNA expression. Red colors correspond to genes upregulated in the met-like primaries. Node sizes are by edge count; Larger nodes have more edges.

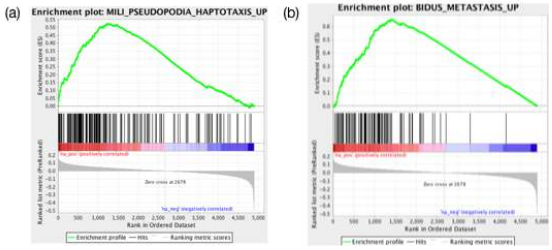

Figure 4.15: Two results from GSEA using the trained primary subtype predictor signatures.

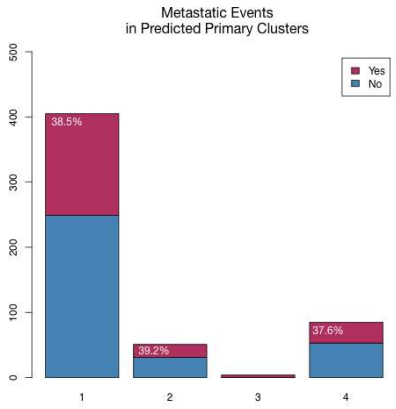

Figure 4.16: Spread of samples that metastasized early, compared to predicted primary clusters.

metastasis.

4.4.4 Conclusions

Identifying which patients have aggressive cancer which will metastasize is vital in treating the disease. Such knowledge enables doctors to aggressively treat the patients early, and will give a better understanding of the survival prognosis of the patient, since metastasis generally leads to patient death. Our analyses have identified a signature for metastatic disease progression. The signature is validated in an external matched-sample dataset. Prognostic markers such as this signature can be used to determine the treatment regimen for new patients, and to help understand how aggressive that patient's disease is expected to be.

4.5 Conclusions

All of the projects in this chapter integrate several genomics data types to find biological drivers of the tumors. Using both integrated analysis and single data type analysis provides distinct perspectives of the disease. For example, the signature derived from PARADIGM MPM subtypes may either replace or supplement histology in the near future. This signature is partially seen in individual platforms, whereas the integrated view captures it much more completely. By combining the data we were able to identify links between patients that was other difficult or impossible to see.

Chapter 5

Multiple View Learning

5.1 Introduction

Cancer is a disease of information in DNA and its 'digital age' has dawned; the plummeting cost of –omics technologies is transforming cancer research from a field limited by data acquisition to one limited by data interpretation. Biomarkers and machine learning classifiers are desperately needed for predicting outcomes, especially those that make use of a battery of different measurement platforms to provide an integrated view of the data.

While the number of genomic datasets have increased dramatically in recent years, there are major complications in using them for inference because each dataset is missing key features. Data platforms and methods often do not overlap between studies, few of which have comprehensive clinical outcomes (*e.g.* survival, drug sensitivity). At the same time many studies have samples that would be useful to analyses

other than their original purpose, yet cannot be included because they lack outcome data. Non-uniformity of large composite datasets, such as The Cancer Genome Atlas (TCGA, *cancergenome.nih.gov*), leads many existing methods to ignore data that it is not available for all samples.

Unfortunately the large number of variables compared to the few samples available leads to many biologically irrelevant solutions [350]. Overfitting issues due to small sample size can be minimized by using prior knowledge-driven feature selection techniques. For example, genes operate in pathways (multi-protein complexes, signaling cascades, transcriptional regulons, shared chromatin domains, etc), and gene modules can be used effectively to summarize activity in individual genes [56]. Concomitantly, several approaches that incorporate database-mined gene-gene interaction information have shown promise for interpreting cancer genomics data and utilizing it to predict outcomes [62, 151, 157, 162, 291]. Additionally, ensembles can be used to reduce error caused by small sample sizes [279].

We present a multiple view learning framework that improves classifier accuracy and interpretation by using multiple biological priors. Each prior generates a specific view of the data and an ensemble of the views provides a more complete and diverse understanding of the underlying biology. The framework 'learns' outcome labels for unlabeled samples, thus including more of the data in the classifiers. It minimizes overfitting caused by small sample sizes both by learning unlabeled samples and by incorporating prior knowledge. This also serves to improve interpretability of the classifier results.

5.2 System and methods

Many current machine learning classifiers train on a subset of samples containing all data [255], impute missing data, or train ensembles based on data availability [86, 392], but are generally restricted to samples with the majority of the data for each sample [192]. Multiview learning (MVL) trains a classifier on single or multiple data platforms (hereafter referred to as a 'view'), then constrains on the unlabeled samples. MVL is a semi-supervised approach that learns missing patient outcome labels, allowing use of all available labeled and unlabeled datasets. By doing this MVL can make predictions on any patient regardless of data availability. This increases overall classifier accuracy while also finding solutions that generalize to the entire population– which has proven extremely difficult in such high-feature, low-sample problems [2, 138, 350].

Multiview learning has been used in bioinformatics problems; in prostate chemical recurrence prediction [114], and in a recent DREAM competition [69] where several competitors use forms of multiview learning to predict drug sensitivity. However, using multiple kernels is not a true multiple-view approach when also using a biological prior as many of the competitors did. While previous work has used multiple kernel learning to combine different biological data types [275, 307], using a biological prior to break data into kernels by, for example, pathways, is more similar to a single view classifier. Another approach in the challenge uses a form of multiview learning, one that cannot use samples with missing data, and furthermore it works by using Canonical Correlation Analysis (CCA). While using CCA is appropriate in situations where views are

highly correlated [53, 371], in biological problems the data is incredibly noisy and different biological perspectives of the data (e.g. immunologic response vs tissue–specific gene interactions) are dissimilar, leading to low-correlated views. Furthermore, both multiple kernel and CCA-based multiview methods do not gracefully handle missing data [276, 376]. The MVL approach in this paper instead uses co-training, which both uses samples with missing data and benefits from views without high correlation. This method is a more flexible form of MVL that is also more tailored to biological problems.

This approach has several important advantages. First, it allows for the use of different classification methods for each data type, capturing the strengths of each data type and increasing flexibility in the framework. Second and third, it is ideal in scenarios with missing data and views with divergent information content. Finally, co-training combines predictions at a later stage in the algorithm, such that the views are trained independently. This is a better scenario for ensemble learning, which as a rich literature has shown, thrives when views are independent even if the accuracy of the views is low [274, 279].

Co-training works by training separate classifiers for each view, making individual predictions, and incorporating disagreement into the loss function. Each view trains on the labeled data, then predicts labels for the common unlabeled set. High confidence labels are passed as truth in the next iteration. Co-training methods iterate until either convergence or some threshold is reached– a minimal change in label definition on the unlabeled samples, or a max number of iterations (for scenarios where the views will never agree on these samples).

After co-training, each view can be used as a standalone classifier that incorporates learning from on one or more data platforms (or feature transformations) without relying solely on that data platform. Since view are trained in conjunction, the trained models will incorporate the perspectives of all views. Thus we have some measure of influence from all views when applying any of the classifiers to new data, without requiring data for those views when making new predictions.

In some cases, a new sample may be missing all or most of the data types. MVL is able to predict outcomes for this sample without retraining, by including only those views for which there is sample data. For example, a sample with only expression data would be predicted using only the expression–based views. Label confidence will be much lower since there are no scores from the missing views, thus labels for samples with missing data will be inferred in later iterations than those with complete data. Since the MVL model was trained on the full set of data types there is still some influence from the missing data. MVL model predictions are constructed such that the user does not need to specify a new weighting for the view predictions. Thus, future predictions are not constrained by available data.

5.2.1 Interpreted Views

Biological experiments of the past century have helped build new perspectives of the underlying rules of biological processes. A plethora of databases have arisen that focus on divergent perspectives. For example, regulation networks for all different kinds of cell states provide information about the ways proteins influence each others

activity. Comprehensive gene annotations make it possible to find connections between distinct genes, *e.g.* shared motifs in non-coding gene sections or genes sharing the same cell location. In a big science community effort, a competition for breast cancer drug sensitivity predictions was recently conducted where the winning method used multiview learning by integrating many biological priors [69]. In addition, multiple recent studies show an overall correlation between the use of outside information and method performance [116, 162, 385].

We create 'interpreted' views (described in Section 5.3.4), each of which incorporates one data type and one biological prior. Several priors are included with MVL, as well as a feature transformation function that takes as input a data set and a prior, and outputs features for that new view. Datasets such as the Cancer Cell Line Encyclopedia (CCLE) [17] have information from multiple data types; copy number, mutation, expression, and clinical/phenotypic covariates. Views can be built from each datatype, although oftentimes there is a discrepancy in datatype performance. Combining the expression data with different biological priors boosts the predictive power and creates many semi-independent views, each with high prediction accuracy. Together, baseline and interpreted views can be used to train an MVL model, one which pulls in multiple prior knowledge databases. This minimizes reliance on user knowledge (poor fit priors have minor effects on the final model due to low model accuracy), while allowing many different perspectives to lend interpretatability.

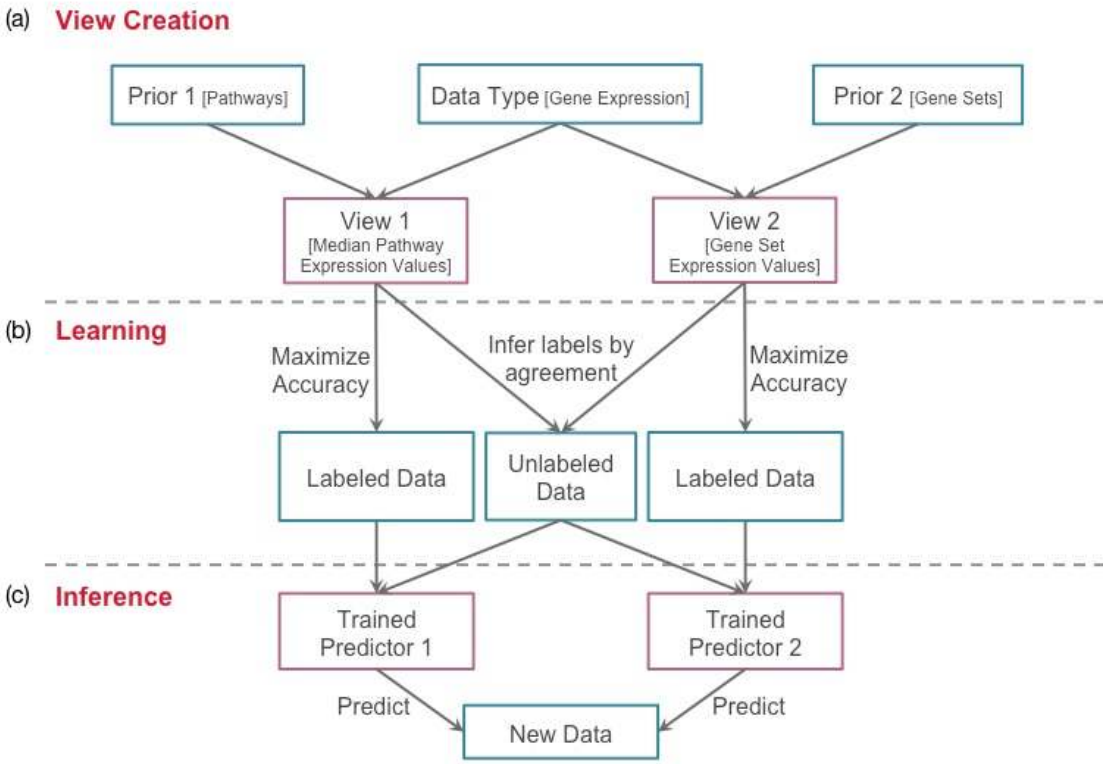

Figure 5.1: Introductory methods figure describing MVL. The MVL framework. This figure shows two views being used in MVL. (a) Creation of the single views using sample data and prior knowledge. (b) The learning process, where each view maximizes prediction accuracy of the labeled samples, and unlabeled samples with high confidence are added to the known sample set. This phase is an iterative process that continues until no new sample labels are learned. (c) Models from the final iteration of MVL training can be applied to new data either independently or using the MVL framework.

5.3 Algorithm

The semi-supervised nature of this MVL framework works by inferring outcome labels from unlabeled data at each training iteration in order to improve the learning process. Fig. 5.1 shows an overview of the MVL framework, where data 'views' are first created from one or multiple data types, which can be combined with one or more priors (Fig. 5.1(a)). While this example uses two views, the number of views used by the

MVL framework is unlimited; we use up to 10 views at a time in our tests. Section 5.3.4 describes the views created in this project.

The second phase in MVL (Fig. 5.1(b)) simultaneously trains the views on labeled data and jointly infers labels for unlabeled samples. Each view is trained to maximize prediction accuracy on the labeled samples. Single views are then combined in an ensemble to predict labels for the unlabeled samples; Labels with high confidence predictions are then added to the existing known labels set. The training process continues until a convergence criterion is met: all labels have been learned, no new labels have been learned in the last iteration, or a user–specified maximum iterations has been reached.

After termination of the learning progress, the trained predictors can be used independently or as an ensemble (Fig. 5.1(c)). In order to validate the learning process, we develop a procedure called *label-learning validation* (LLV, Section 5.3.2, Fig. 5.2(c-d)) which masks out labeled samples, then re-learns the masked labels. This process is similar to cross-validation.

5.3.1 Implementation

There are 2 main inputs to the MVL algorithm: (1) binary outcome labels ('sensitive' or 'non-sensitive') of the labeled samples and (2) the data view objects. View objects contain the feature matrix, learning algorithm type, optimized parameters for the algorithm (optional), and MVL weight for that view (optional).

$$w(a) = -log(1 - \frac{a - 0.5}{0.5}) \tag{5.1}$$

In each iteration the views are first trained on the labeled data, then used to predict the unlabeled data. The vote from each view is weighted either by a user-provided value or by AUC. Weights can be static or updated at each iteration, as specified by the user. Accuracy is rescaled from $[0.5, 1]$ to $[0, 1]$ and log-scaled (Eq. 5.1, where a =accuracy and w =weight). Views with an accuracy lower than 0.5 are given a weight of 0, since it indicates worse than random predictions. Predicted label confidence for each sample is then derived from the weighted votes sum.

$$\lambda \sum_{v \in V} w(a_v) \tag{5.2}$$

At every iteration, MVL updates the confidence threshold defining whether or not a predicted label is added to the known label set. Votes for each sample are summed up separately for the two possible class labels. The maximum sum over all samples ($max.vote$) defines the threshold a sample has to meet in order to be included in the training data, thus assuring that only the most confident predictions are chosen. In order to favor missing data over a contrary prediction, we define a second threshold based on the minimal contrary prediction vote among all samples that meet the $max.vote$ requirement ($min.max.vote$). Samples meeting both requirements are added to the labeled data. The training iteration procedure stops when either the $max.vote$ value drops below a user-defined threshold ($\lambda = majority.threshold$) or no unlabeled samples

remain. By default λ is 75% of the maximal reachable voting value (Eq. 5.2).

5.3.2 Label-learning Validation

To validate the label learning process we create a function for k-fold validation of the labeled data. Similar to cross-fold validation approaches, label-learning validation (LLV) masks a subset of the labels then trains the model using the remaining labeled samples. Masked samples are treated as unlabeled data . MVL then learns labels for the masked samples. LLV compares the learned labels to the (masked) true labels. Views are trained using the unlabeled data and the $(k-1)$ folds of labeled samples. After performing all k folds, labeled samples have an additional learned label (except in cases where it could not be learned either because of strong disagreement between the views or extensively missing data).

Fig. 5.2(c-d) shows a label-learning visualization of PD-0325901. In this visualization, MVL has correctly relearned the majority of labels. This example may have benefited from stopping at at earlier iteration of MVL, since the majority of incorrectly learned labels are from later iterations. A user parameter sets the maximum number of MVL iterations and LLV is a useful tool for choosing an appropriate cutoff. Fig. 5.2(d) also shows that labels are not learned linearly. LLV highlights MVL confidence in the predicted labels, and also that MVL does not force a label on samples. Fig. A.19 shows the LLV visualization for each of the 24 CCLE drugs. For each drug, MVL successfully learns the majority of sample labels correctly. The learning processes differs between drugs and so we postulate that there is no specific number of iterations that MVL should

run. By default it runs for 100 iterations. However, by using LLV a user can see how MVL performs on the labeled data and can then extrapolate its performance on the full (labeled and unlabeled) data.

5.3.3 Baseline Views

'Baseline' views do not use prior knowledge. The four baseline views in this paper are derived from platform-specific data; Copy number, mutation, expression, and clinical/phenotypic. The mutations view contains a binary mutation feature vector for each gene, where '1' represents presence a non-silent mutation in that gene within in a given cell line. Data in the expression view is reduced to the 5,000 genes with the highest variance over all cell lines, in essence the 5,000 most frequently mutated genes, since mutation data is so sparse. Similarly the CNV data view includes the 5,000 most variable genes by copy number. Note that in the example, PD-0235901, in Fig. 5.2(b), 3 of the baseline views are top performers and used in the MVL models.

The 'clinical' view is a collection of binary features based on tissue, histologic type, and gender. Additionally, it includes two categorical features representing the genomic instability of the cancer: The number of mutated genes, and the sum over all absolute CNV values in each cell line. These are grouped into 'low', 'medium', and 'high' instability.

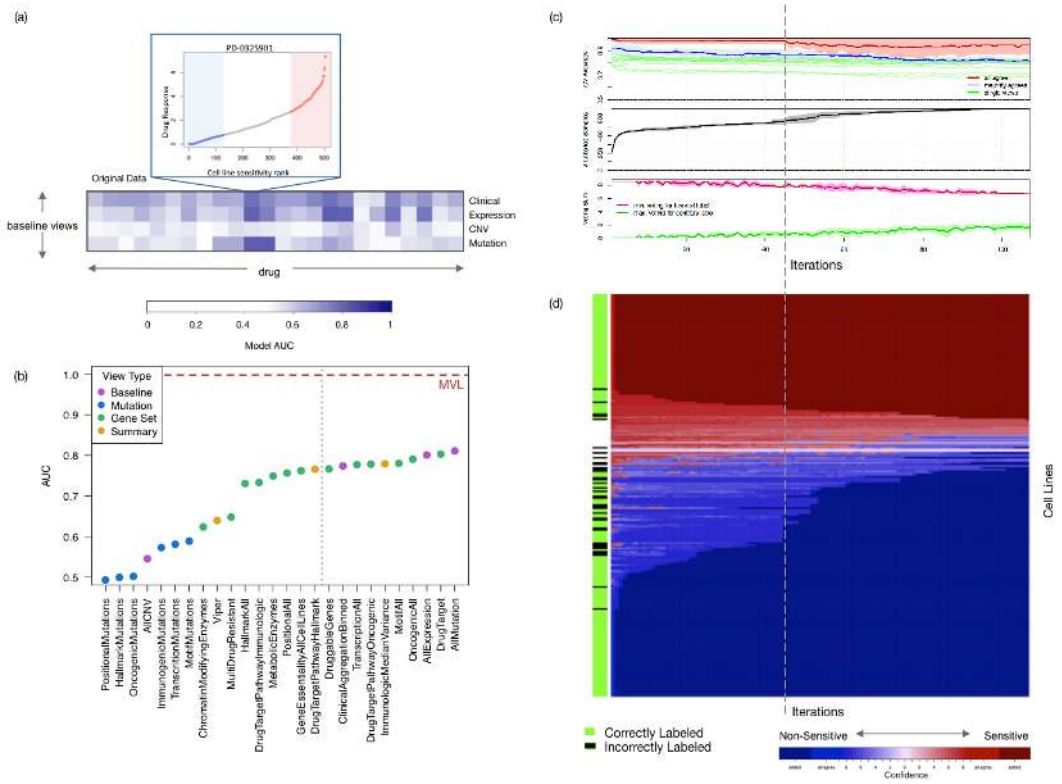

Figure 5.2: An example MVL run on the drug PD-0325901. (a) Baseline view AUC for each drug in CCLE, with the PD-0325901 sensitivity rank plot showing the binary labels, (b) AUC for PD-0325901 sensitivity predictions for each view, colored by view type (top 10 views are used in (c-d)), (c) Label-learning validation plots, top to bottom: the accuracy in the labeled set at each iteration for two types of MVL models and for each view, the number of samples for which labels have been learned, and the amount of disagreement in the label scores. (d) visualization of LLV showing confidence of predictions for each cell line at every iteration. (c-d) We added a dashed vertical line to show a likely user-defined stopping point for the method, where the overall disagreement in predictions between the views has started to increase and before the model AUC starts to significantly decrease.

5.3.4 Interpreted Views

Genes operate in pathways (multi-protein complexes, signaling cascades, transcriptional regulons, shared chromatin domains, etc), and gene modules can be used to summarize activity in groups of genes [56]. Concomitantly, several approaches that incorporate database-mined gene-gene interaction information have shown promise for interpreting cancer genomics data and utilizing it to predict outcomes [62, 151, 291]. We use these to add prior knowledge-based views, called 'interpreted' views.

Using multiple views both helps improve accuracy and, if using a relevant biological prior, can help with downstream analysis and interpretation. Views can be data platform specific (ex. expression), integrated data platforms (ex. both expression and copy number), and can use a biological prior to transform the data platform features into features based on that prior. View signatures can be combined to help identify new potential drug targets. Furthermore, each view brings in a distinct biological perspective and, when combined with expert knowledge, can help guide treatment decisions. MVL has the flexibility to add and remove biological priors easily, and builds on current state-of-the-art methods by integrating more than one prior and machine learning method.

Prior knowledge databases mitigate overfitting issues and to help with downstream analyses. Furthermore, incorporating several biological priors adds expert knowledge to the learning process. We generate feature sets based on expression data that has been transformed using one of several prominent biological priors.

5.3.4.1 Biologically-Driven Gene Network Views

Pathway–based predictors are often easier to interpret and more descriptive of the processes being disrupted in cancer. We include several biologically-driven views to capture these underlying structures. Users can create new views using the MVL function to calculate gene set values, which takes as input the data, prior knowledge information, and summary statistic. We test mean, median, and kurtosis of gene sets over many different prior knowledge databases. Views are composed of each of these values for each gene set.

Spatial form of the genome and chromatin structure are closely tied to biological function and, as such we create a view based on physical structural proximity. A recent study found that chromatin interaction domains are both highly stable and have few boundaries that differ between cell types [81]. It also deconvolves tissue-specific noise, which have been strongly correlated with expression [148, 297, 308].

We use the Molecular Signatures Database (MSigDB) [214] to create a pathway-specific view to help elucidate pathways dysregulated in cancers. Inclusion of the Drug-Gene Interaction database [126] clarifies the effects of mutations on the response of cell lines to chemical agents. Both of these databases provide insights into the function of drug sensitivity, and knowing which genes a drug interacts with helps focus attention to that subset of genes and their interactors. The use of pathways as gene sets has been shown [330] to be effective at increasing interpretability of the systemic changes in the system by the cancer, thus motivating their use as a view.

Additionally, master regulator analysis provides a view that summarizes aggregate expression of a transcription factor's downstream targets, which has been shown to identify key transcriptional regulators driving the cancer phenotype [207].

5.3.4.2 Biological Gene Sets

Biological gene sets can be used as priors, where a subset of the data (*e.g.* a list of chromatin–modifying genes) is used to create a new MVL view. In this paper, the gene set views are built with expression data. Gene sets used in this paper are described in detail in Section A.2.2. We also created these views with CNV data, however they are excluded due to poor predictive power (average AUC 0.53).

5.4 Data

The Cancer Cell Line Encyclopedia (CCLE, *www.broadinstitute.org/ccle/home*) contains genomic, pharmacological, clinical, and other annotation data for about 1,000 cancer cell lines [17] (Section A.2.3). At time of download there was drug sensitivity data for approximately 500 cell lines and 24 drugs. Drug response was converted to a binary label, in order to transform the regression problem into a classification problem. For each compound, cell lines were divided in quartiles ranked by ActArea; The bottom 25% are assigned to the 'non-sensitive' class and the top 25% to the 'sensitive' class. Cell lines lying in the middle are marked with 'intermediate' and considered unlabeled in this analysis (see Fig. 5.2(a) and Fig. A.15).

5.5 Discussion

5.5.1 Experiments

We ran MVL using 3, 5, 7, and 10 views for each of the 24 CCLE drugs. Fig. 5.3 shows the highest accuracy MVL run as well as each of the single view scores. In almost all cases MVL significantly outperforms single view models, most notably the MEK inhibitors AZD6244 and PD-0325901, and HDAC inhibitor Panobinostat. Furthermore, within 10 iterations most MVL runs added 90% or more of the unlabeled cell lines to the labeled set, effectively doubling the number of samples on which the models trained. We look more closely at the results from the best overall performing MVL model, PD-0325901, in Section 5.5.2, as well as important features from each of the models.

Fig. 5.2(d) and Fig. A.19 show LLV for each of the 24 drugs in CCLE. While most of the drug models learn labels correctly, cases with noisy labels (*e.g.* no/few sensitive cell lines) tend to learn new labels incorrectly. Over many iterations this can lead to a model where the majority of labels are learned incorrectly, such as Nutlin-3 (see Fig. A.19).

5.5.2 Results

PD-0325901 has the highest MVL scores of all experiments in this paper, with near–perfect accuracy.. It was initially tested in papillary thyroid carcinoma cell lines [144] and is known to be especially effective in cell lines with BRAF mutations.

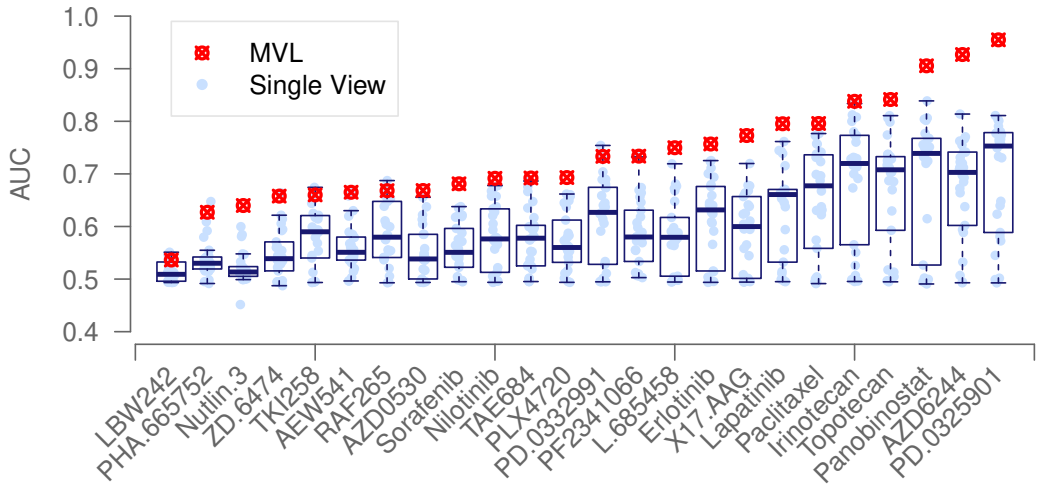

Figure 5.3: MVL results. Boxplot showing performance (in AUC) sorted by MVL score, of all single views and the best MVL score. MVL score for each drug is the highest from the 3,5,7, and 10 view MVL runs.

Since these are frequent in the CCLE data, the high accuracy of the single view models is expected. However, all MVL tests in this drug have a significantly higher AUC than single view models. A more in depth look at the features forming these models (Fig. 5.2 and Fig. 5.5) shows that there is more than simply a mutation plus drug combination that affects sensitivity.

Part of the success of the MVL framework is in the ensemble learning. It is well established that combining multiple weak but independent models will result in much higher model accuracy [192, 279]. The way in which models are combined is a key part of an ensemble; We tested several approaches. Initially we used a predetermined weighting where each view contributed to the final prediction, however this made the

model sensitive to information-poor views. For the analyses in this paper we used AUC-weighted voting. This allows the user to include a large number of near random views, in the hope that they will become more informative as more cell lines are added to the training set, and so that the user does not have to perform a pre-processing step of selecting views. Furthermore the results from these models can help with downstream interpretation.

Similarly, previous work has shown that using multiple biological priors can help minimize the ever-present noise in biological data [62, 151, 157, 291, 298, 347]. By using an ensemble of different biological priors we both improve the model accuracy and help minimize downstream analysis required. Since the user is already provided with pathways, functions, and genes associated with the predicted outcome, there is no need to re-identify them.

In coming years, CRISPR knockout screens have helped to more fully map individual interactions between genes. Cell lines with induced mutations in possible drug targets will open up a new world of targetable drugs without requiring living patients with those mutations. This will generate a huge data resource that is patient agnostic but that would dramatically change the way that we determine patient treatment.

One drug, LBW242, had almost random scores in all models (including MVL, Fig. 5.3). However this is unusual; In other drugs with near random single view model scores, for example PHA-665752 and Nutlin-3, the MVL models have much higher accuracy. LBW242 scores are likely low because there are few CCLE cell lines that are sensitive to that drug, meaning that the binary class labels are not reflective of the data.

It's also possible that the metric for drug sensitivity is ineffective. Traditional methods to quantify sensitivity are dependent on population growth and thus slow-growing cell lines may appear to be resistant to all drugs [135]. In future work we would like to use multiple measures to calculate drug sensitivity in each of the drugs.

These results are consistent with previous findings; sensitivity to some compounds is easier to predict than others [69]. Fig. A.16 shows an overview of the prediction accuracy of single data views. The two MEK inhibitors (PD-0325901, AZD6244) and Panobinostat have higher overall accuracy in the single view models. In Panobinostat, the 'Chromatin Modifiers' and 'Positional Gene Set' views have higher single view accuracy than the baseline expression view, which suggests that there is an epigenetic effect from chromatin modifiers. We postulate that a small region of the genome has been unwound, lending sensitivity to Panobinostat.

5.5.3 Tissue-Specific MVL

By restricting to one tissue we can identify tissue–specific drug sensitivity factors. While this reduces the number of samples used in analysis, it can help uncover signals that are only present in the given tissue. Fig. A.17 shows results from models trained on single tissue types. Unfortunately most tissues have too few samples to run robust analysis, and so we include lung and blood specific analysis. Since the number of cell lines influences model AUC, we label the number of samples included in each analysis as well as predictor and view types. In most cases there are fewer than 50 labeled cell lines in the tissue-specific models (Fig. 5.4).

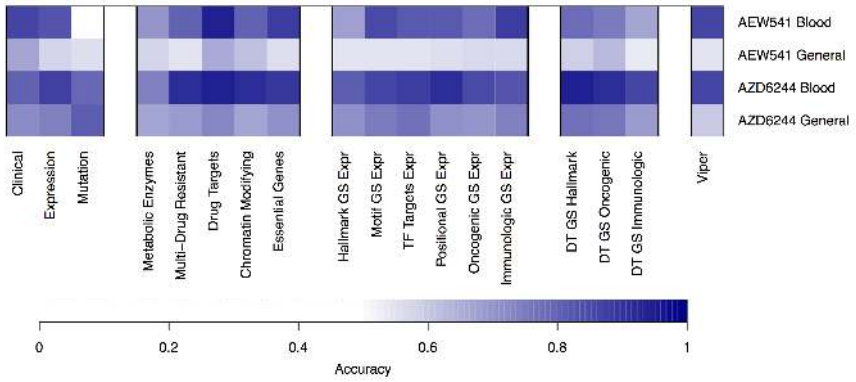

Figure 5.4: Cross-validated AUC of single views with their optimized parameter settings. This compares the tissue–specific setting using blood cancer cell lines to the complete CCLE on AEW541 and AZD6244.

Blood MVL (180 cell lines) had less success than experiments on the full CCLE data. Most of the drug models converged within 10 iterations. Despite this, the single view accuracies before LLV are higher than the full data (Fig. 5.4 and Fig. A.20). This is in part due to the number of samples (see Fig. A.17). AZD6244 (MEK inhibitor) and AEW541 (Tyrosine Kinase inhibitor) especially perform well in the blood data.

5.5.4 Pan-Tissue MVL

The clinical view has the highest performance of all the baseline views in most drugs, with AUC ranging from 0.6 to 0.8. In some drugs (*e.g.* MEK inhibitors) the mutation view is effective. None of the CNV views have high AUC and thus are excluded from the rest of analysis, except as the 'aggregated copy number changes' feature in the Clinical view.

Interpreted views, however, often outperform the clinical view. There are many examples of a biological prior view outperforming the data-specific view, *e.g.* Metabolic

Enzymes, Drug Targets, and Chromatin Modifying Enzymes are more useful for Lapatinib sensitivity prediction than the Expression predictor. The Drug Target Gene Set Hallmark predictor outperforms data-specific views in Irinotecan and Panobinostat sensitivity predictions. Such examples can be found for all compounds except for the MEK inhibitors, for which the mutation predictor is always the top performer.

In general, views incorporating expression data seem to help in many cases (Fig. A.16), whereas the views using mutation data are comparable to a random prediction. These views were built by measuring mutation enrichment in the different collections of gene sets and do not capture drug sensitivity information. The Drug Target Mutation predictor is more accurate compared to the simple Annotated Target Mutation predictor, suggesting that it overcomes limitations identified in Section A.2.1.

In all drugs except AZD6244, the Drug Target Expression view is more accurate than the Drug Target Mutation view. This might be due to the expression of a gene being closer to biological effects in the cell than a mutation, which can only translate into an effect if the resulting altered protein is expressed or the expression rate is influenced by the mutation. Overall, mutation-based views have low accuracy despite mutations being key to drug sensitivity, indicating that other representations of this data should be explored in future work, perhaps using a diffusion kernel in a support vector machine model.

The Drug Target Gene Set views created from Molecular Signatures Database (MSigDB) gene set collections (Fig. A.16) perform well overall, and especially on Irinotecan, Topotecan, and Panobinostat. For most compounds the Drug Target Gene Set

Hallmark is more accurate compared to Oncogenic and Immunologic. A possible reason is that these gene sets are from the Hallmark collection; They are re-occurring, high reliability gene sets built from combinations of other gene collections. Therefore, the similar performance could be due to overlap in the gene sets [184]. A future improvement to MVL would be to test for, and subsequently remove, highly correlated views before running label-learning. In addition to the MSigDB gene set views, master regulator-based predictors via Virtual Inference of Protein-activity by Enriched Regulon analysis (Viper) [4, 5] were tested but are not among the top performing ones for any compound.

VIPER performance is much higher in the tissue–specific MVL runs than in the full CCLE MVL runs (Figs. 5.4, A.20). This could be caused by our use of a generic regulon as VIPER input rather than a tissue–specific one.

5.5.5 Key Features from MVL Models

Each machine learning algorithm used by an MVL view has its own internal feature selection. We extracted features from each model to find the most informative features. Fig. 5.5 shows highly ranked features from the top 10 views for each drug, after MVL training. We normalized feature weights across models and select the top 10 features per view. Many of the highly ranked features are known oncogenes, for example ETV4 is known to be fused with TMPRSS2 in prostate cancer gene fusions [210, 333, 359]. However there are less than 10 prostate cell lines in CCLE. ETV4 was also previously found to be correlated with MEK inhibitor sensitivity [210].

All of the genes labeled in Fig. 5.5 are all related to at least one type of cancer. CDT1, DUSP6, SPRY2, and ETV4, are associated with many types of cancers and are involved in at least one key cancer-related pathway [106]. DUSP6 negatively regulates MAPK family genes, which are associated with cell proliferation [169, 206]. Deactivating this gene allows MAPK genes to be over-expressed and MAPK pathway over-expression has been associated with aggressive disease [37, 39, 40, 84]. SPRY2, another kinase inhibitor, is associated with many forms of cancer including prostate. It is also associated with signaling pathways, several microRNAs associated with cancer, and glioblastoma signaling pathways [33, 42]. Lactate Dehydrogenase B (LDHB) has the largest negative correlation to AZD6244 sensitivity. It a known oncogene for different cancer types [208, 395] especially in combination with KRAS amplification [235]. The Raf-MEK-ERK pathway targeted by AZD6244 is also suspected to be involved in DNA damage response [370].

CDT1 has large weights in several single view models after MVL training, and is associated with the RB pathway (implicated in aggressive cancers), as well as with genomic instability [263, 411]. DUSP6, ETV4, and RNF125 are associated with metastasis [396]. RNF125 inhibits several receptor tyrosine kinase signaling proteins [141, 180]. ARHGAP19 is associated with cell migration, differentiation, and proliferation [226]. Finally, CDCA7 is a c-Myc responsive gene that behaves as a direct c-Myc target gene and is linked to breast cancer [240]. Overexpression of CDCA7 is involved in lymphoblastoid cell transformation [129].

5.6 Conclusions

Features extracted from MVL-trained models are frequently known oncogenes. Together with the interpreted view features (*e.g.* RB pathway activation), they reiterate known cancer models. By combining extracted features from each of the MVL model views, the user is provided a clearer picture of the key facets of sensitivity to each drug. Furthermore, in almost every instance the MVL models outperform the single views. In all cases MVL AUC is higher than the majority of single views, often by more than one standard deviation of the expected scores. Thus it makes sense to use MVL for training, since it never hurts performance and is likely to significantly increase overall performance. MVL also returns several sets of features independently, which can be used in a variety of downstream analysis.

Several cell lines are never learned by the MVL label-learning process. These may be of more interest in future studies since they are likely very different even in the interpreted view feature spaces. Perhaps they are an unusual form of cancer with distinctly different mutations. Outlier detection methods could help identify the unique aspects of these cell lines.

Label-learning validation (Fig. 5.2(c-d)) is effective in many of the drugs. A few cases, such as Nutlin-3 (Fig. A.19), have little success with learning new labels, which as previously mentioned are likely due to the lack of CCLE cell lines sensitive to the drug in question. Overall, LLV shows the effectiveness of bringing in new labels, improves model stability by nearly doubling the number of cell lines included in analysis,

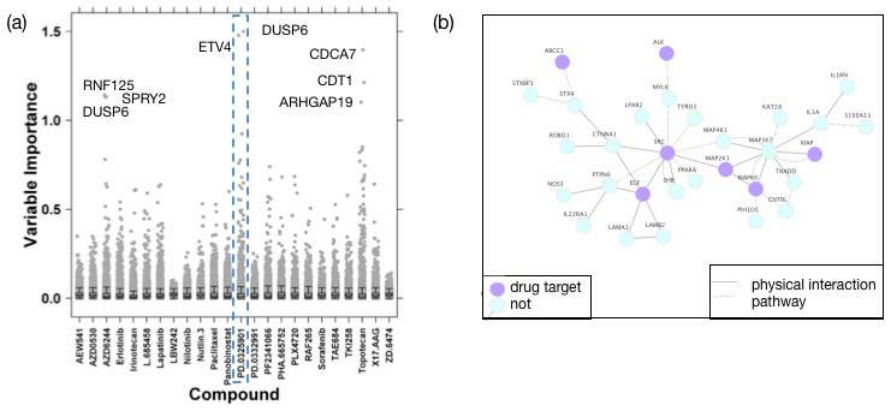

Figure 5.5: (a) Top 10 features for each view in each drug, with weights rescaled to be [0,1.5] and (b) GeneMania [368] plot showing an interaction network for the PD-0325901 MVL features, with the known drug targets highlighted in purple.

and helps to identify outlier cell lines.

MVL includes both unlabeled samples and missing data. When compared to a traditional ensemble and to single view predictors, MVL has often higher prediction accuracy (Fig. 5.3). Furthermore, the biological perspectives innately provided in the interpreted views make analysis of the models easier. Label-learning validation shows that, in most cases, labels are learned correctly and improve model performance. This also helps identify outlier samples in the unlabeled data, since the MVL model will return a vector of 'unlearnable' samples which can then be more closely analyzed. The MVL models have higher accuracy than the majority of single views, and are able to incorporate significantly more cell lines. Combined with the ability to highlight outlier cell lines (not learned during label-learning), and the greater interpretability provided by the use of many biological priors, makes MVL an effective choice.

5.7 Identifying Patients with Rare Histology in a Combined Treatment–Resistant Prostate Cancer Clinical Trial

We participate in a prostate cancer project through Stand Up to Cancer (SU2C, *www.standup2cancer.org*). Reduction of testosterone can slow tumor progession in some patients diagnosed with prostate cancer, however in some cases patients do not respond as expected. These are called treatment–resistant prostate cancer (TRPC). To investigate this more, we collaborated with several West coast labs to treat and analyze a cohort of TRPC patients. New treatment approaches are developed by scientists such as myself, then presented to the medical team who use the new information to guide patient treatment. In order to join this study, patients must no longer be responding to current treatment options.

While the prostate cancer dream teams are split by the two coasts (the East Coast Dream Team (ECDT) and West Coast Dream Team (WCDT)) each with a different cohort of patients and researchers, there is collaboration between the groups. I was asked to perform joint analysis on the two datasets. The goal was to identify a rare subtype of patients within the ECDT samples, which do not yet have histology calls. WCDT has been working to subtype their patients and has identified a new rare subtype.

There are 235 samples in the combined dataset. We have training data for 46 samples; 10 small cell and 36 adenocarcinomas. There are also labels for some mixed

histology samples, which we use in later validation.

I ran ComBat [167] on the combined WCDT&ECDT data, and used the combined data for downstream analysis. With this, I trained the 9 single views described in Table 5.1. For each view I trained the model 100 times using the same folds for crossfold validation. The three views with higher than 0.80 AUC on average were used to train the MVL model: 'Hallmark Genes', 'Expr5k', and 'Chromatin–Modifying Enzymes'.

MVL trained for 8 iterations with the user–specified parameter for label confidence set to 70%. Initial training set included 10 small cell samples and 36 adeno samples. After training, MVL had learned 10 additional small cell samples and 178 adeno, for a total of 214 adeno samples and 20 small cell. Only 1 sample, TP2061 from the ECDT data, remained unlabeled. Furthermore, after label learning all of the training samples are correctly predicted, whereas beforehand there is 1 mislabeled sample (Figure 5.1).

Of note, I included samples with mixed histology in the unlabeled sample set (Figure 5.7). Two of these MVL predicts to be small cell; DTB-120 is Adeno&IAC, and DTB-040 is IAC&Small Cell. We do not know the histology calls for the other eight samples but hope to validate them in the near future.

An MVL test using all views was unable to identify more small cell samples, likely because of the small training set size and high label–confidence threshold. A lower threshold test recapitulated some of the small cell predictions in the 3 view test. However, these tests all converged in a few iterations.

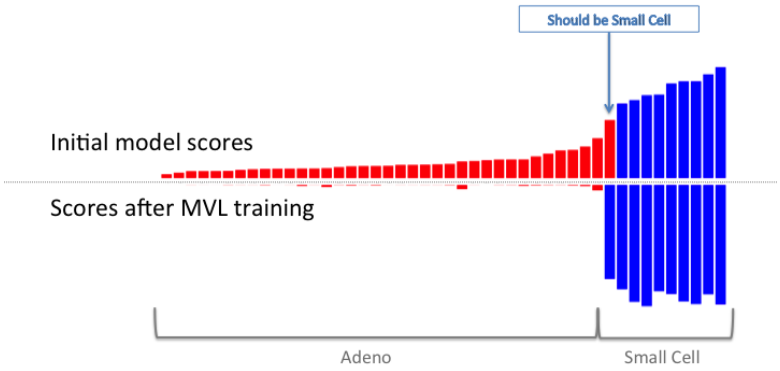

Figure 5.6: Scores for the training set labels before and after MVL label learning.

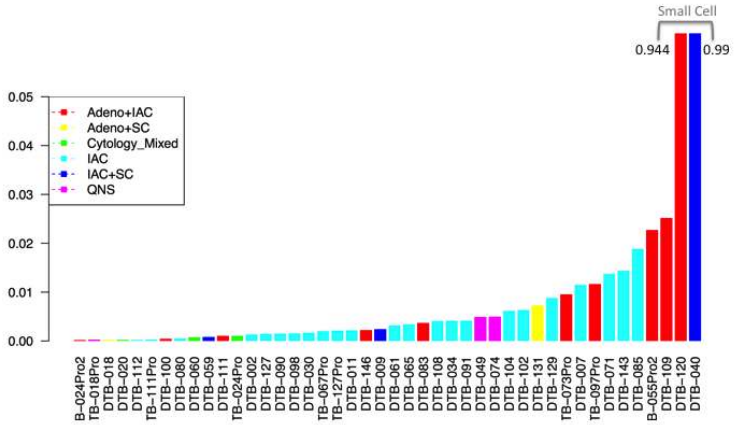

Figure 5.7: MVL was not given labels for mixed histology samples. Several are small cell mixed samples, and are predicted small cell by MVL.

Table 5.1: Single views considered for the combined run of ECDT&WCDT data. AUC is average calculated from 100 tests each with a unique sets of folds. The same fold sets were used on all views. Views with greater than 0.8 AUC are included in the MVL experiments. Type labeled as 's' for summary and 'gs' for geneset.

View Name	Type	# Features	Origin	AUC
Hallmark Genes	s	50	MSigDB	0.87
Expr5k	gs	5,000	Varying Genes	0.84
Chrom. Mod. Enzymes	gs	65	Allis et al 2007	0.81
Oncogenic Genes	s	189	MSigDB	0.79
Positional Gene Sets	s	343	MSigDB	0.77
Druggable Genes	gs	4,963	DrugBank,DGIdb,TTD	0.75
Trans. Factor Targets	s	615	MSigDB	0.72
Motif Gene Sets	s	836	MSigDB	0.70
Immunological Gene Sets	s	1,910	MSigDB	0.62

Chapter 6

Conclusions

Cancer is a complex set of diseases, questions about which may appear simple yet be incredibly difficult to answer. And yet, it is one of the most common causes of death in the US [10]. Someday, when we understand much more about the human body, these questions will be easier to answer. Until then computational techniques are absolutely essential. As Carl Zimmer once said, "Early telescopes weren't terribly accurate, either, and yet they still allowed astronomers to discover new planets, galaxies, and even the expansion of the universe."

Just as radio, infrared and ultraviolet telescopes provided new perspectives on the universe that were not available from the classic optical telescopes. So too, new techniques allow us to measure what we do understand about cancer, and I hope that the work in this dissertation has helped us to understand more. I worked to add new types of data to our analyses; Imaging data is often ignored because it is so large and requires yet another type of medical expertise. Once we were able to

use this data well, I combined it with more traditional data types to analyze tumors through a integrated–data perspective. Finally, I combined together all of these data and develop a semi–supervised prediction framework which uses prior knowledge when making decisions. By doing this, I help make it easier to repurpose genomic data. Each of these methods has both found new results and reiterated findings from the literature. For example, HOCUS subtypes are biologically- and treatment-relevant, and are undetected by equivalent approaches using standard clustering.

Such findings can inspire new biological questions. While each publication may seem to provide only a tiny change in our understanding of the world, together we have created incredible change.

Bibliography

[1] Charu Aggarwal. Social Network Data Analytics. In Charu Aggarwal, editor, *Social Network Data Analytics*, pages 1–15. Springer, 2011 edition, 2011.

[2] Antti Airola and T Pahikkala. A comparison of AUC estimators in small-sample studies. *... workshop on Machine ...*, pages 3–13, 2009.

[3] C David Allis, Shelley L Berger, Jacques Cote, Sharon Dent, Thomas Jenuwien, Tony Kouzarides, Lorraine Pillus, Danny Reinberg, Yang Shi, Ramin Shiekhattar, et al. New nomenclature for chromatin-modifying enzymes. *Cell*, 131(4):633–636, 2007.

[4] Mariano J Alvarez, Federico Giorgi, and Andrea Califano. Using viper, a package for virtual inference of protein-activity by enriched regulon analysis, 2014.

[5] Mariano J Alvarez, Yao Shen, Federico M Giorgi, Alexander Lachmann, B Belinda Ding, B Hilda Ye, and Andrea Califano. Functional characterization of somatic mutations in cancer using network-based inference of protein activity. *Nature Genetics*, 2016.

[6] Elena Alvarez Moreno, Mar Jimenez de la Peña, and Raquel Cano Alonso. Role of New Functional MRI Techniques in the Diagnosis, Staging, and Followup of Gynecological Cancer: Comparison with PET-CT. *Radiology research and practice*, 2012:219546, January 2012.

[7] M Amini, Nicolas Usunier, and Cyril Goutte. Learning from multiple partially observed views-an application to multilingual text categorization. *Advances in neural information ...*, 2009.

[8] Massih-Reza Amini and Cyril Goutte. A co-classification approach to learning from multilingual corpora. *Machine Learning*, 79(1-2):105–121, October 2009.

[9] John Ashburner and Stefan Klöppel. Multivariate models of inter-subject anatomical variability. *NeuroImage*, 56(2):422–39, May 2011.

[10] Atlanta: American Cancer Society. American cancer society: Cancer facts and figures 2016. www.cancer.org/research/cancerfactsstatistics/cancerfactsfigures2014, 2016.

[11] PC Austin and EW Steyerberg. Interpreting the concordance statistic of a logistic regression model: relation to the variance and odds ratio of a continuous explanatory variable. *BMC medical research ...*, 12(1):82, January 2012.

[12] BB Avants, PA Cook, Lyle Ungar, JC Gee, and Murray Grossman. Dementia induces correlated reductions in white matter integrity and cortical thickness: a

multivariate neuroimaging study with sparse canonical correlation analysis. *Neuroimage*, 50(3):1004–1016, 2010.

[13] P Baas, H Schouwink, and FAN Zoetmulder. Malignant pleural mesothelioma. *Annals of oncology*, 9(2):139–150, 1998.

[14] Francis R Bach. Exploring large feature spaces with hierarchical multiple kernel learning. In *Advances in neural information processing systems*, pages 105–112, 2009.

[15] Francis R Bach, Gert RG Lanckriet, and Michael I Jordan. Multiple kernel learning, conic duality, and the smo algorithm. In *Proceedings of the twenty-first international conference on Machine learning*, page 6. ACM, 2004.

[16] Elisa V Bandera, Lawrence H Kushi, and Lorna Rodriguez-Rodriguez. Nutritional factors in ovarian cancer survival. *Nutrition and cancer*, 61(5):580–6, January 2009.

[17] Jordi Barretina, Giordano Caponigro, Nicolas Stransky, Kavitha Venkatesan, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, 483(7391):603–7, March 2012.

[18] Jirina Bartkova, Christina E Hoei-Hansen, Katerina Krizova, Petra Hamerlik, Niels E Skakkebæk, Ewa Rajpert-De Meyts, and Jiri Bartek. Patterns of dna damage response in intracranial germ cell tumors versus glioblastomas reflect cell

of origin rather than brain environment: Implications for the anti-tumor barrier concept and treatment. *Molecular oncology*, 8(8):1667–1678, 2014.

[19] Benjamin Beck and Cédric Blanpain. Unravelling cancer stem cell potential. *Nature reviews. Cancer*, 13(10):727–38, October 2013.

[20] S Ben-David, J Blitzer, and K Crammer. A theory of learning from different domains. *Machine learning*, 2010.

[21] Shai Ben-David, Tyler Lu, and Dávid Pál. Does unlabeled data provably help? worst-case analysis of the sample complexity of semi-supervised learning. In *COLT*, pages 33–44, 2008.

[22] Sean C Bendall, Kara L Davis, El-Ad David Amir, Michelle D Tadmor, Erin F Simonds, Tiffany J Chen, Daniel K Shenfeld, Garry P Nolan, and Dana Pe'er. Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. *Cell*, 157(3):714–25, April 2014.

[23] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

[24] Rameen Beroukhim, Gad Getz, Leia Nghiemphu, Jordi Barretina, Teli Hsueh, David Linhart, Igor Vivanco, Jeffrey C Lee, Julie H Huang, Sethu Alexander, et al. Assessing the significance of chromosomal aberrations in cancer: methodology

and application to glioma. *Proceedings of the National Academy of Sciences*, 104(50):20007–20012, 2007.

[25] Francesco Bertolini, Paola Marighetti, and Yuval Shaked. Cellular and soluble markers of tumor angiogenesis: from patient selection to the identification of the most appropriate postresistance therapy. *Biochimica et biophysica acta*, 1806(2):131–7, December 2010.

[26] Sujeeth Bharadwaj. Multiview feature learning for speech recognition. 2012.

[27] S. Bickel and T. Scheffer. Multi-View Clustering. *Fourth IEEE International Conference on Data Mining (ICDM'04)*, pages 19–26, 2004.

[28] Erhan Bilal, Janusz Dutkowski, Justin Guinney, et al. Improving breast cancer survival analysis through competition-based multidimensional modeling. *PLoS computational biology*, 9(5):e1003047, May 2013.

[29] A Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. *Proceedings of the eleventh annual conference on ...*, pages 92–100, 1998.

[30] Hamid Bolouri, Lue Ping Zhao, and Eric C Holland. Big data visualization identifies the multidimensional molecular landscape of human gliomas. *Proceedings of the National Academy of Sciences of the United States of America*, 113(19):5394–9, 2016.

[31] Paul C Boutros, Adam A Margolin, Joshua M Stuart, Andrea Califano, and Gustavo Stolovitzky. Toward better benchmarking: challenge-based methods assessment in cancer genomics. *Genome biology*, 15(9):1, 2014.

[32] Cameron W Brennan, Roel G W Verhaak, Aaron McKenna, Benito Campos, Houtan Noushmehr, Sofie R Salama, Siyuan Zheng, Debyani Chakravarty, J Zachary Sanborn, Samuel H Berman, et al. The somatic genomic landscape of glioblastoma. *Cell*, 155(2):462–77, October 2013.

[33] Cameron W Brennan, Roel G W Verhaak, Aaron McKenna, et al. The somatic genomic landscape of glioblastoma. *Cell*, 155(2):462–77, October 2013.

[34] David Bryder, Derrick J Rossi, and Irving L Weissman. Hematopoietic stem cells: the paradigmatic tissue-specific stem cell. *The American journal of pathology*, 169(2):338–46, August 2006.

[35] Changmeng Cai, Housheng Hansen He, Sen Chen, Ilsa Coleman, Hongyun Wang, Zi Fang, Shaoyong Chen, Peter S Nelson, X Shirley Liu, Myles Brown, et al. Androgen receptor gene expression in prostate cancer is directly suppressed by the androgen receptor through recruitment of lysine-specific demethylase 1. *Cancer cell*, 20(4):457–471, 2011.

[36] Yidong Cai, Haoxuan Zheng, Wei Gong, Yufang Che, and Bo Jiang. The role of hedgehog signaling pathway in liver regeneration. *Hepato-gastroenterology*, 58(112):2071–2076, 2010.

[37] Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70, 2012.

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

[39] Cancer Genome Atlas Research Network. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113–1120, 2013.

[40] Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. *Nature*, 511(7511):543–550, 2014.

[41] Cancer Genome Atlas Research Network. Integrative clinical genomics of advanced prostate cancer. *Cell*, 161(5):1215–1228, 2015.

[42] Cancer Genome Atlas Research Network. Molecular profiling reveals biologically discrete subsets and pathways of progression in diffuse glioma. *Cell*, 164(3):550–563, 2016.

[43] Hongbao Cao, Dongdong Lin, Junbo Duan, Yu-Ping Wang, and Vince Calhoun. Bio marker identification for diagnosis of schizophrenia with integrated analysis of fmri and snps. In *Bioinformatics and Biomedicine (BIBM), 2012 IEEE International Conference on*, pages 1–6. IEEE, 2012.

[44] Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R Hruschka Jr, and Tom M Mitchell. Toward an architecture for never-ending language learning. In *AAAI*, volume 5, page 3, 2010.

[45] Hannah Carter, Matan Hofree, and Trey Ideker. Genotype to phenotype via network analysis. *Current opinion in genetics & development*, 23(6):611–21, December 2013.

[46] Ramon Casanova, Christopher T Whitlow, Benjamin Wagner, Jeff Williamson, Sally a Shumaker, Joseph a Maldjian, and Mark a Espeland. High dimensional classification of structural MRI Alzheimer's disease data based on large scale regularization. *Frontiers in neuroinformatics*, 5(October):22, January 2011.

[47] M.D. Anderson Cancer Center. Phase ii trial of alisertib (mln8237) in salvage malignant mesothelioma. In: ClinicalTrials.gov [Internet]. Bethesda (MD): National Library of Medicine (US). 2000-2016. Available from: clinicaltrials.gov/ct2/show/NCT00004451 NLM Identifier: NCT00004451.

[48] Ethan Cerami, Jianjiong Gao, Ugur Dogrusoz, Benjamin E Gross, Selcuk Onur Sumer, Bülent Arman Aksoy, et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer discovery*, 2(5):401–4, may 2012.

[49] Nicolò Cesa-Bianchi, David R. Hardoon, and Gayle Leen. Guest Editorial: Learning from multiple sources. *Machine Learning*, 79(1-2):1–3, February 2010.

[50] Uma R Chandran, Changqing Ma, Rajiv Dhir, Michelle Bisceglia, Maureen Lyons-Weiler, Wenjing Liang, George Michalopoulos, Michael Becich, and Federico A

Monzon. Gene expression profiles of prostate cancer reveal involvement of multiple molecular pathways in the metastatic process. *BMC cancer*, 7(1):1, 2007.

[51] Hang Chang, Ju Han, Alexander Borowsky, Leandro Loss, Joe W Gray, Paul T Spellman, and Bahram Parvin. Invariant delineation of nuclear architecture in glioblastoma multiforme for clinical and molecular association. *IEEE transactions on medical imaging*, 32(4):670–82, April 2013.

[52] Kamalika Chaudhuri, Sham M. Kakade, Karen Livescu, and Karthik Sridharan. Multi-view clustering via canonical correlation analysis. *Proceedings of the 26th Annual International Conference on Machine Learning - ICML '09*, pages 1–8, 2009.

[53] Ning Chen, J Zhu, and EP Xing. Predictive subspace learning for multi-view data: a large margin approach. *Advances in neural ...*, 2010.

[54] Xi Chen and Hemant Ishwaran. Pathway hunting by random survival forests. *Bioinformatics (Oxford, England)*, 29(1):99–105, January 2013.

[55] Yong Chen, Jingjing Hao, Wei Jiang, Tong He, Xuegong Zhang, Tao Jiang, and Rui Jiang. Identifying potential cancer driver genes by genomic data integration. *Scientific reports*, 3:3538, January 2013.

[56] Wei-Yi Cheng, Tai-Hsien Ou Yang, and Dimitris Anastassiou. Biomolecular events in cancer revealed by attractor metagenes. *PLoS computational biology*, 9(2):e1002920, January 2013.

[57] Chit Fang Cheok, Chandra S Verma, José Baselga, and David P Lane. Translating p53 into the clinic. *Nature reviews Clinical oncology*, 8(1):25–37, 2011.

[58] SS Choi, SH Cha, and CC Tappert. A survey of binary similarity and distance measures. *Journal of Systemics, Cybernetics and ...*, 0, 2010.

[59] C Christoudias, R Urtasun, and T Darrell. Multi-view learning in the presence of view disagreement. *arXiv preprint arXiv:1206.3242*, 2012.

[60] Carlton Chu, Yizhao Ni, Geoffrey Tan, Craig J Saunders, and John Ashburner. Kernel regression for fMRI pattern prediction. *NeuroImage*, 56(2):662–73, May 2011.

[61] Gil Chu, Jun Li, Balasubramanian Narasimhan, Robert Tibshirani, and Virginia Tusher. Significance analysis of microarrays users guide and technical document. 2001.

[62] Han-Yu Chuang, Eunjung Lee, Yu-Tsueng Liu, Doheon Lee, and Trey Ideker. Network-based classification of breast cancer metastasis. *Molecular systems biology*, 3(140):140, January 2007.

[63] Jon H Chung, Andrew R Larsen, Evan Chen, and Fred Bunz. A ptch1 homolog transcriptionally activated by p53 suppresses hedgehog signaling. *Journal of Biological Chemistry*, 289(47):33020–33031, 2014.

[64] Giovanni Ciriello, Martin L Miller, Bülent Arman Aksoy, Yasin Senbabaoglu,

Nikolaus Schultz, and Chris Sander. Emerging landscape of oncogenic signatures across human cancers. *Nature genetics*, 45(10):1127–1133, September 2013.

[65] Giovanni Ciriello, Martin L Miller, Bülent Arman Aksoy, Yasin Senbabaoglu, Nikolaus Schultz, and Chris Sander. Emerging landscape of oncogenic signatures across human cancers. *Nature genetics*, 45(10):1127–1133, 2013.

[66] Rivka R Colen, Mark Vangel, Jixin Wang, David a Gutman, et al. Imaging genomic mapping of an invasive MRI phenotype predicts patient outcome and metabolic dysfunction: a TCGA glioma phenotype research group project. *BMC medical genomics*, 7(1):30, January 2014.

[67] Ivan G Costa, Stefan Roepcke, Christoph Hafemeister, and Alexander Schliep. Inferring differentiation pathways from gene expression. *Bioinformatics (Oxford, England)*, 24(13):i156–64, July 2008.

[68] Ivan G Costa, Stefan Roepcke, and Alexander Schliep. Gene expression trees in lymphoid development. *BMC immunology*, 8:25, January 2007.

[69] James C Costello, Laura M Heiser, Elisabeth Georgii, Mehmet Gönen, et al. A community effort to assess and improve drug sensitivity prediction algorithms. *Nature biotechnology*, June 2014.

[70] Prue A. Cowin, Joshy George, Sian Fereday, Elizabeth Loehrer, Peter Van Loo, Carleen Cullinane, Dariush Etemadmoghadam, Sarah Ftouni, Laura Galletta, Michael S. Anglesio, Joy Hendley, Leanne Bowes, Karen E. Sheppard, Eliza-

beth L. Christie, Richard B. Pearson, Paul R. Harnett, Viola Heinzelmann-Schwarz, Michael Friedlander, Orla McNally, Michael Quinn, Peter Campbell, Anna DeFazio, and David D L Bowtell. LRP1B deletion in high-grade serous ovarian cancers is associated with acquired chemotherapy resistance to liposomal doxorubicin. *Cancer Research*, 72(16):4060–4073, 2012.

[71] Glenn S Cowley, Barbara A Weir, Francisca Vazquez, Pablo Tamayo, Justine A Scott, Scott Rusin, Alexandra East-Seletsky, Levi D Ali, William FJ Gerath, Sarah E Pantel, et al. Parallel genome-scale loss of function screens in 216 cancer cell lines for the identification of context-specific genetic dependencies. *Scientific data*, 1, 2014.

[72] Nello Cristianini, Jaz Kandola, Andre Elisseeff, and John Shawe-Taylor. On kernel target alignment. In DawnE. Holmes and LakhmiC. Jain, editors, *Innovations in Machine Learning*, volume 194 of *Studies in Fuzziness and Soft Computing*, pages 205–256. Springer Berlin Heidelberg, 2006.

[73] Mark Culp and George Michailidis. A cotraining algorithm for multiview data with applications in data fusion. *Journal of chemometrics*, pages 1–20, 2009.

[74] JPC Da Cunha, PAF Galante, JE De Souza, RF De Souza, PM Carvalho, DT Ohara, RP Moura, SM Oba-Shinja, SKN Marie, WA Silva, et al. Bioinformatics construction of the human cell surfaceome. *Proceedings of the National Academy of Sciences*, 106(39):16752–16757, 2009.

[75] Phuong Dao, Kendric Wang, Colin Collins, Martin Ester, Anna Lapuk, and S Cenk Sahinalp. Optimally discriminative subnetwork markers predict response to chemotherapy. *Bioinformatics (Oxford, England)*, 27(13):i205–13, July 2011.

[76] Sanjoy Dasgupta. PAC generalization bounds for co-training. *Advances in neural ...*, 2002.

[77] FG Davis. Toward determining the lifetime occurrence of metastatic brain tumors estimated from 2007 United States cancer incidence data. *Neuro- ...*, 14(9):1171–1177, 2012.

[78] Pasquale De Bonis, Carmelo Anile, Angelo Pompucci, Alba Fiorentino, Mario Balducci, Silvia Chiesa, Libero Lauriola, Giulio Maira, and Annunziato Mangiola. The influence of surgery on recurrence pattern of glioblastoma. *Clinical neurology and neurosurgery*, 115(1):37–43, January 2013.

[79] Oguz Demirci, Vincent P. Clark, and Vince D. Calhoun. A projection pursuit algorithm to classify individuals using fMRI data: Application to schizophrenia. *NeuroImage*, 39(4):1774–1782, February 2008.

[80] Peter B Dirks. Brain tumour stem cells: the undercurrents of human brain cancer and their relationship to neural stem cells. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 363(1489):139–52, January 2008.

[81] Jesse R Dixon, Siddarth Selvaraj, Feng Yue, Audrey Kim, Yan Li, Yin Shen, Ming Hu, Jun S Liu, and Bing Ren. Topological domains in mammalian genomes

identified by analysis of chromatin interactions. *Nature*, 485(7398):376–80, May 2012.

[82] P K Douglas, Sam Harris, Alan Yuille, and Mark S Cohen. Performance comparison of machine learning algorithms and number of independent components used in fMRI decoding of belief vs. disbelief. *NeuroImage*, 56(2):544–53, May 2011.

[83] Sylvia Drabycz, Gloria Roldán, Paula de Robles, Daniel Adler, John B McIntyre, Anthony M Magliocco, J Gregory Cairncross, and J Ross Mitchell. An analysis of image texture, tumor location, and MGMT promoter methylation in glioblastoma using magnetic resonance imaging. *NeuroImage*, 49(2):1398–405, January 2010.

[84] J. M. Drake, N. a. Graham, T. Stoyanova, a. Sedghi, a. S. Goldstein, H. Cai, D. a. Smith, H. Zhang, E. Komisopoulou, J. Huang, T. G. Graeber, and O. N. Witte. Oncogene-specific activation of tyrosine kinase networks during prostate cancer progression. *Proceedings of the National Academy of Sciences*, 109(5):1643–1648, 2012.

[85] Michael J Duffy, Naoise C Synnott, Patricia M McGowan, John Crown, Darran OConnor, and William M Gallagher. p53 as a target for the treatment of cancer. *Cancer treatment reviews*, 40(10):1153–1160, 2014.

[86] Janusz Dutkowski and Trey Ideker. Protein networks as logic functions in development and cancer. *PLoS computational biology*, 7(9):e1002180, September 2011.

[87] Benjamin M Ellingson, Timothy F Cloughesy, Albert Lai, Phioanh L Nghiemphu, Linda M Liau, and Whitney B Pope. High order diffusion tensor imaging in human glioblastoma. *Academic radiology*, 18(8):947–54, August 2011.

[88] Benjamin M Ellingson, Timothy F Cloughesy, Whitney B Pope, Taryar M Zaw, Heidi Phillips, Shadi Lalezari, Phioanh L Nghiemphu, et al. Anatomic localization of O6-methylguanine DNA methyltransferase (MGMT) promoter methylated and unmethylated tumors: a radiographic study in 358 de novo human glioblastomas. *NeuroImage*, 59(2):908–16, January 2012.

[89] N ello Cristianini, Andre Elisseeff, John Shawe-Taylor, and Jaz Kandola. On kernel-target alignment. *NIPS*, 2001.

[90] D Erhan, Y Bengio, and A Courville. Why does unsupervised pre-training help deep learning? *... of Machine Learning ...*, 11:625–660, 2010.

[91] Nicholas Erho, Anamaria Crisan, Ismael a Vergara, Anirban P Mitra, Mercedeh Ghadessi, Christine Buerki, Eric J Bergstralh, Thomas Kollmeyer, et al. Discovery and validation of a prostate cancer genomic classifier that predicts early metastasis following radical prostatectomy. *PloS one*, 8(6):e66855, January 2013.

[92] Georg Eschenburg, Angelika Eggert, Alexander Schramm, Holger N Lode, and Patrick Hundsdoerfer. Smac mimetic lbw242 sensitizes xiap-overexpressing neuroblastoma cells for tnf-α–independent apoptosis. *Cancer research*, 72(10):2645–2656, 2012.

[93] Alan C Evans, D Louis Collins, SR Mills, ED Brown, RL Kelly, and Terry M Peters. 3d statistical neuroanatomical models from 305 mri volumes. In *Nuclear Science Symposium and Medical Imaging Conference, 1993., 1993 IEEE Conference Record.*, pages 1813–1817. IEEE, 1993.

[94] Ali Faisal and R Louhimo. Biomarker discovery via dependency analysis of multi-view functional genomics data. *Advances in ...*, pages 1–7, 2011.

[95] Lulu Fan, Zhonghu Yuan, Xiaowei Han, and Wenwu Hua. Overview of content-based image feature extraction methods. In *International Conference on Computer, Networks and Communication Engineering (ICCNCE 2013)*. Atlantis Press, 2013.

[96] Giovanna Fattovich, Tommaso Stroffolini, Irene Zagni, and Francesco Donato. Hepatocellular carcinoma in cirrhosis: incidence and risk factors. *Gastroenterology*, 127(5):S35–S50, 2004.

[97] Andriy Fedorov, Reinhard Beichel, Jayashree Kalpathy-Cramer, Julien Finet, Jean-Christophe Fillion-Robin, Sonia Pujol, Christian Bauer, Dominique Jennings, Fiona Fennessy, Milan Sonka, et al. 3d slicer as an image computing platform for the quantitative imaging network. *Magnetic resonance imaging*, 30(9):1323–1341, 2012.

[98] Maurizio Filippone, Francesco Camastra, Francesco Masulli, and Stefano Rovetta.

A survey of kernel and spectral methods for clustering. *Pattern recognition*, 41(1):176–190, 2008.

[99] Thomas Finley and Thorsten Joachims. Supervised clustering with support vector machines. In *Proceedings of the 22nd international conference on Machine learning*, pages 217–224. ACM, 2005.

[100] Santo Fortunato and Marc Barthélemy. Resolution limit in community detection. *Proceedings of the National Academy of Sciences of the United States of America*, 104(1):36–41, 2007.

[101] Dean P Foster, Sham M Kakade, and Tong Zhang. Multi-view dimensionality reduction via canonical correlation analysis. In *Technical Report TR-2008-4, TTI-Chicago*. TTI-Chicago, 2008.

[102] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1, 2010.

[103] Anna V Galkin, Jonathan S Melnick, Sungjoon Kim, Tami L Hood, Nanxin Li, Lintong Li, Gang Xia, Ruo Steensma, Greg Chopiuk, Jiqing Jiang, et al. Identification of nvp-tae684, a potent, selective, and efficacious inhibitor of npm-alk. *Proceedings of the National Academy of Sciences*, 104(1):270–275, 2007.

[104] Hui K Gan, Andrew H Kaye, and Rodney B Luwor. The EGFRvIII variant in

glioblastoma multiforme. *Journal of clinical neuroscience : official journal of the Neurosurgical Society of Australasia*, 16(6):748–54, June 2009.

[105] Maxime Garcia, Raphaelle Millat-Carus, François Bertucci, Pascal Finetti, Daniel Birnbaum, and Ghislain Bidaut. Interactome-transcriptome integration for predicting distant metastasis in breast cancer. *Bioinformatics (Oxford, England)*, 28(5):672–8, March 2012.

[106] Mathew J Garnett, Elena J Edelman, Sonja J Heidorn, et al. Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature*, 483(7391):570–5, March 2012.

[107] Levi A Garraway and Eric S Lander. Lessons from the cancer genome. *Cell*, 153(1):17–37, 2013.

[108] The Gene and Ontology Consortium. The Gene Ontology: enhancements for 2011. *Nucleic acids research*, 40(Database issue):D559–64, January 2012.

[109] Thomas a Gerds, Michael W Kattan, Martin Schumacher, and Changhong Yu. Estimating a timedependent concordance index for survival prediction models with covariate dependent censoring. *Statistics in medicine*, 32(13):2173–84, June 2013.

[110] Gianluigi Giannelli, Erica Villa, and Michael Lahn. Transforming growth factor-β as a therapeutic target in hepatocellular carcinoma. *Cancer research*, 74(7):1890–1894, 2014.

[111] Michelle Girvan and Mark EJ Newman. Community structure in social and biological networks. *Proceedings of the national academy of sciences*, 99(12):7821–7826, 2002.

[112] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 513–520, 2011.

[113] Mary Goldman, Brian Craft, Teresa Swatloski, Melissa Cline, Olena Morozova, Mark Diekhans, David Haussler, and Jingchun Zhu. The ucsc cancer genomics browser: update 2015. *Nucleic acids research*, page gku1073, 2014.

[114] Abhishek Golugula, George Lee, Stephen R Master, Michael D Feldman, John E Tomaszewski, David W Speicher, and Anant Madabhushi. Supervised regularized canonical correlation analysis: integrating histologic and proteomic measurements for predicting biochemical recurrence following prostate surgery. *BMC bioinformatics*, 12(1):483, January 2011.

[115] M Gönen. A Bayesian Multiple Kernel Learning Framework for Single and Multiple Output Regression. *ECAI*, pages 354–359, 2012.

[116] M Gonen. Bayesian efficient multiple kernel learning. *arXiv preprint arXiv:1206.6465*, 2012.

[117] M Gönen and E Alpaydn. Multiple kernel learning algorithms. *The Journal of Machine Learning Research*, 12:2211–2268, 2011.

[118] Mehmet Gönen. Predicting drug-target interactions from chemical and genomic kernels using Bayesian matrix factorization. *Bioinformatics (Oxford, England)*, 28(18):2304–10, September 2012.

[119] Mehmet Gönen. Coupled dimensionality reduction and classification for supervised and semi-supervised multilabel learning. *Pattern recognition letters*, 38:132–141, March 2014.

[120] Mehmet Gönen and Ethem Alpaydn. Localized algorithms for multiple kernel learning. *Pattern Recognition*, 46(3):795–807, March 2013.

[121] Mehmet Gönen and Adam A Margolin. Localized data fusion for kernel k-means clustering with application to cancer biology. In *Advances in Neural Information Processing Systems*, pages 1305–1313, 2014.

[122] Abel Gonzalez-Perez, Ville Mustonen, Boris Reva, Graham R S Ritchie, et al. Computational approaches to identify functional genetic variants in cancer genomes. *Nature methods*, 10(8):723–9, August 2013.

[123] MG Grabherr, BJ Haas, and M Yassour. Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nature . . .*, 29(7):644–652, 2013.

[124] Kiley Graim. *Leveraging expression and network data for protein function prediction*. PhD thesis, Colorado State University, 2007.

[125] Catherine S Grasso, Yi-Mi Wu, Dan R Robinson, Xuhong Cao, Saravana M Dhanasekaran, Amjad P Khan, Michael J Quist, Xiaojun Jing, Robert J Lonigro, J Chad Brenner, et al. The mutational landscape of lethal castration-resistant prostate cancer. *Nature*, 487(7406):239–243, 2012.

[126] Malachi Griffith, Obi L Griffith, Adam C Coffman, et al. DGIdb: mining the druggable genome. *Nature methods*, (october):1–7, October 2013.

[127] F Peter Guengerich. Cytochrome p450 and chemical toxicology. *Chemical research in toxicology*, 21(1):70–83, 2007.

[128] Carlos Guestrin. Co-Training for Semi- supervised learning (cont .) semi-supervised learning, 2007. Lecture given at Cernegie Mellon University on April 23, 2007.

[129] Jordi Guiu, Dylan JM Bergen, Emma De Pater, Abul BMMK Islam, Verónica Ayllón, Leonor Gama-Norton, Cristina Ruiz-Herguido, Jessica González, Nuria López-Bigas, Pablo Menendez, et al. Identification of cdca7 as a novel notch transcriptional target involved in hematopoietic stem cell emergence. *The Journal of experimental medicine*, 211(12):2411–2423, 2014.

[130] Guoji Guo, Sidinh Luc, Eugenio Marco, Ta-Wei Lin, Cong Peng, Marc a Kerenyi, Semir Beyaz, Woojin Kim, Jian Xu, Partha Pratim Das, Tobias Neff, Keyong Zou,

Guo-Cheng Yuan, and Stuart H Orkin. Mapping cellular hierarchy by single-cell analysis of the cell surface repertoire. *Cell stem cell*, 13(4):492–505, October 2013.

[131] David A Gutman, Lee AD Cooper, Scott N Hwang, Chad A Holder, JingJing Gao, Tarun D Aurora, William D Dunn Jr, Lisa Scarpace, Tom Mikkelsen, Rajan Jain, et al. Mr imaging predictors of molecular profile and survival: multi-institutional study of the tcga glioblastoma data set. *Radiology*, 267(2):560–569, 2013.

[132] David A Gutman, Lee AD Cooper, Scott N Hwang, Chad A Holder, JingJing Gao, Tarun D Aurora, William D Dunn Jr, Lisa Scarpace, Tom Mikkelsen, Rajan Jain, et al. Mr imaging predictors of molecular profile and survival: multi-institutional study of the tcga glioblastoma data set. *Radiology*, 267(2):560–569, 2013.

[133] Brian J Haas, Alexie Papanicolaou, Moran Yassour, Manfred Grabherr, et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature protocols*, 8(8):1494–512, August 2013.

[134] Kerstin Hackmack, Friedemann Paul, Martin Weygandt, Carsten Allefeld, and John-Dylan Haynes. Multi-scale classification of disease using structural MRI and wavelet transform. *NeuroImage*, 62(1):48–58, August 2012.

[135] Marc Hafner, Mario Niepel, Mirra Chung, and Peter K Sorger. Growth rate inhibition metrics correct for confounders in measuring sensitivity to cancer drugs. *Nature Methods*, 13(October 2015):1–11, 2016.

[136] David K Hammond, Pierre Vandergheynst, and Rémi Gribonval. Wavelets on graphs via spectral graph theory. *Applied and Computational Harmonic Analysis*, 30(2):129–150, 2011.

[137] Douglas Hanahan and Robert A Weinberg. Hallmarks of cancer: the next generation. *cell*, 144(5):646–674, 2011.

[138] Blaise Hanczar, Jianping Hua, Chao Sima, John Weinstein, Michael Bittner, and Edward R Dougherty. Small-sample precision of ROC-related estimates. *Bioinformatics (Oxford, England)*, 26(6):822–30, March 2010.

[139] D Hardoon, Sandor Szedmak, and J Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, 2003.

[140] Tatsunori Hashimoto, Tommi Jaakkola, Richard Sherwood, Esteban O Mazzoni, Hynek Wichterle, and David Gifford. Lineage-based identification of cellular states and expression programs. *Bioinformatics (Oxford, England)*, 28(12):i250–7, June 2012.

[141] Sebastian D Hayes, Han Liu, Ewan MacDonald, Christopher M Sanderson, Judy M Coulson, Michael J Clague, and Sylvie Urbé. Direct and indirect control of mitogen-activated protein kinase pathway-associated components, brap/imp e3 ubiquitin ligase and craf/raf1 kinase, by the deubiquitylating enzyme usp15. *Journal of Biological Chemistry*, 287(51):43007–43018, 2012.

[142] Jingrui He and Rick Lawrence. A graph-based framework for multi-task multi-view

learning. *Proceedings of the 28th International Conference on Machine Learning, ICML 2011*, pages 25–32, 2011.

[143] Laura M Heiser, Anguraj Sadanandam, Wen-Lin Kuo, Stephen C Benz, Theodore C Goldstein, Sam Ng, William J Gibb, Nicholas J Wang, Safiyyah Ziyad, Frances Tong, et al. Subtype and pathway specific responses to anticancer compounds in breast cancer. *Proceedings of the National Academy of Sciences*, 109(8):2724–2729, 2012.

[144] Ying C Henderson, Yunyun Chen, Mitchell J Frederick, Stephen Y Lai, and G. L. Clayman. MEK Inhibitor PD0325901 Significantly Reduces the Growth of Papillary Thyroid Carcinoma Cells In vitro and In vivo. *Molecular Cancer Therapeutics*, 9(7):1968–1976, jul 2010.

[145] Tracy SP Heng, Michio W Painter, Kutlu Elpek, Veronika Lukacs-Kornek, Nora Mauermann, Shannon J Turley, Daphne Koller, et al. The immunological genome project: networks of gene expression in immune cells. *Nature immunology*, 9(10):1091–1094, 2008.

[146] Roy S Herbst, Jean-Charles Soria, Marcin Kowanetz, Gregg D Fine, Omid Hamid, Michael S Gordon, Jeffery A Sosman, David F McDermott, John D Powderly, Scott N Gettinger, et al. Predictive correlates of response to the anti-pd-l1 antibody mpdl3280a in cancer patients. *Nature*, 515(7528):563–567, 2014.

[147] Steven M Hill, Laura M Heiser, Thomas Cokelaer, Michael Unger, Nicole K Nesser,

Daniel E Carlin, Yang Zhang, Artem Sokolov, Evan O Paull, Chris K Wong, et al. Inferring causal molecular networks: empirical assessment through a community-based effort. *Nature methods*, 13(4):310–318, 2016.

[148] Katherine A Hoadley, Christina Yau, Denise M Wolf, Andrew D Cherniack, David Tamborero, Sam Ng, Max DM Leiserson, Beifang Niu, Michael D McLellan, Vladislav Uzunangelov, et al. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell*, 158(4):929–944, 2014.

[149] Matan Hofree, John P Shen, Hannah Carter, Andrew Gross, and Trey Ideker. Network-based stratification of tumor mutations. *Nature methods*, 10(11):1108–15, November 2013.

[150] E C Holland. Glioblastoma multiforme: the terminator. *Proceedings of the National Academy of Sciences of the United States of America*, 97(12):6242–4, June 2000.

[151] Chao-Hui Huang, Antoine Veillard, Ludovic Roux, Nicolas Loménie, and Daniel Racoceanu. Time-efficient sparse analysis of histopathological whole slide images. *Computerized medical imaging and graphics : the official journal of the Computerized Medical Imaging Society*, 35(7-8):579–91, 2011.

[152] HUGO Gene Nomenclature Committee (HGNC). Gene family: Chromatin-modifying enzymes. Accessed July 15, 2015.

[153] Ilkka Huopaniemi, Tommi Suvitaival, Janne Nikkilä, Matej Oresic, and Samuel Kaski. Multivariate multi-way analysis of multi-source data. *Bioinformatics (Oxford, England)*, 26(12):i391–8, June 2010.

[154] Frank Hutter, H Hoos, and K Leyton-Brown. An Efficient Approach for Assessing Hyperparameter Importance. *. . . of The 31st International Conference on . . .*, 32, 2014.

[155] SJ Hwang, Kristen Grauman, and F Sha. Semantic Kernel Forests from Multiple Taxonomies. *Advances in Neural Information . . .*, 2012.

[156] Trey Ideker, Janusz Dutkowski, and Leroy Hood. Boosting signal-to-noise in complex biology: prior knowledge is power. *Cell*, 144(6):860–3, March 2011.

[157] Francesco Iorio, Theo A Knijnenburg, Daniel J Vis, Julio Saez-Rodriguez, Ultan Mcdermott, Mathew J Garnett Correspondence, Graham R Bignell, Michael P Menden, et al. A Landscape of Pharmacogenomic Interactions in Cancer. *Cell*, 16613616(421):1–15, 2016.

[158] Hemant Ishwaran and UB Kogalur. Random survival forests for highdimensional data. *Statistical analysis and . . .*, 2011.

[159] Anil K Jain. Data clustering: 50 years beyond k-means. *Pattern recognition letters*, 31(8):651–666, 2010.

[160] Rajan Jain, Laila Poisson, Jayant Narang, David Gutman, Lisa Scarpace, Scott N Hwang, Chad Holder, Max Wintermark, Rivka R Colen, Justin Kirby, John Frey-

mann, Daniel J Brat, Carl Jaffe, and Tom Mikkelsen. Genomic mapping and survival prediction in glioblastoma: molecular subclassification strengthened by hemodynamic imaging biomarkers. *Radiology*, 267(1):212–20, April 2013.

[161] Neema Jamshidi, Maximilian Diehn, Markus Bredel, and MD Kuo. Illuminating Radiogenomic Characteristics of Glioblastoma Multiforme through Integration of MR Imaging, Messenger RNA Expression, and DNA Copy Number Variation. *Radiology*, 270(1), 2013.

[162] IS Jang, Rodrigo Dienstmann, Adam A Margolin, and Justin Guinney. Stepwise group sparse regression (sgsr): gene-set-based pharmacogenomic predictive models with stepwise selection of functional priors. In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, volume 20, pages 32–43. NIH Public Access, 2014.

[163] Rodolphe Jenatton, Alexandre Gramfort, Vincent Michel, Guillaume Obozinski, Francis Bach, and Bertrand Thirion. Multi-scale Mining of fMRI Data with Hierarchical Structured Sparsity. *2011 International Workshop on Pattern Recognition in NeuroImaging*, 0(2):69–72, May 2011.

[164] Peng Jiang, Hongfang Wang, Wei Li, Chongzhi Zang, Bo Li, Yinling J. Wong, Cliff Meyer, Jun S. Liu, Jon C. Aster, and X. Shirley Liu. Network analysis of gene essentiality in functional genomics experiments. *Genome Biology*, 16(1):239, 2015.

[165] Gaole Jin and Raviv Raich. On surrogate supervision multiview learning. *Machine Learning for Signal Processing*, 2012.

[166] Aart G Jochemsen. Reactivation of p53 as therapeutic intervention for malignant melanoma. *Current opinion in oncology*, 26(1):114–119, 2014.

[167] W Evan Johnson, Cheng Li, and Ariel Rabinovic. Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*, 8(1):118–127, 2007.

[168] Adrian M Jubb and Adrian L Harris. Biomarkers to predict the clinical efficacy of bevacizumab in cancer. *The lancet oncology*, 11(12):1172–83, December 2010.

[169] Aleksandra Jurek, Kenichi Amagasaki, Agnieszka Gembarska, Carl-Henrik Heldin, and Johan Lennartsson. Negative and positive regulation of mapk phosphatase 3 controls platelet-derived growth factor-induced erk activation. *Journal of Biological Chemistry*, 284(7):4626–4634, 2009.

[170] Bogumił Kaczkowski, Elfar Torarinsson, Kristin Reiche, Jakob Hull Havgaard, Peter F. Stadler, and Jan Gorodkin. Structural profiles of human miRNA families from pairwise clustering. *Bioinformatics*, 25(3):291–294, 2009.

[171] Sham M Kakade and Dean P Foster. Multi-view regression via canonical correlation analysis. In *International Conference on Computational Learning Theory*, pages 82–96. Springer, 2007.

[172] Sham M Kakade and Dean P Foster. Multi-view regression via canonical correlation analysis. In *International Conference on Computational Learning Theory*, pages 82–96. Springer, 2007.

[173] Melih Kandemir, Akos Vetek, Mehmet Gönen, Arto Klami, and Samuel Kaski. Multi-task and multi-view learning of user state. *Neurocomputing*, 139:97–106, September 2014.

[174] Cyriac Kandoth, Michael D McLellan, Fabio Vandin, Kai Ye, Beifang Niu, Charles Lu, Mingchao Xie, Qunyuan Zhang, Joshua F McMichael, Matthew A Wyczalkowski, et al. Mutational landscape and significance across 12 major cancer types. *Nature*, 502(7471):333–339, 2013.

[175] Dietrich Keppler. Multidrug resistance proteins (mrps, abccs): importance for pathophysiology and drug therapy. In *Drug Transporters*, pages 299–323. Springer, 2011.

[176] Siddharth Khullar, Andrew Michael, Nicolle Correa, Tulay Adali, Stefi a Baum, and Vince D Calhoun. Wavelet-based fMRI analysis: 3-D denoising, signal separation, and validation metrics. *NeuroImage*, 54(4):2867–84, February 2011.

[177] Ekta Khurana, Yao Fu, Jieming Chen, and Mark Gerstein. Interpretation of genomic variants using a unified biological network approach. *PLoS Comput Biol*, 9(3):e1002886, 2013.

[178] Patrick J Killela, Zachary J Reitman, Yuchen Jiao, Chetan Bettegowda, et al.

TERT promoter mutations occur frequently in gliomas and a subset of tumors derived from cells with low rates of self-renewal. *Proceedings of the National Academy of Sciences of the United States of America*, 110(15):6021–6, April 2013.

[179] PJ Killela, CJ Pirozzi, Patrick Healy, and ZJ Reitman. Mutations in IDH1, IDH2, and in the TERT promoter define clinically distinct subgroups of adult malignant gliomas. *Oncotarget*, 5(6), 2014.

[180] Hyungsoo Kim, Dennie T Frederick, Mitchell P Levesque, Zachary A Cooper, Yongmei Feng, Clemens Krepler, Laurence Brill, Yardena Samuels, Nicholas K Hayward, Ally Perlina, et al. Downregulation of the ubiquitin ligase rnf125 underlies resistance of melanoma cells to braf inhibitors via jak1 deregulation. *Cell reports*, 11(9):1458–1473, 2015.

[181] S Kim, S Nowozin, P Kohli, and CD Yoo. Higher-order correlation clustering for image segmentation. *Advances in Neural Information ...*, 34(3):533–40, March 2011.

[182] Sungwoong Kim, Sebastian Nowozin, Pushmeet Kohli, and Yoo Chang D. Higher-Order Correlation Clustering for Image Segmentation. *Nips*, pages 1–9, 2011.

[183] Tae Hoon Kim, Kyoung Mu Lee, and Sang Uk Lee. Learning full pairwise affinities for spectral segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 35(7):1690–1703, 2013.

[184] Daniel C Kirouac, Julio Saez-Rodriguez, Jennifer Swantek, John M Burke, Dou-

glas A Lauffenburger, and Peter K Sorger. Creating and analyzing pathway and protein interaction compendia for modelling signal transduction networks. *BMC systems biology*, 6:29, jan 2012.

[185] Arto Klami, S Virtanen, and S Kaski. Bayesian canonical correlation analysis. *The Journal of Machine Learning ...*, 14:965–1003, 2013.

[186] Jon M Kleinberg. Navigation in a small world. *Nature*, 406(6798):845, 2000.

[187] Younhee Ko, Seth a Ament, James a Eddy, Juan Caballero, John C Earls, Leroy Hood, and Nathan D Price. Cell type-specific genes show striking and distinct patterns of spatial expression in the mouse brain. *Proceedings of the National Academy of Sciences of the United States of America*, 110(8):3095–100, February 2013.

[188] Daphne Koller, Nir Friedman, Lise Getoor, and Ben Taskar. Graphical Models in a Nutshell. *Introduction to Statistical Relational Learning*, page 43, 2007.

[189] J. Kong, O. Sertel, H. Shimada, K.L. Boyer, J.H. Saltz, and M.N. Gurcan. Computer-aided evaluation of neuroblastoma on whole-slide histology images: Classifying grade of neuroblastic differentiation. *Pattern Recognition*, 42(6):1080–1092, June 2009.

[190] Jun Kong, Lee a D Cooper, Fusheng Wang, David a Gutman, Jingjing Gao, Candace Chisolm, and Ashish Sharma. Integrative, Multi-model Analysis of Glioblas-

toma Using TCGA Molecular Data, Pathology Images and Clinical Outcomes. *IEEE transactions on bio-medical engineering*, 58(12):3469–3474, 2012.

[191] Jun Kong, Lee a D Cooper, Fusheng Wang, David a Gutman, Jingjing Gao, Candace Chisolm, Ashish Sharma, Tony Pan, Erwin G Van Meir, Tahsin M Kurc, Carlos S Moreno, Joel H Saltz, and Daniel J Brat. Integrative, multimodal analysis of glioblastoma using TCGA molecular data, pathology images, and clinical outcomes. *IEEE transactions on bio-medical engineering*, 58(12):3469–74, December 2011.

[192] Hans-Peter Kriegel and Arthur Zimek. Subspace clustering, ensemble clustering, alternative clustering, multiview clustering: what can we learn from each other. In *Proceedings of the 1st international workshop on discovering, summarizing and using multiple clusterings (MultiClust) held in conjunction with KDD*, 2010.

[193] Vessela N Kristensen, Ole Christian Lingjærde, Hege G Russnes, Hans Kristian M Vollan, Arnoldo Frigessi, Anne-Lise Børresen-Dale, and Hans Kristian M Vollan. Principles and methods of integrative genomic analyses in cancer. *Nature reviews. Cancer*, 14(5):299–313, 2014.

[194] Damjan Krstajic, Ljubomir J Buturovic, David E Leahy, and Simon Thomas. Cross-validation pitfalls when selecting and assessing regression and classification models. *Journal of cheminformatics*, 6(1):10, March 2014.

[195] Abhishek Kumar and H Daumé. A co-training approach for multi-view spectral clustering. *Proceedings of the 28th ...*, 2011.

[196] MD Kuo and N Jamshidi. Behind the Numbers: Decoding Molecular Phenotypes with Radiogenomics– Guiding Principles and Technical Considerations. *Radiology*, 270(2), 2014.

[197] Olcay Kursun and Ethem Alpaydin. Canonical correlation analysis for multiview semisupervised feature extraction. *Artificial Intelligence and Soft Computing*, pages 430–436, 2010.

[198] Roselyne M Labbé, Manuel Irimia, Ko W Currie, Alexander Lin, Shu Jun Zhu, David D R Brown, Eric J Ross, et al. A comparative transcriptomic analysis reveals conserved features of stem cell pluripotency in planarians and mammals. *Stem cells (Dayton, Ohio)*, 30(8):1734–45, August 2012.

[199] Peter Langfelder and Steve Horvath. WGCNA: an R package for weighted correlation network analysis. *BMC bioinformatics*, 9:559, 2008.

[200] Peter Langfelder and Steve Horvath. Wgcna: an r package for weighted correlation network analysis. *BMC bioinformatics*, 9(1):1, 2008.

[201] James Larkin, Vanna Chiarion-Sileni, Rene Gonzalez, Jean Jacques Grob, C Lance Cowey, Christopher D Lao, Dirk Schadendorf, Reinhard Dummer, Michael Smylie, Piotr Rutkowski, et al. Combined nivolumab and ipilimumab or monotherapy in untreated melanoma. *N Engl J Med*, 2015(373):23–34, 2015.

[202] Louise Laurent, Eleanor Wong, Guoliang Li, Tien Huynh, Aristotelis Tsirigos, Chin Thing Ong, Hwee Meng Low, Ken Wing Kin Sung, Isidore Rigoutsos, Jeanne Loring, and Chia-Lin Wei. Dynamic changes in the human methylome during differentiation. *Genome research*, 20(3):320–31, March 2010.

[203] Vivian Law, Craig Knox, Yannick Djoumbou, Tim Jewison, An Chi Guo, Yifeng Liu, Adam Maciejewski, David Arndt, Michael Wilson, Vanessa Neveu, et al. Drugbank 4.0: shedding new light on drug metabolism. *Nucleic acids research*, 42(D1):D1091–D1097, 2014.

[204] Michael S Lawrence, Petar Stojanov, Craig H Mermel, James T Robinson, Levi A Garraway, Todd R Golub, Matthew Meyerson, Stacey B Gabriel, Eric S Lander, and Gad Getz. Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature*, 505(7484):495–501, 2014.

[205] Michael S. Lawrence, Petar Stojanov, Paz Polak, Gregory V. Kryukov, Kristian Cibulskis, Andrey Sivachenko, Scott L. Carter, Chip Stewart, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, 499(7457):214–218, 07 2013.

[206] Eun Kyung Lee, Ki-Wook Chung, Sun Kyung Yang, Min Ji Park, Hye Sook Min, Seok Won Kim, and Han Sung Kang. Dna methylation of mapk signal-inhibiting genes in papillary thyroid carcinoma. *Anticancer research*, 33(11):4833–4839, 2013.

[207] Celine Lefebvre, Presha Rajbhandari, Mariano J Alvarez, Pradeep Bandaru, et al. A human B-cell interactome identifies MYB and FOXM1 as master regulators of proliferation in germinal centers. *Molecular systems biology*, 6(377):377, June 2010.

[208] A Leiblich, SS Cross, JWF Catto, JT Phillips, HY Leung, FC Hamdy, and I Rehman. Lactate dehydrogenase-b is silenced by promoter hypermethylation in human prostate cancer. *Oncogene*, 25(20):2953–2960, 2006.

[209] Ed S Lein, Michael J Hawrylycz, Nancy Ao, Mikael Ayres, Amy Bensinger, et al. Genome-wide atlas of gene expression in the adult mouse brain. *Nature*, 445(7124):168–76, January 2007.

[210] C Chang-Yew Leow, Steven Gerondakis, and Andrew Spencer. Mek inhibitors as a chemotherapeutic intervention in multiple myeloma. *Blood cancer journal*, 3(3):e105, 2013.

[211] a. Lex, M. Streit, H.-J. Schulz, C. Partl, D. Schmalstieg, P.J. Park, and N. Gehlenborg. StratomeX: Visual Analysis of Large-Scale Heterogeneous Genomics Data for Cancer Subtype Characterization. *Computer Graphics Forum*, 31(3pt3):1175–1184, jun 2012.

[212] Chenwei Li, David G Heidt, Piero Dalerba, Charles F Burant, Lanjing Zhang, Volkan Adsay, Max Wicha, Michael F Clarke, and Diane M Simeone. Identification of pancreatic cancer stem cells. *Cancer research*, 67(3):1030–7, February 2007.

[213] Yue-Ming Li, Min Xu, Ming-Tain Lai, Qian Huang, José L Castro, Jillian DiMuzio-Mower, Timothy Harrison, Colin Lellis, Alan Nadin, Joseph G Neduvelil, et al. Photoactivated γ-secretase inhibitors directed to the active site covalently label presenilin 1. *Nature*, 405(6787):689–694, 2000.

[214] Arthur Liberzon, Aravind Subramanian, Reid Pinchback, Helga Thorvaldsdóttir, Pablo Tamayo, and Jill P Mesirov. Molecular signatures database (msigdb) 3.0. *Bioinformatics*, 27(12):1739–1740, 2011.

[215] Dongdong Lin, Hongbao Cao, Vince D. Calhoun, and Yu-Ping Wang. Sparse models for correlative and integrative analysis of imaging and genetic data. *Journal of Neuroscience Methods*, 237:69–78, 2014.

[216] Chong Liu, Jonathan C Sage, Michael R Miller, Roel G W Verhaak, Simon Hippenmeyer, Hannes Vogel, Oded Foreman, et al. Mosaic analysis with double markers reveals tumor cell of origin in glioma. *Cell*, 146(2):209–21, July 2011.

[217] Jing Liu, Yu Jiang, Zechao Li, Zhi-hua Zhou, and Hanqing Lu. Partially Shared Latent Factor Learning With Multiview Data. *IEEE Transaction on Neural Networks and Learning Systems*, pages 1–14, 2014.

[218] TT Liu, AS Achrol, LA Mitchell, WA Du, JJ Loya, SA Rodriguez, A Feroze, EM Westbroek, KW Yeom, JM Stuart, et al. Computational identification of tumor anatomic location associated with survival in 2 large cohorts of human primary glioblastomas. *American Journal of Neuroradiology*, 2016.

[219] Y Liu, Z Xing, and Chao Deng. Automatically detecting lung nodules based on shape descriptor and semi-supervised learning. *Computer Application and ...*, (Iccasm):V1–647–V1–650, October 2010.

[220] Josep M Llovet and Jordi Bruix. Molecular targeted therapies in hepatocellular carcinoma. *Hepatology*, 48(4):1312–1327, 2008.

[221] Josep M Llovet and Virginia Hernandez-Gea. Hepatocellular carcinoma: reasons for phase iii failure and novel perspectives on trial design. *Clinical Cancer Research*, 20(8):2072–2079, 2014.

[222] Josep M Llovet, Sergio Ricci, Vincenzo Mazzaferro, Philip Hilgard, Edward Gane, Jean-Frédéric Blanc, Andre Cosme de Oliveira, Armando Santoro, Jean-Luc Raoul, Alejandro Forner, et al. Sorafenib in advanced hepatocellular carcinoma. *New England journal of medicine*, 359(4):378–390, 2008.

[223] John Lonsdale, Jeffrey Thomas, Mike Salvatore, Rebecca Phillips, Edmund Lo, Saboor Shad, Richard Hasz, Gary Walters, Fernando Garcia, Nancy Young, et al. The genotype-tissue expression (gtex) project. *Nature genetics*, 45(6):580–585, 2013.

[224] Riku Louhimo, Viljami Aittomäki, Ali Faisal, and Marko Laakso. Systematic use of computational methods allows stratifying treatment responders in glioblastoma multiforme. 2011.

[225] Jan Luts, Arend Heerschap, Johan a K Suykens, and Sabine Van Huffel. A com-

bined MRI and MRSI based multiclass system for brain tumour recognition using LS-SVMs with class probabilities and feature selection. *Artificial intelligence in medicine*, 40(2):87–102, June 2007.

[226] Lei Lv, Jian Xu, Shuo Zhao, Chunjing Chen, Xin Zhao, Shaohua Gu, Chaoneng Ji, Yi Xie, Lei Lv, Jian Xu, et al. Sequence analysis of a human rhogap domain-containing gene and characterization of its expression in human multiple tissues: Full length research paper. *DNA Sequence*, 18(3):184–189, 2007.

[227] Laura E MacConaill, Catarina D Campbell, Sarah M Kehoe, Adam J Bass, Charles Hatton, Lili Niu, Matt Davis, Keluo Yao, Megan Hanna, Chandrani Mondal, et al. Profiling critical cancer gene mutations in clinical tumor samples. *PloS one*, 4(11):e7887, 2009.

[228] Julien Mairal and Bin Yu. Supervised feature selection in graphs with path coding penalties and network flows. *Journal of Machine Learning Research*, 14(1):2449–2485, 2013.

[229] D Mantini, L Marzetti, M Corbetta, G L Romani, and C Del Gratta. Multimodal integration of fMRI and EEG data for high spatial and temporal resolution analysis of brain networks. *Brain topography*, 23(2):150–8, June 2010.

[230] Jean-Christophe Marine, Sarah Francoz, Marion Maetens, G Wahl, F Toledo, and G Lozano. Keeping p53 in check: essential and synergistic functions of mdm2 and mdm4. *Cell death and differentiation*, 13(6):927–934, 2006.

[231] Rafal T Marszalek. Cancer genomics just got personal. *Genome Biology*, 15(9):464, 2014.

[232] Elizabeth a Mason, Jessica C Mar, Andrew L Laslett, Martin F Pera, John Quackenbush, Ernst Wolvetang, and Christine a Wells. Gene expression variability as a unifying element of the pluripotency network. *Stem cell reports*, 3(2):365–77, August 2014.

[233] Volker Matys, Olga V Kel-Margoulis, Ellen Fricke, Ines Liebich, Sigrid Land, A Barre-Dirrie, Ingmar Reuter, D Chekmenev, Mathias Krull, Klaus Hornischer, et al. Transfac® and its module transcompel®: transcriptional gene regulation in eukaryotes. *Nucleic acids research*, 34(suppl 1):D108–D110, 2006.

[234] Andreas Mayr and Matthias Schmid. Boosting the concordance index for survival data–a unified framework to derive and evaluate biomarker combinations. *PloS one*, 9(1):e84483, January 2014.

[235] Mark L McCleland, Adam S Adler, Laura Deming, Ely Cosino, Leslie Lee, Elizabeth M Blackwood, Margaret Solon, Janet Tao, Li Li, David Shames, et al. Lactate dehydrogenase b is required for the growth of kras-dependent lung adenocarcinomas. *Clinical Cancer Research*, 19(4):773–784, 2013.

[236] Roland Memisevic. On multi-view feature learning. *arXiv preprint arXiv:1206.4609*, 2012.

[237] Bjoern H. Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Key-

van Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, et al. The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). *IEEE Transactions on Medical Imaging*, 34(10):1993–2024, 2015.

[238] Craig H Mermel, Steven E Schumacher, Barbara Hill, Matthew L Meyerson, Rameen Beroukhim, and Gad Getz. Gistic2. 0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome biology*, 12(4):1, 2011.

[239] Alexander G Miamen, Haidong Dong, and Lewis R Roberts. Immunotherapeutic approaches to hepatocellular carcinoma treatment. *Liver cancer*, 1(3-4):226–237, 2012.

[240] Kyriaki Michailidou, Per Hall, Anna Gonzalez-Neira, Maya Ghoussaini, Joe Dennis, Roger L Milne, Marjanka K Schmidt, Jenny Chang-Claude, Stig E Bojesen, Manjeet K Bolla, et al. Large-scale genotyping identifies 41 new loci associated with breast cancer risk. *Nature genetics*, 45(4):353–361, 2013.

[241] Tom M Mitchell. Machine Learning Today: When can Unlabeled Data Help Learn, 2011.

[242] Qianxing Mo, Sijian Wang, Venkatraman E Seshan, Adam B Olshen, Nikolaus Schultz, Chris Sander, R Scott Powers, Marc Ladanyi, and Ronglai Shen. Pattern discovery and cancer gene identification in integrated cancer genomic data.

Proceedings of the National Academy of Sciences of the United States of America, 110(11):4245–50, mar 2013.

[243] UB Mogensen, H Ishwaran, TA Gerds, and B Afdeling. *Evaluating random forests for survival analysis using prediction error curves*, volume 50. 2010.

[244] John Moult, Jan T Pedersen, Richard Judson, and Krzysztof Fidelis. A large-scale experiment to assess protein structure prediction methods. *Proteins: Structure, Function, and Bioinformatics*, 23(3), 1995.

[245] FJ Müller, LC Laurent, Dennis Kostka, and Igor Ulitsky. Regulatory networks define phenotypic classes of human stem cell lines. *Nature*, 455(7211):401–405, 2008.

[246] Ion Muslea, Steven Minton, and Craig A Knoblock. Active learning with multiple views. *Journal of Artificial Intelligence Research*, 27:203–233, 2006.

[247] Viswam S VS Nair, Olivier Gevaert, Guido Davidzon, Sandy Napel, Edward E Graves, Chuong D Hoang, et al. Prognostic PET 18F-FDG uptake imaging features are associated with major oncogenomic alterations in patients with resected non-small cell lung cancer. *Cancer research*, 72(15):3725–3734, August 2012.

[248] Kristopher L Nazor, Gulsah Altun, Candace Lynch, et al. Recurrent variations in DNA methylation in human pluripotent stem cells and their differentiated derivatives. *Cell stem cell*, 10(5):620–34, May 2012.

[249] Masatoshi Nei. The Neighbor-joining Method: A New Method for Reconstructing Phylogenetic Trees'. *Molecular Biology and Evolution*, 4(4):406–425, 1987.

[250] Yulia Newton, Adam Novak, Teresa Swatloski, Duncan C McColl, Sahil Chopra, Kiley Graim, Alana S Weinstein, Robert Baertsch, Sofie R Salama, et al. Ucsc tumor map: Exploring cancer signatures on an interactive dynamic landscape. *in review*, 2016.

[251] Sam Ng, Eric a Collisson, Artem Sokolov, Theodore Goldstein, Abel Gonzalez-Perez, Nuria Lopez-Bigas, Christopher Benz, David Haussler, and Joshua M Stuart. PARADIGM-SHIFT predicts the function of mutations in multiple cancers using pathway impact analysis. *Bioinformatics (Oxford, England)*, 28(18):i640–i646, September 2012.

[252] Naosuke Nonoguchi, Takashi Ohta, Ji-Eun Oh, Young-Ho Kim, Paul Kleihues, and Hiroko Ohgaki. TERT promoter mutations in primary and secondary glioblastomas. *Acta neuropathologica*, 126(6):931–7, December 2013.

[253] Kenneth a Norman, Sean M Polyn, Greg J Detre, and James V Haxby. Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends in cognitive sciences*, 10(9):424–30, September 2006.

[254] Houtan Noushmehr, Daniel J Weisenberger, Kristin Diefes, Heidi S Phillips, et al. Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma. *Cancer cell*, 17(5):510–22, May 2010.

[255] Fabio Parisi, Francesco Strino, Boaz Nadler, and Yuval Kluger. Ranking and combining multiple predictors without labeled data. *Proceedings of the National Academy of Sciences of the United States of America*, 111(4):1253–8, January 2014.

[256] John K Park, Tiffany Hodges, Leopold Arko, Michael Shen, Donna Dello Iacono, Adrian McNabb, Nancy Olsen Bailey, et al. Scale to predict survival after surgery for recurrent glioblastoma multiforme. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, 28(24):3838–43, August 2010.

[257] D Williams Parsons, Siân Jones, Xiaosong Zhang, Jimmy Cheng-Ho Lin, et al. An integrated genomic analysis of human glioblastoma multiforme. *Science (New York, N.Y.)*, 321(5897):1807–12, September 2008.

[258] Harvey I Pass, Nicholas Vogelzang, Steven Hahn, and Michele Carbone. Malignant pleural mesothelioma. *Current problems in cancer*, 28(3):93–174, 2004.

[259] Anoop P Patel, Itay Tirosh, John J Trombetta, Alex K Shalek, et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science (New York, N.Y.)*, (June):1–9, June 2014.

[260] Vishal N Patel, Giridharan Gokulrangan, Salim a Chowdhury, Yanwen Chen, Andrew E Sloan, Mehmet Koyutürk, Jill Barnholtz-Sloan, and Mark R Chance. Network signatures of survival in glioblastoma multiforme. *PLoS computational biology*, 9(9):e1003237, January 2013.

[261] Evan O Paull, Daniel E Carlin, Mario Niepel, Peter K Sorger, David Haussler, and Joshua M Stuart. Discovering causal pathways linking genomic events to transcriptional states using tied diffusion through interacting events (tiedie). *Bioinformatics*, 29(21):2757–2764, 2013.

[262] Francisco Pereira, Tom Mitchell, and Matthew Botvinick. Machine learning classifiers and fMRI: a tutorial overview. *NeuroImage*, 45(1 Suppl):S199–209, March 2009.

[263] Chariklia Petropoulou, Panorea Kotantaki, Dimitris Karamitros, and Stavros Taraviras. Cdt1 and geminin in cancer: markers or triggers of malignant transformation? *Frontiers in bioscience: a journal and virtual library*, 13:4485–4494, 2007.

[264] Floriane Pez, Anaïs Lopez, Miran Kim, Jack R Wands, Claude Caron de Fromentel, and Philippe Merle. Wnt signaling and hepatocarcinogenesis: molecular targets for the development of innovative anticancer drugs. *Journal of hepatology*, 59(5):1107–1117, 2013.

[265] Yongjun Piao, Minghao Piao, Kiejung Park, and Keun Ho Ryu. An ensemble correlation-based gene selection algorithm for cancer classification with gene expression data. *Bioinformatics (Oxford, England)*, 28(24):3306–15, December 2012.

[266] Gemma Piella. A general framework for multiresolution image fusion: from pixels to regions. *Information Fusion*, 4(4):259–280, December 2003.

[267] Paulo S Pinheiro, Melanie Williams, Eric a Miller, Stephanie Easterday, Sheniz Moonie, and Edward J Trapido. Cancer survival among Latinos and the Hispanic Paradox. *Cancer causes & control : CCC*, 22(4):553–61, April 2011.

[268] Kornelia Polyak and Robert A Weinberg. Transitions between epithelial and mesenchymal states: acquisition of malignant and stem cell traits. *Nature Reviews Cancer*, 9(4):265–273, 2009.

[269] Thomas Powles, Joseph Paul Eder, Gregg D Fine, Fadi S Braiteh, Yohann Loriot, Cristina Cruz, Joaquim Bellmunt, Howard A Burris, Daniel P Petrylak, Siew-leng Teng, et al. Mpdl3280a (anti-pd-l1) treatment leads to clinical activity in metastatic bladder cancer. *Nature*, 515(7528):558–562, 2014.

[270] Buyue Qian, Xiang Wang, and Ian Davidson. A Reconstruction Error Formulation for Semi-Supervised Multi-task and Multi-view Learning. *arXiv preprint arXiv:1202.0855*, 2, 2012.

[271] Peng Qiu, Andrew J Gentles, and Sylvia K Plevritis. Discovering biological progression underlying microarray samples. *PLoS computational biology*, 7(4):e1001123, April 2011.

[272] Novi Quadrianto and CH Lampert. Learning multi-view neighborhood preserving projections. *Proceedings of the ...*, 2011.

[273] MH Quang, L Bazzani, and Vittorio Murino. A unifying framework for vector-

valued manifold regularization and multi-view learning. *Proceedings of the ...*, 28, 2013.

[274] Brian Quanz. *Learning with Low-Quality Data: Multi-View Semi-Supervised Learning with Missing Views*. PhD thesis, 2012.

[275] Predrag Radivojac, Wyatt T Clark, Tal Ronnen Oron, Alexandra M Schnoes, Tobias Wittkop, Artem Sokolov, Kiley Graim, Christopher Funk, Karin Verspoor, et al. A large-scale evaluation of computational protein function prediction. *Nature methods*, 10(3):221–7, March 2013.

[276] P Rai, Anusua Trivedi, H Daumé III, and SL DuVall. Multiview Clustering with Incomplete Views. *NIPS Workshop on Machine ...*, pages 1–7, 2010.

[277] Arvind Rao and Tcga Phenotype Group. Exploring relationships between multivariate radiological phenotypes and genetic features : A case- study in Glioblastoma using the Cancer Genome Atlas. *IEEE*, pages 69–72, 2013.

[278] Mathias Rask-Andersen, Surendar Masuram, and Helgi B Schiöth. The druggable genome: evaluation of drug targets in clinical trials suggests major shifts in molecular class and indication. *Annual review of pharmacology and toxicology*, 54:9–26, 2014.

[279] Lior Rokach. Ensemble-based classifiers. *Artificial Intelligence Review*, 33(1-2):1–39, 2010.

[280] Nicolas P. Rougier, Michael Droettboom, and Philip E. Bourne. Ten Simple Rules for Better Figures. *PLoS Computational Biology*, 10(9):e1003833, September 2014.

[281] Maria Ruden and Neelu Puri. Novel anticancer therapeutics targeting telomerase. *Cancer treatment reviews*, 39(5):444–456, 2013.

[282] S Rüping and Tobias Scheffer. Learning with multiple views. In Stefan Rüping and Tobias Scheffer, editors, *Learning with Multiple Views*, Bonn, Germany, 2005. ICML.

[283] Jan Rupnik and J Shawe-Taylor. Multi-view canonical correlation analysis. *Conference on Data Mining and Data ...*, (1):2–5, 2010.

[284] Virginia R. Sa, Patrick W. Gallagher, Joshua M. Lewis, and Vicente L. Malave. Multi-view kernel construction. *Machine Learning*, 79(1-2):47–71, November 2009.

[285] Moumita Saha. A Graph Based Approach to Multiview. *Springer-Verlag*, pages 128–133, 2013.

[286] M. M. F. Savitski, F. B. M. Reinhard, H. Franken, T. Werner, D. Eberhard, D. M. Molina, R. Jafari, R. B. Dovega, et al. Tracking cancer drugs in living cells by thermal profiling of the proteome. *Science*, 346(6205):1255784–1255784, October 2014.

[287] Eric E Schadt. Molecular networks as sensors and drivers of common human diseases. *Nature*, 461(7261):218–23, September 2009.

[288] Matthias Schmid, Hans a Kestler, and Sergej Potapov. On the validity of time-dependent AUC estimators. *Briefings in bioinformatics*, September 2013.

[289] J. Schrouff, J. Cremers, G. Garraux, L. Baldassarre, J. Mourao-Miranda, and C. Phillips. Localizing and Comparing Weight Maps Generated from Linear Kernel Machine Learning Models. *2013 International Workshop on Pattern Recognition in Neuroimaging*, pages 124–127, June 2013.

[290] Friedhelm Schwenker and Edmondo Trentin. Pattern classification and clustering: A review of partially supervised learning approaches. *Pattern Recognition Letters*, 37(1):4–14, 2014.

[291] José a Seoane, Ian N M Day, Tom R Gaunt, and Colin Campbell. A pathway-based data integration framework for prediction of disease progression. *Bioinformatics (Oxford, England)*, 30(6):838–45, March 2014.

[292] B Settles. Multimodal Deep Learning. *Proceedings of the 28th International Conference on Machine Learning*, 2011.

[293] Burr Settles. Active learning literature survey. *University of Wisconsin, Madison*, 2010.

[294] Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S Baliga, Jonathan T Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, 13(11):2498–2504, 2003.

[295] a. Sharma, a. Kumar, H. Daume, and D. W. Jacobs. Generalized Multiview Analysis: A discriminative latent space. *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2160–2167, June 2012.

[296] Kerby Shedden, Jeremy M G Taylor, Steven a Enkemann, Ming-Sound Tsao, et al. Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. *Nature medicine*, 14(8):822–7, August 2008.

[297] Nathan C Sheffield, Robert E Thurman, Lingyun Song, Alexias Safi, John a Stamatoyannopoulos, Boris Lenhard, Gregory E Crawford, and Terrence S Furey. Patterns of regulatory activity across diverse human cell types predict tissue identity, transcription factor binding, and long-range interactions. *Genome research*, 23(5):777–88, May 2013.

[298] 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.

[299] Noam Shental, Assaf Zomet, Tomer Hertz, and Yair Weiss. Pairwise Clustering and Graphical Models. *Advances in Neural Information Processing Systems*, page 8, 2003.

[300] Liran I Shlush, Sasan Zandi, Amanda Mitchell, Weihsu Claire Chen, Joseph M

Brandwein, et al. Identification of pre-leukaemic haematopoietic stem cells in acute leukaemia. *Nature*, 506(7488):328–333, February 2014.

[301] Rebecca L Siegel, Kimberly D Miller, and Ahmedin Jemal. Cancer statistics, 2016. *CA: a cancer journal for clinicians*, 66(1):7–30, 2016.

[302] N Simon, J Friedman, and Trevor Hastie. A blockwise descent algorithm for group-penalized multiresponse and multinomial regression. *arXiv preprint arXiv:1311.6529*, 2013.

[303] JR Simpson, J Horton, and C Scott. extent of surgical resection on survival of patients with glioblastoma multiforme: results of three consecutive Radiation Therapy Oncology Group (RTOG) clinical trials. *... of Radiation Oncology* ...*, (December 1992):239–244, 1993.

[304] JR Simpson, J Horton, C Scott, WJ Curran, P Rubin, J Fischbach, S Isaacson, M Rotman, SO Asbell, JS Nelson, et al. Influence of location and extent of surgical resection on survival of patients with glioblastoma multiforme: results of three consecutive radiation therapy oncology group (rtog) clinical trials. *International Journal of Radiation Oncology* Biology* Physics*, 26(2):239–244, 1993.

[305] Michael E Smoot, Keiichiro Ono, Johannes Ruscheinski, Peng-Liang Wang, and Trey Ideker. Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics*, 27(3):431–432, 2011.

[306] American Cancer Society. Cancer facts & figures 2014. *Atlanta: American Cancer Society*, 2014.

[307] Artem Sokolov, Christopher Funk, Kiley Graim, Karin Verspoor, and Asa Ben-Hur. Combining heterogeneous data sources for accurate functional annotation of proteins. *BMC bioinformatics*, 14 Suppl 3(Suppl 3):S10, January 2013.

[308] Lingyun Song, Zhancheng Zhang, Linda L Grasfeder, Alan P Boyle, Paul G Giresi, Bum-Kyu Lee, Nathan C Sheffield, Stefan Gräf, et al. Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. *Genome research*, 21(10):1757–67, October 2011.

[309] Xiaomu Song and Alice M Wyrwicz. Unsupervised spatiotemporal fMRI data analysis using support vector machines. *NeuroImage*, 47(1):204–12, August 2009.

[310] Thierry Soussi. The tp53 gene network in a postgenomic era. *Human mutation*, 35(6):641–642, 2014.

[311] Karen Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21, 1972.

[312] Nora K Speicher and Nico Pfeifer. Integrating different data types by regularized unsupervised multiple kernel learning with application to cancer subtype discovery. In *Bioinformatics*, volume 31, pages i268–i275, 2015.

[313] Karthik Sridharan and SM Kakade. An Information Theoretic Framework for Multi-view Learning. *COLT*, 2008.

[314] Andreas M Stark, Julia van de Bergh, Jürgen Hedderich, H Maximilian Mehdorn, and Arya Nabavi. Glioblastoma: clinical characteristics, prognostic factors and survival in 492 patients. *Clinical neurology and neurosurgery*, 114(7):840–5, September 2012.

[315] Cynthia M Stonnington, Carlton Chu, Stefan Klöppel, Clifford R Jack, John Ashburner, and Richard S J Frackowiak. Predicting clinical scores from magnetic resonance scans in Alzheimer's disease. *NeuroImage*, 51(4):1405–13, July 2010.

[316] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545–15550, 2005.

[317] Jing Sui, Tülay Adali, Godfrey D Pearlson, and Vince D Calhoun. An ICA-based method for the identification of optimal FMRI features and components using combined group-discriminative techniques. *NeuroImage*, 46(1):73–86, May 2009.

[318] Jing Sui, Godfrey Pearlson, Arvind Caprihan, Tülay Adali, Kent A Kiehl, Jingyu Liu, Jeremy Yamamoto, and Vince D Calhoun. Discriminating schizophrenia and bipolar disorder by fusing fMRI and DTI in a multimodal CCA+ joint ICA model. *NeuroImage*, 57(3):839–55, August 2011.

[319] Kumar Sukhdeo, Dolores Hambardzumyan, and Jeremy N Rich. Glioma development: where did it all go wrong? *Cell*, 146(2):187–8, July 2011.

[320] Pavel Sumazin, Xuerui Yang, Hua-Sheng Chiu, Wei-Jen Chung, Archana Iyer, David Llobet-Navas, Presha Rajbhandari, et al. An extensive microRNA-mediated network of RNA-RNA interactions regulates established oncogenic pathways in glioblastoma. *Cell*, 147(2):370–81, October 2011.

[321] Shiliang Sun. A survey of multi-view machine learning. *Neural Computing and Applications*, 23(7-8):2031–2038, February 2013.

[322] Shiliang Sun and Feng Jin. Robust co-training. *International Journal of Pattern Recognition and Artificial Intelligence*, 25(07):1113–1126, 2011.

[323] Marie Szafranski, Yves Grandvalet, and Alain Rakotomamonjy. Composite kernel learning. *Machine learning*, 79(1-2):73–103, 2010.

[324] E. Tabouret, M. Labussi??re, A. Alentorn, Y. Schmitt, Y. Marie, and M. Sanson. LRP1B deletion is associated with poor outcome for glioblastoma patients. *Journal of the Neurological Sciences*, 358(1-2):440–443, 2015.

[325] David Tamborero, Abel Gonzalez-Perez, Christian Perez-Llamas, Jordi Deu-Pons, Cyriac Kandoth, Jüri Reimand, Michael S Lawrence, Gad Getz, Gary D Bader, Li Ding, et al. Comprehensive identification of mutational cancer driver genes across 12 tumor types. *Scientific reports*, 3:2650, 2013.

[326] Chad Tang, Xiaohong Wang, Hendrick Soh, Steven Seyedin, Maria Angelica

Cortez, Sunil Krishnan, et al. Combining Radiation and Immunotherapy: A New Systemic Therapy for Solid Tumors? *Cancer Immunology Research*, 2(9):831–838, sep 2014.

[327] Yangzhong Tang, Tiao Xie, Stefan Florian, Nathan Moerke, Caroline Shamu, Cyril Benes, and Timothy J Mitchison. Differential determinants of cancer cell insensitivity to antimitotic drugs discriminated by a one-step cell imaging assay. *Journal of biomolecular screening*, 18(9):1062–71, October 2013.

[328] Ben Taskar, Vassil Chatalbashev, Daphne Koller, and Carlos Guestrin. Learning structured prediction models: A large margin approach. In *Proceedings of the 22nd international conference on Machine learning*, pages 896–903. ACM, 2005.

[329] Barry S Taylor, Nikolaus Schultz, Haley Hieronymus, Anuradha Gopalan, Yonghong Xiao, Brett S Carver, Vivek K Arora, Poorvi Kaushik, Ethan Cerami, Boris Reva, et al. Integrative genomic profiling of human prostate cancer. *Cancer cell*, 18(1):11–22, 2010.

[330] Paul D Thomas, Valerie Wood, Christopher J Mungall, Suzanna E Lewis, and Judith a Blake. On the Use of Gene Ontology Annotations to Assess Functional Similarity among Orthologs and Paralogs: A Short Report. *PLoS computational biology*, 8(2):e1002386, January 2012.

[331] B Thompson. Canonical correlation analysis. *Encyclopedia of statistics in behavioral science*, 25:81–99, 2005.

[332] Ze Tian, TaeHyun Hwang, and Rui Kuang. A hypergraph-based learning algorithm for classifying gene expression and arraycgh data with prior knowledge. *Bioinformatics*, 25(21):2831–2838, 2009.

[333] Scott A. Tomlins, Rohit Mehra, Daniel R. Rhodes, Lisa R. Smith, Diane Roulston, Beth E. Helgeson, Xuhong Cao, John T. Wei, Mark A. Rubin, Rajal B. Shah, and Arul M. Chinnaiyan. TMPRSS2:ETV4 gene fusions define a third molecular subtype of prostate cancer. *Cancer Research*, 66(7):3396–3400, 2006.

[334] Elfar Torarinsson, Jakob H Havgaard, and Jan Gorodkin. Multiple structural alignment and clustering of rna sequences. *Bioinformatics*, 23(8):926–932, 2007.

[335] Anh N Tran, Albert Lai, Sichen Li, Whitney B Pope, Stephanie Teixeira, Robert J Harris, Davis C Woodworth, Phioanh L Nghiemphu, Timothy F Cloughesy, and Benjamin M Ellingson. Increased sensitivity to radiochemotherapy in idh1 mutant glioblastoma as demonstrated by serial quantitative mr volumetry. *Neuro-oncology*, page not198, 2013.

[336] A Tripathi, A Klami, and S Kaski. Using dependencies to pair samples for multi-view learning. *Acoustics, Speech and Signal . . .*, 2009.

[337] Evanthia E Tripoliti, Dimitrios I Fotiadis, Maria Argyropoulou, and George Manis. A six stage approach for the diagnosis of the Alzheimer's disease based on fMRI data. *Journal of biomedical informatics*, 43(2):307–20, April 2010.

[338] Ioannis Tsochantaridis. Large margin methods for structured and interdependent output variables. *Journal of Machine ...*, 6:1453–1484, 2005.

[339] Paul C Tumeh, Christina L Harview, Jennifer H Yearley, I Peter Shintaku, Emma JM Taylor, Lidia Robert, Bartosz Chmielowski, Marko Spasic, Gina Henry, Voicu Ciobanu, et al. Pd-1 blockade induces responses by inhibiting adaptive immune resistance. *Nature*, 515(7528):568–571, 2014.

[340] Sevin Turcan, Daniel Rohle, Anuj Goenka, Logan a Walsh, et al. IDH1 mutation is sufficient to establish the glioma hypermethylator phenotype. *Nature*, 483(7390):479–83, March 2012.

[341] Cagatay Turkay, A Lex, and Marc Streit. Characterizing Cancer Subtypes Using Dual Analysis in Caleydo StratomeX. *... computer graphics and ...*, 2014.

[342] Brandon M Turner, Birte U Forstmann, Eric-Jan Wagenmakers, Scott D Brown, Per B Sederberg, and Mark Steyvers. A Bayesian framework for simultaneously modeling neural and behavioral data. *NeuroImage*, 72:193–206, May 2013.

[343] Virginia Goss Tusher, Robert Tibshirani, and Gilbert Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences*, 98(9):5116–5121, 2001.

[344] Hajime Uno, Tianxi Cai, Michael J Pencina, Ralph B D'Agostino, and L J Wei. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in medicine*, 30(10):1105–17, May 2011.

[345] Sooryanarayana Varambally, Jianjun Yu, Bharathi Laxman, Daniel R Rhodes, Rohit Mehra, Scott A Tomlins, Rajal B Shah, Uma Chandran, Federico A Monzon, Michael J Becich, et al. Integrative genomic and proteomic analysis of prostate cancer reveals signatures of metastatic progression. *Cancer cell*, 8(5):393–406, 2005.

[346] G Varoquaux, S Sadaghiani, P Pinel, A Kleinschmidt, J B Poline, and B Thirion. A group model for stable multi-subject ICA on fMRI datasets. *NeuroImage*, 51(1):288–99, May 2010.

[347] Charles J Vaske, Stephen C Benz, J Zachary Sanborn, Dent Earl, Christopher Szeto, Jingchun Zhu, David Haussler, and Joshua M Stuart. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics (Oxford, England)*, 26(12):i237–45, June 2010.

[348] Lyubomir T Vassilev, Binh T Vu, Bradford Graves, Daisy Carvajal, Frank Podlaski, Zoran Filipovic, Norman Kong, Ursula Kammlott, Christine Lukacs, Christian Klein, et al. In vivo activation of the p53 pathway by small-molecule antagonists of mdm2. *Science*, 303(5659):844–848, 2004.

[349] Marina Velikova and PJF Lucas. A decision support system for breast cancer detection in screening programs. *ECAI*, pages 658–662, 2008.

[350] David Venet, Jacques E Dumont, and Vincent Detours. Most random gene ex-

pression signatures are significantly associated with breast cancer outcome. *PLoS computational biology*, 7(10):e1002240, October 2011.

[351] Roel G W Verhaak, Katherine a Hoadley, Elizabeth Purdom, Victoria Wang, et al. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer cell*, 17(1):98–110, January 2010.

[352] Rogier Versteeg. Cancer: Tumours outside the mutation box. *Nature*, 506(7489):438–9, February 2014.

[353] João Vinagre, Ana Almeida, Helena Pópulo, Rui Batista, et al. Frequency of TERT promoter mutations in human cancers. *Nature communications*, 4:2185, January 2013.

[354] Seppo Virtanen, Arto Klami, Suleiman A Khan, and Samuel Kaski. Bayesian group factor analysis. In *AISTATS*, pages 1269–1277, 2012.

[355] Jane E Visvader. Cells of origin in cancer. *Nature*, 469(7330):314–22, January 2011.

[356] Antonia Vlahou, John O. Schorge, Betsy W. Gregory, and Robert L. Coleman. Diagnosis of Ovarian Cancer Using Decision Tree Classification of Mass Spectral Data. *Journal of biomedicine & biotechnology*, 2003(5):308–314, January 2003.

[357] Pamela Volkel, Barbara Dupret, Xuefen Le Bourhis, and Pierre-Olivier Angrand.

Diverse involvement of ezh2 in cancer epigenetics. *Am J Transl Res*, 7(2):175–193, 2015.

[358] Hua Wang, Feiping Nie, and Heng Huang. Multi-view clustering and feature learning via structured sparsity. In *ICML (3)*, pages 352–360, 2013.

[359] Jian-Jiang Wang, Yue-Xin Liu, Wei Wang, Wei Yan, Yu-Peng Zheng, Lu-Dong Qiao, Dan Liu, and Shan Chen. Fusion between tmprss2 and ets family members (erg, etv1, etv4) in prostate cancers from northern china. *Asian Pacific journal of cancer prevention: APJCP*, 13(10):4935–4938, 2011.

[360] W Wang, R Arora, K Livescu, and J Bilmes. On deep multi-view representation learning. *32st Int. Conf. Machine Learning (. . .*, 37, 2015.

[361] Wei Wang and Zhi-Hua Zhou. Analyzing co-training style algorithms. In *European Conference on Machine Learning*, pages 454–465. Springer, 2007.

[362] Wei Wang and Zhi-Hua Zhou. A New Analysis of Co-Training. *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 1135–1142, 2010.

[363] Wei Wang and Zhi-Hua Zhou. A new analysis of co-training. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 1135–1142, 2010.

[364] Wenting Wang, Veerabhadran Baladandayuthapani, Jeffrey S Morris, Bradley M Broom, Ganiraju Manyam, and Kim-Anh Do. iBAG: integrative Bayesian anal-

ysis of high-dimensional multiplatform genomics data. *Bioinformatics (Oxford, England)*, 29(2):149–59, January 2013.

[365] YY Wang, T Zhang, SW Li, TY Qian, X Fan, XX Peng, J Ma, L Wang, and T Jiang. Mapping p53 mutations in low-grade glioma: a voxel-based neuroimaging analysis. *American Journal of Neuroradiology*, 36(1):70–76, 2015.

[366] YY Wang, T Zhang, SW Li, TY Qian, X Fan, XX Peng, J Ma, L Wang, and T Jiang. Mapping p53 mutations in low-grade glioma: a voxel-based neuroimaging analysis. *American Journal of Neuroradiology*, 36(1):70–76, 2015.

[367] Zhe Wang, Songcan Chen, and Tingkai Sun. MultiK-MHKS: a novel multiple kernel learning algorithm. *IEEE transactions on pattern analysis and machine intelligence*, 30(2):348–53, February 2008.

[368] David Warde-Farley, Sylva L Donaldson, Ovi Comes, Khalid Zuberi, Rashad Badrawi, Pauline Chao, Max Franz, Chris Grouios, Farzana Kazi, Christian Tannus Lopes, et al. The genemania prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic acids research*, 38(suppl 2):W214–W220, 2010.

[369] Joshua D Webster, Eleanor R Simpson, Aleksandra M Michalowski, Shelley B Hoover, and R Mark Simpson. Quantifying histological features of cancer biospecimens for biobanking quality assurance using automated morphometric pattern

recognition image analysis algorithms. *Journal of biomolecular techniques : JBT*, 22(3):108–18, September 2011.

[370] F Wei, J Yan, and D Tang. Extracellular signal-regulated kinases modulate dna damage response-a contributing factor to using mek inhibitors in cancer therapy. *Current medicinal chemistry*, 18(35):5476–5482, 2011.

[371] Martha White and Xinhua Zhang. Convex multi-view subspace learning. *Advances in Neural . . .*, pages 1–14, 2012.

[372] Matthew D Wilkerson and D Neil Hayes. Consensusclusterplus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics*, 26(12):1572–1573, 2010.

[373] Daniela M Witten, Robert Tibshirani, and Trevor Hastie. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics (Oxford, England)*, 10(3):515–34, July 2009.

[374] Stefan Wuchty, Alice Zhang, Jennifer Walling, Susie Ahn, Aiguo Li, Martha Quezado, Carl Oberholtzer, Jean-Claude Zenklusen, and Howard a Fine. Gene pathways and subnetworks distinguish between major glioma subtypes and elucidate potential underlying biology. *Journal of biomedical informatics*, 43(6):945–52, December 2010.

[375] Xiaohui Xie, Jun Lu, EJ Kulbokas, Todd R Golub, Vamsi Mootha, Kerstin Lindblad-Toh, Eric S Lander, and Manolis Kellis. Systematic discovery of regu-

latory motifs in human promoters and 3 utrs by comparison of several mammals. *Nature*, 434(7031):338–345, 2005.

[376] Chang Xu, Dacheng Tao, and Chao Xu. A survey on multi-view learning. *arXiv preprint arXiv:1304.5634*, 2013.

[377] Chen Xu and Zhengchang Su. Identification of cell types from single-cell transcriptomes using a novel clustering method. *Bioinformatics*, page btv088, 2015.

[378] Jiaming Xu, R Wu, Kai Zhu, Bruce Hajek, R Srikant, and L Ying. Jointly Clustering Rows and Columns of Binary Matrices: Algorithms and Trade-offs. *Urbana*, 2013.

[379] Zenglin Xu, R Jin, and Haiqin Yang. Simple and efficient multiple kernel learning by group lasso. *Proceedings of the ...*, 2010.

[380] Honghui Yang, Jingyu Liu, Jing Sui, Godfrey Pearlson, and Vince D Calhoun. A Hybrid Machine Learning Method for Fusing fMRI and Genetic Data: Combining both Improves Classification of Schizophrenia. *Frontiers in human neuroscience*, 4(October):192, January 2010.

[381] Ju Dong Yang and Lewis R Roberts. Hepatocellular carcinoma: a global view. *Nature Reviews Gastroenterology and Hepatology*, 7(8):448–458, 2010.

[382] Yanan Yang, Marie Wislez, Nobukazu Fujimoto, Ludmila Prudkin, Julie G Izzo, Futoshi Uno, Lin Ji, Amy E Hanna, Robert R Langley, Diane Liu, et al. A selec-

tive small molecule inhibitor of c-met, pha-665752, reverses lung premalignancy induced by mutant k-ras. *Molecular cancer therapeutics*, 7(4):952–960, 2008.

[383] Jieping Ye, Teresa Wu, Jing Li, and K Chen. Machine learning approaches for the neuroimaging study of Alzheimer's disease. *IEEE*, (April):99–101, 2011.

[384] Jarkko Ylipaavalniemi and Ricardo Vigário. Analyzing consistency of independent components: an fMRI illustration. *NeuroImage*, 39(1):169–80, January 2008.

[385] Matthew D Young, Matthew J Wakefield, Gordon K Smyth, and Alicia Oshlack. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome biology*, 11(2):R14, January 2010.

[386] H A N Yu and Rachael Hageman Blair. A framework for attribute-based community detection with applications to integrated functional genomics. pages 69–80, 2016.

[387] Jun Yu, Yong Rui, and Yuan Yan Tang. High-Order Distance-Based Multiview Stochastic Learning in Image Classification. *IEEE Transactions on Cybernetics*, 44(12):2431–2442, 2014.

[388] Jun Yu, Yong Rui, Yuan Yan Tang, and Dacheng Tao. High-order distance-based multiview stochastic learning in image classification. *IEEE transactions on cybernetics*, 44(12):2431–2442, 2014.

[389] Jun Yu, Meng Wang, and Dacheng Tao. Semisupervised multiview distance metric

learning for cartoon synthesis. *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society*, 21(11):4636–48, November 2012.

[390] Shipeng Yu, Balaji Krishnapuram, Romer Rosales, and R Bharat Rao. Active Sensing. *AISTATS*, pages 639–646, 2009.

[391] Shipeng Yu, Balaji Krishnapuram, Romer Rosales, and RB Bharat Rao. Bayesian co-training. *The Journal of Machine ...*, 12(Ml):2649–2680, 2011.

[392] Lei Yuan, Yalin Wang, Paul M Thompson, Vaibhav a Narayan, and Jieping Ye. Multi-source feature learning for joint analysis of incomplete multiple heterogeneous neuroimaging data. *NeuroImage*, 61(3):622–32, July 2012.

[393] Majd Zayzafoon, Sarki A Abdulkadir, and Jay M McDonald. Notch signaling and erk activation are important for the osteomimetic properties of prostate cancer bone metastatic cell lines. *Journal of Biological Chemistry*, 279(5):3662–3670, 2004.

[394] Tao Zeng, Hanbo Chen, Ahmed Fakhry, Xiaoping Hu, Tianming Liu, and Shuiwang Ji. Allen mouse brain atlases reveal different neural connection and gene expression patterns in cerebellum gyri and sulci. *Brain structure & function*, June 2014.

[395] Xiaojun Zha, Fang Wang, Ying Wang, Shaozong He, Yanling Jing, Xueyan Wu, and Hongbing Zhang. Lactate dehydrogenase b is critical for hyperactive mtor-mediated tumorigenesis. *Cancer research*, 71(1):13–18, 2011.

[396] Xiaolei Zhai, Qianhe Han, Zhongjie Shan, Xiaowei Qu, Liang Guo, and Yudong Zhou. Dual specificity phosphatase 6 suppresses the growth and metastasis of prostate cancer cells. *Molecular medicine reports*, 10(6):3052–3058, 2014.

[397] He Zhang, M Gönen, Z Yang, and E Oja. Predicting Emotional States of Images Using Bayesian Multiple Kernel Learning. *Neural Information Processing*, pages 274–282, 2013.

[398] Jing Zhang, Daniel P. Barboriak, Hasan Hobbs, and Maciej a. Mazurowski. A fully automatic extraction of magnetic resonance image features in glioblastoma patients. *Medical Physics*, 41(4):042301, April 2014.

[399] Naiqian Zhang, Haiyun Wang, Yun Fang, Jun Wang, Xiaoqi Zheng, and X. Shirley Liu. Predicting Anticancer Drug Responses Using a Dual-Layer Integrated Cell Line-Drug Network Model. *PLoS Computational Biology*, 11(9):1–18, 2015.

[400] Tan Zhang, Yinyan Wang, Xing Fan, Jun Ma, Shaowu Li, Tao Jiang, and Lei Wang. Anatomical localization of p53 mutated tumors: A radiographic study of human glioblastomas. *Journal of the neurological sciences*, 346(1):94–98, 2014.

[401] Tan Zhang, Yinyan Wang, Xing Fan, Jun Ma, Shaowu Li, Tao Jiang, and Lei Wang. Anatomical localization of p53 mutated tumors: A radiographic study of human glioblastomas. *Journal of the neurological sciences*, August 2014.

[402] W Zhang, Ke Zhang, Pan Gu, and Xiangyang Xue. Multi-view embedding learning

for incompletely labeled data. *... of the Twenty-Third international joint ...*, pages 1910–1916, 2013.

[403] Yudong Zhang, Zhengchao Dong, Lenan Wu, and Shuihua Wang. A hybrid method for MRI brain image classification. *Expert Systems with Applications*, 38(8):10049–10053, August 2011.

[404] Yanchang Zhao. *R and Data Mining: Examples and Case Studies*. Number December 2012. 2012.

[405] S Zhe, Z Xu, and Y Qi. Supervised Heterogeneous Multiview Learning for Joint Association Study and Disease Diagnosis. *arXiv preprint arXiv:1304.7284*, pages 1–19, 2013.

[406] Shandian Zhe, Zenglin Xu, Yuan Qi, and Peng Yu. Joint association discovery and diagnosis of Alzheimer's disease by supervised heterogeneous multiview learning. *Pacific Symposium on Biocomputing. ...*, pages 300–311, 2013.

[407] Dengyong Zhou and Christopher J. C. Burges. Spectral clustering and transductive learning with multiple views. *Proceedings of the 24th international conference on Machine learning - ICML '07*, pages 1159–1166, 2007.

[408] Zhenyu Zhou, Mingzhou Ding, Yonghong Chen, Paul Wright, Zuhong Lu, and Yijun Liu. Detecting directional influence in fMRI connectivity analysis using PCA based Granger causality. *Brain research*, 1289:22–9, September 2009.

[409] Zhi-Hua Zhou and Ming Li. Semi-supervised learning by disagreement. *Knowledge and Information Systems*, 24(3):415–439, 2010.

[410] Feng Zhu, Zhe Shi, Chu Qin, Lin Tao, Xin Liu, Feng Xu, Li Zhang, Yang Song, Xianghui Liu, Jingxian Zhang, et al. Therapeutic target database update 2012: a resource for facilitating target-oriented drug discovery. *Nucleic acids research*, page gkr797, 2011.

[411] Wenge Zhu and Melvin L DePamphilis. Selective killing of cancer cells by suppression of geminin activity. *Cancer research*, 69(11):4870–4877, 2009.

[412] Xiaojin Zhu. *Semi-supervised learning literature survey.* PhD thesis, University of Wisconsin– Madison, 2005.

[413] Fuzhen Zhuang, George Karypis, Xia Ning, Qing He, and Zhongzhi Shi. Multi-view learning via probabilistic latent semantic analysis. *Information Sciences*, 199:20–30, September 2012.

[414] a. Zibakhsh and M. Saniee Abadeh. Gene selection for cancer tumor detection using a novel memetic algorithm with a multi-view fitness function. *Engineering Applications of Artificial Intelligence*, 26(4):1274–1281, April 2013.

[415] Pascal O Zinn, Bhanu Mahajan, Bhanu Majadan, Pratheesh Sathyan, Sanjay K Singh, Sadhan Majumder, Ferenc a Jolesz, and Rivka R Colen. Radiogenomic mapping of edema/cellular invasion MRI-phenotypes in glioblastoma multiforme. *PloS one*, 6(10):e25451, January 2011.

[416] 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, April 2005.

Appendix A

Some Ancillary Stuff

A.1 Supplement for: HOCUS: Higher-Order Correlations to Uncover Subtypes

A.1.1 Processing the Data

HOCUS can also be improved by filtering the data. We explore filtering MR image data. Mutations and copy number data may benefit from filtering, for example excluding patients with more than 200 mutations as is often done in TCGA analysis. Removing singleton gene mutations would also be beneficial, since a friend-less gene provides little insight as to the mutational network structure. Even without filtering, HOCUS applies well to mutation data as well as MR image data.

MR images required high levels of filtering (Fig. A.7) to mask out the non-brain regions of the images since brains have slightly different shapes and also because tumor appears in only white tissue, not grey. Similarly, many genes are mutated in only

1 sample in a cohort, which makes them difficult to use as a social connection; they are essentially friendless and can never contribute to building similarities between pairs of patients. Eliminating them would reduce similarity score inflation.

A.1.2 Imaging

A.1.2.1 MR Image Preprocessing

Size of the MR images is prohibitive to analysis. We filtered the MR images to remove non-informative voxels prior to clustering. Because GBM occurs primarily in the white matter of the brain, most voxels elsewhere (ex. gray matter) do not contain tumor. By filtering out voxel locations in which no patients have GBM we eliminate 80% of the voxels without losing any information.

After this step 1 million voxels per sample remain, a large number of features that is impractical for use in most clustering algorithms. Especially since we have so few samples, we add another step to the filtering. Voxels that have little variation across the cohort – usually those in which only a few patients had tumors that overlapped the voxel – are uninformative for clustering patients into subtypes. Retaining only voxels having a minimum frequency of tumor occurrences across the cohort (Eq. A.1) results in a set of voxels with a spectrum of tumor event frequencies. We measured in units of tumor volume loss (Fig. A.7).

We denoise the voxel data by applying a tumor threshold filter. For each threshold, voxels having fewer tumor events than the threshold are masked from the data. M is the samples by voxels matrix, D is the denoised matrix. x and y are the

row and column identifiers and i is the specified threshold.

$$D_{x,y} = \begin{cases} M_{x,y}, & \text{if } \sum M_y \geq 1 \\ 0, & \text{otherwise} \end{cases} \tag{A.1}$$

We calculate tumor volume loss at each threshold, for sample x. X is the vector of voxels (binary values) and n is the length of the voxel vector.

$$\log_{10} \sum_{i=1}^{n} X_i == 1 \tag{A.2}$$

A minimum threshold of 15 tumor occurrences preserved voxels with an enrichment for high variability (high entropy) that would be informative for subtyping patients. Restricting to these voxels provided sufficient data for most of the patients in the cohort (Fig. A.7) and did not seem to bias the set around any particular subset of tumors. Tumor voxels removed by this threshold are those that buttress the most dense tumor regions. By removing these voxels we minimize relationships defined by the uncommon tumor regions– in most patients these will be completely non-tumor, inflating the patient-patient similarity scores. This also forces the clustering methods to focus on the core of common tumor regions, which are of more interest in this study.

A.1.2.2 2nd-Order HOCUS GBM Imaging– Location-Finding

While appearing to be symmetric, clusters of patients on either side of the brain (Fig. A.8(a)), clusters 4 and 5 are distinct from 3 in that there are 2 focal points of the tumors. Cluster 3 tumors show mid-region focus on the right side of the brain, whereas

clusters 4 and 5 split into 2 clusters on the left side. Tumor focal points of the patient groups are on the edges of that region– cluster 5 patients have tumor higher in the brain whereas cluster 4 patients have tumor much lower and closer to the base of the skull. Thus we do not see symmetry in tumor growth based on brain hemisphere. Despite close physical proximity, clusters 4 and 5 have wildly different survival prognoses– cluster 5 having a long projected survival whereas clusters 3 and 4 have the worst in the cohort (Fig. A.8(d)).

In previous works, Jain *et al* [160] and Liu *et al* [218] find a correlation between volume of tumor in MR images and patient survival. Our data shows a similar trend. To illustrate this we divided the tumors groups based on tumor volume (Supplemental Fig. A.9(a)), finding poor prognosis for large tumors. HOCUS clusters show larger separation in survival than the volume-based groups (Supplemental Table A.2), indicating that tumor volume is one of many vital components of the image data. Furthermore, there is no distinction in survival separation between tumors grouped by the expert annotations of anatomic location (p-value 0.596, Fig. A.10). HOCUS clusters span multiple locations specified by expert anatomic annotations.

There is little enrichment in HOCUS clusters of age, race, ethnicity, tumor status, gender, or surgical resection (Supplemental Table A.2). These are also independent of tumor volume. However, the clinical marker denoting ability of self-care, Karnofsky Performance Score, is enriched in first-order (direct comparison between samples) HOCUS clusters. First-order clusters are heavily volume-dependent.

A.1.2.3 Alternative Similarity Metrics

We tested alternative similarity metrics on the MR Imaging data– voxel-frequency scaled, Jaccard similarity, and tumor volume scaled (Supplemental Table A.3, Fig. A.11). Of these, scaling Hamming similarity by the volumes of the two tumors being compared led to inflated similarity scores of the same values which could not be clustered into meaningful groups. Voxel-frequency scaled clusters appeared to negate the volume-dependence of Hamming HOCUS, however it ignores negative matches (voxels where neither patient had tumor) and thus loses the anatomic dependence.

We selected voxel-frequency weighting because clustering using Hamming similarities can to build a cluster of small volume tumors, without appropriate filtering. To avoid this, we add weighting to the voxel matching, using term (aka voxel) frequency [311] such that in addition to summing the number of matched voxels, we scale the weight of the matches by the relative frequency of tumor in that voxel within the cohort. The similarity metric now also ignores tumor-free voxels when doing patient-patient comparisons, greatly reducing the number of voxels compared.

The two solutions are nearly identical except in two respects. First, voxel-frequency HOCUS disperses the Hamming HOCUS small-tumor cluster into the other 4 clusters. Second, it creates a new cluster with the 3 samples that have no visible tumor in the post-processed tumor data. Voxel-frequency clusters are still correlated with tumor volume and lose the location-dependence of the Hamming HOCUS clusters (Fig. A.1).

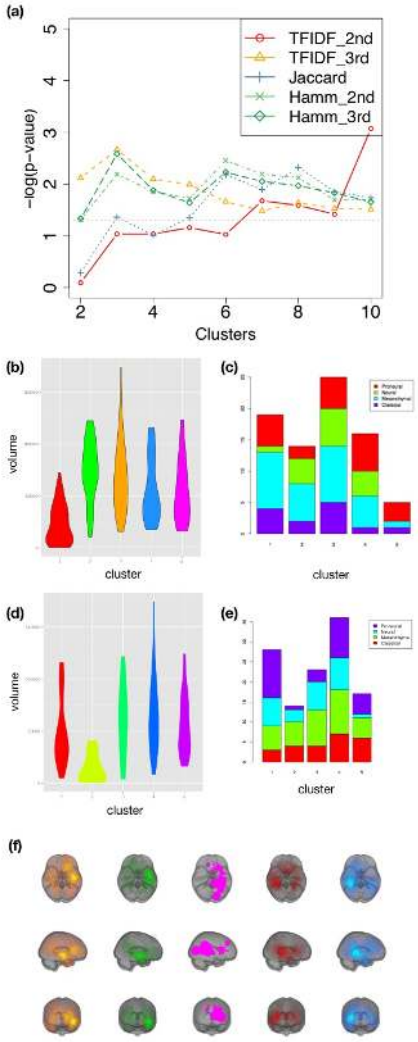

Figure A.1: Alternative similarity metrics used to compare patients. (a) P-values of survival differences between clusters for each similarity metric over a range of clusters, (b) tumor volumes of second-order TFIDF clusters, (c) molecular subtypes within each second-order TFIDF cluster. (d) Tumors volumes of Jaccard clusters, (e) barplot of molecular subtypes by Jaccard cluster, (d) brain images of Jaccard clusters.

Jaccard similarity is often used in analyzing social networks. Using this metric, HOCUS finds 5 clusters that are volume dependent and molecular subtype independent clusters (Fig. A.1). Survival separation is comparable to the Hamming HOCUS clusters, however the most different (and best surviving) cluster is composed mostly of small tumors.

Supplementary Table A.3 shows the correlations between each similarity metric clustering solution and tumor volume, molecular subtype, and survival. Equations for calculating each similarity metric are also in this table. Figure A.11 shows patient cluster membership changes between all clustering solutions.

A.1.3 Genomics

A.1.3.1 HOCUS of TCGA BRCA Mutations

Because of their demonstrated clinical import, BRCA subtypes are often defined using gene expression data. It has been shown that clusters based on mRNA transcription data readily identify luminal-like from the more aggressive basal-like tumors. While luminal tumors tend to be associated with the expression and presentation of the estrogen receptor, the basals tend to be less differentiated and lack this and other hormone receptors (such as progesterone receptor).

The BRCA subtypes identified by the second order HOCUS algorithm are distinct from the more established expression-based subtypes (e.g. basals and luminals). However, since the mutation-based clusters provide independent information for predicting patient survival in a multivariate analysis ($P < 5.428232e - 17$), it is critical

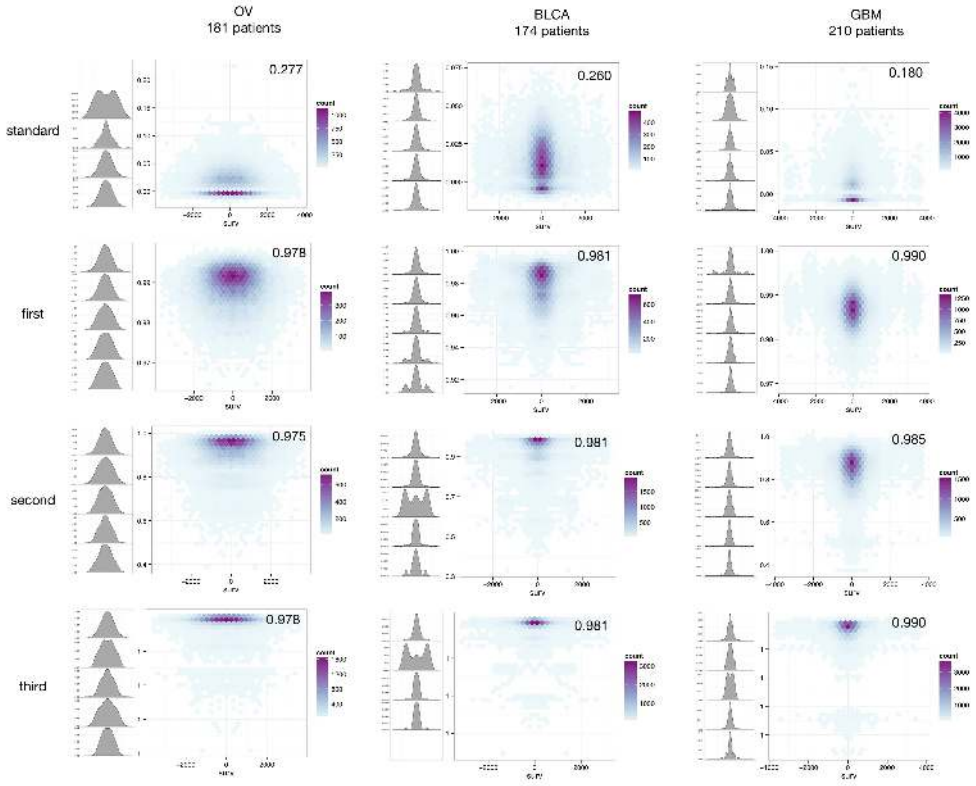

Figure A.2: Visualization of association between mutation-based and outcome-based similarity measures for TCGA cohorts: a) OV, b) BLCA, and c) GBM. Data was restricted to patients with a death event, then pairwise correlations were calculated in each feature space (pearson, 1st-, 2nd-, 3rd-order HOCUS) as well as difference in the length of survival, in days, between each pair of patients. A series of plots, one for each metric (pearson correlation, hamming similarity, or higher-order) for three different tumor analyses. In each plot, the joint density is shown in which the distribution of all sample pairs are depicted as density maps. On the left-hand side of each plot, a series of plots are shown in which the feature-based measure is divided into five bands of equal size, and differences in survival time (the outcome metric) are plotted in histograms for those samples restricted to each band. In every case tested, a higher-order metric could be found that had a positive association with the survival similarity metric, whereas pearson correlation, based on the original features, had seemed to have a low and sometimes negative association. For example, the surprising negative association of the pearson-based first-order measure is evident where most highly correlated sample pairs actually show an appreciable increase in samples with very different survival outcomes (seen as the introduction of extra "modes" in the top histograms). For BLCA and GBM cohorts the higher-order clustering solutions revealed subtypes with better survival separation than first-order metrics. For OV, the higher-order metrics performed comparably with Pearson just outperforming.

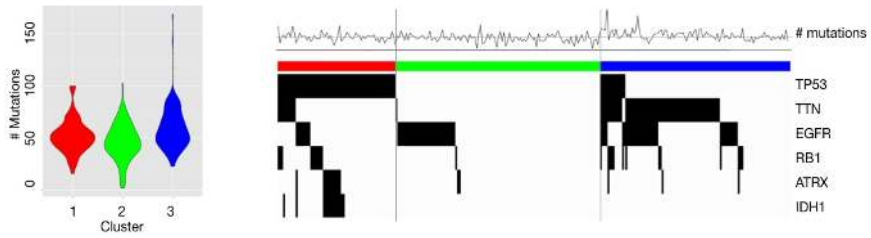

Figure A.3: Oncoprint showing a subset of mutations in GBM that are associated with cluster 1 via a χ^2 test and are mutated in at least 10 samples. Line plot above the oncoprint shows the total number of mutations per sample, and the grey line indicates median mutational load across the entire cohort. We show 5 frequently mutated genes that are associated with GBM mutations HOCUS result via a χ^2 test of independence, cluster 1.

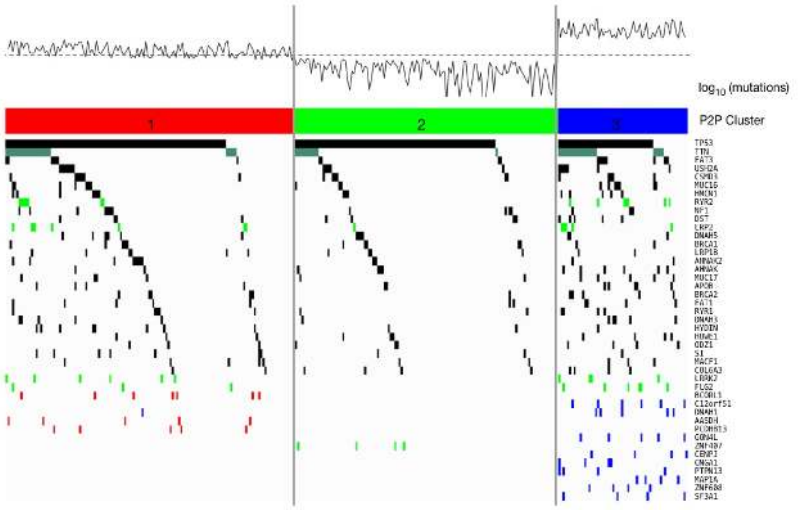

Figure A.4: Oncoprint showing a subset of mutations in OV. Line plots above the oncoprint shows the total number of mutations per sample. The grey dotted lines indicate median mutational load across the cohort. A combination of the most frequently mutated genes in the OV cohort (colored black) and the genes significantly associated with any 1st-order HOCUS cluster through a χ^2 test of independence are shown. Colors in the oncoprint indicate which cluster the mutation is associated with. TTN, a known passenger mutation, is associated with clusters 2 and 3.

to incorporate the SNV-based subtypes into categorization as has been shown by [65]. As might be expected, the first major split of the data separates tumors with elevated mutation rates (cluster 1 in Fig. A.12) from those with fewer mutations, such as cluster 3 that has the least. Cluster 4 is predominantly luminal A while the other clusters contain a mixture of the expression-based subtypes (see Fig. A.12). The samples in the cluster with elevated mutation rate are enriched for mutations in TP53 (50%), which may be an early event in these tumors enabling global loss of genome integrity. Consistent with this statement is the observation that these tumors also have the highest levels of copy number alterations as well (Fig. A.13). The hyper-mutated group also contains PIK3CA mutations that are often mutually-exclusive with TP53. In addition, the cluster contains mutations in what are thought to be passenger genes frequently mutated in some cancers possibly due to their location with respect to late replication fork timing during mitosis [107] such as TTN and MUC16. Taken together, the results suggest cluster 1 represents more advanced tumors.

On the other hand, cluster 3 not only contains the fewest mutations but also lacks mutations in TP53 or PIK3CA. Nearly all normal-like tumors fall into this group, which recapitulates the expression-based designation for these tumors. Samples in cluster 3 have mutations characteristic of luminal-A tumors such as in the PI3-kinase pathway (MAP3K1 and CDH1) and GATA3. Interestingly, several tumors classified as basal using expression data fall into this group. It would be interesting to determine if these tumors do indeed have TP53 mutations that were not detected through the TCGA's whole exome sequencing analysis (e.g. through regulatory mutations in pro-

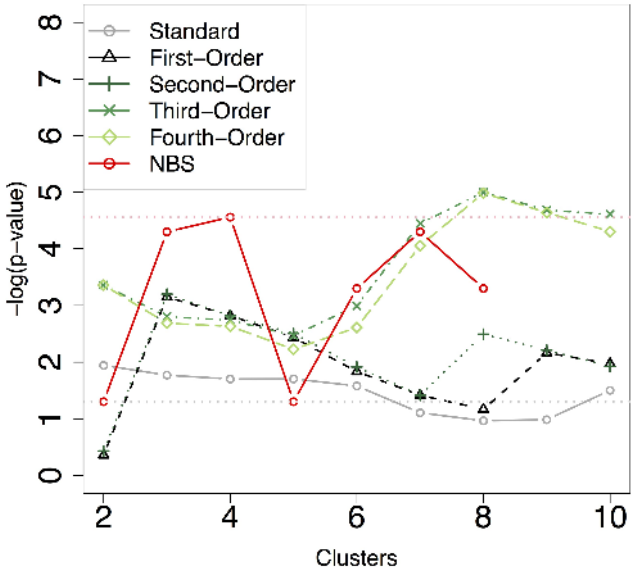

Figure A.5: Comparison to Network-Based Stratification [149] using the TCGA OV data used in their publication, and the same filtering.

moter elements) or through other mechanisms (e.g. epigenetic silencing) or genes that can influence TP53 function.

The HOCUS method identifies a PIK3CA mutated group (cluster 4), a clear hallmark of luminal-A breast cancers. This group also has enrichment for other PI-3-kinase pathway mutations such as in MAP3K1 and CDH1, underscoring the selective pressure to enhance signaling in this growth-related pathway for this tumor type.

Cluster 2 samples are notable for mutational frequencies equal to that across the cohort, and yet having neither TTN nor PIK3CA mutations. Approximately 30% of these samples have TP53 mutations.

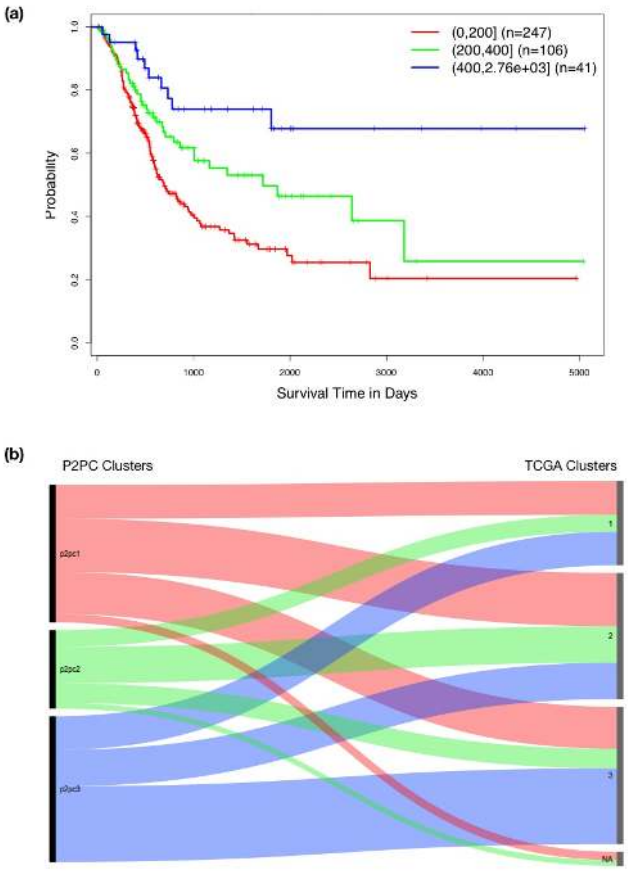

Figure A.6: (a)KM plot where samples are grouped by overall mutational frequency. P-value $4.7e-05$ (compared to HOCUS p-value $1.59e-05$), and (b) Alluvial diagram showing the difference in HOCUS 1st-order BLCA clusters and the TCGA-defined clusters based on mutation and CNA data. P-value 0.128 in a χ^2 test of independence. This diagram compares the 125 samples that are defined in both cluster sets.

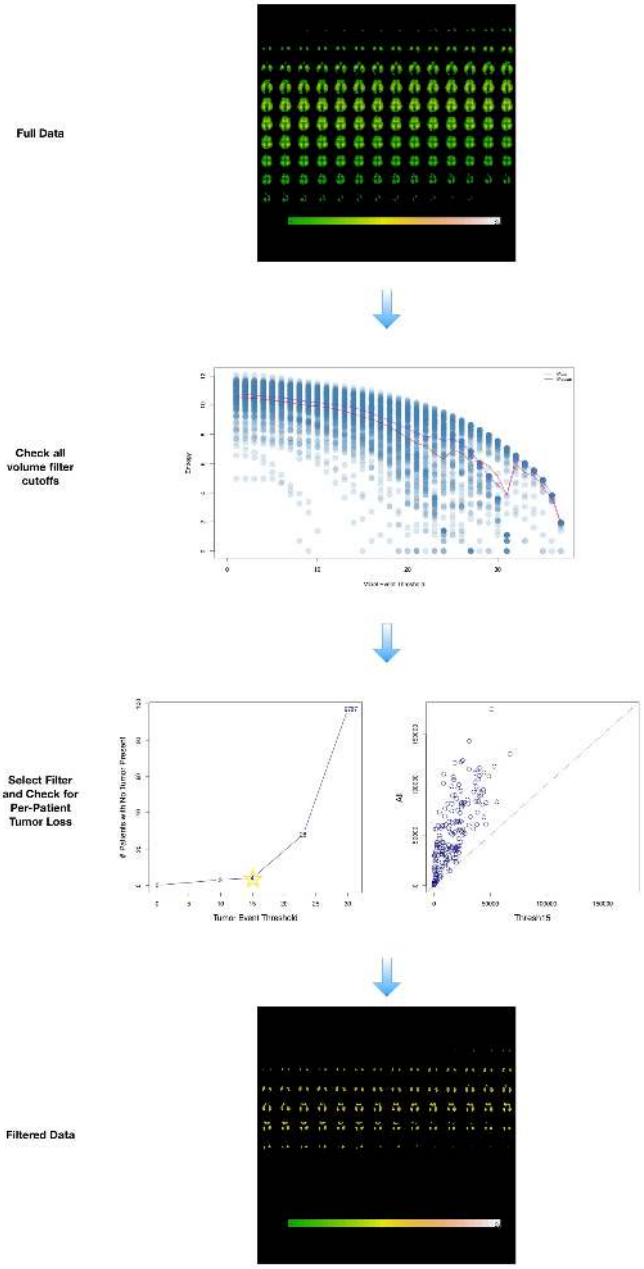

Figure A.7: MR images were filtered on a voxel level to indicate presence/absence of tumor in that region, after images were fit the the brain atlas. At each level of activity (number of patients having tumor in a given voxel) we calculate the $-log_{10}(tumor)$ visible after filtering below the given threshold. Cutoff was selected based on tumor loss per patient.

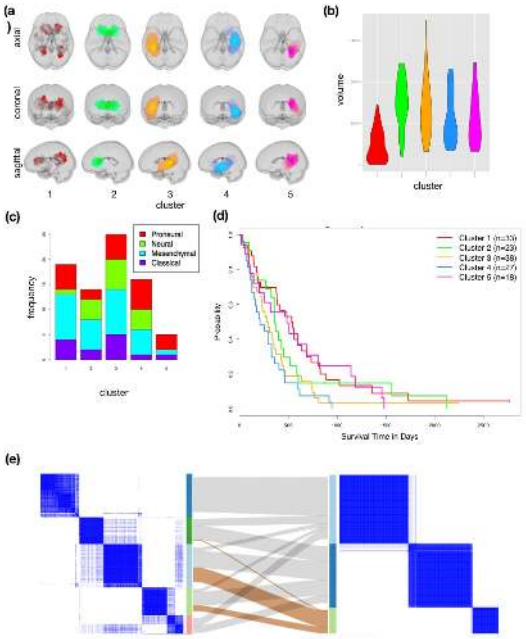

Figure A.8: (a) Sagittal, coronal, and axial views of the tumors within each image cluster (b) Violin plots of tumors volumes for each cluster. (c) Comparison to molecular subtypes defined by TCGA. (d) Kaplan-Meier plot of image clusters, showing clusters 3 and 4 to have poorer overall survival. (e) Consensus clustering matrices for 2nd- and 3rd-order HOCUS clusters, connected by an alluvial diagram showing that the majority of patients in 2nd-order clusters 3 and 4 (the poor survivors) make up the 3rd-order cluster 3.

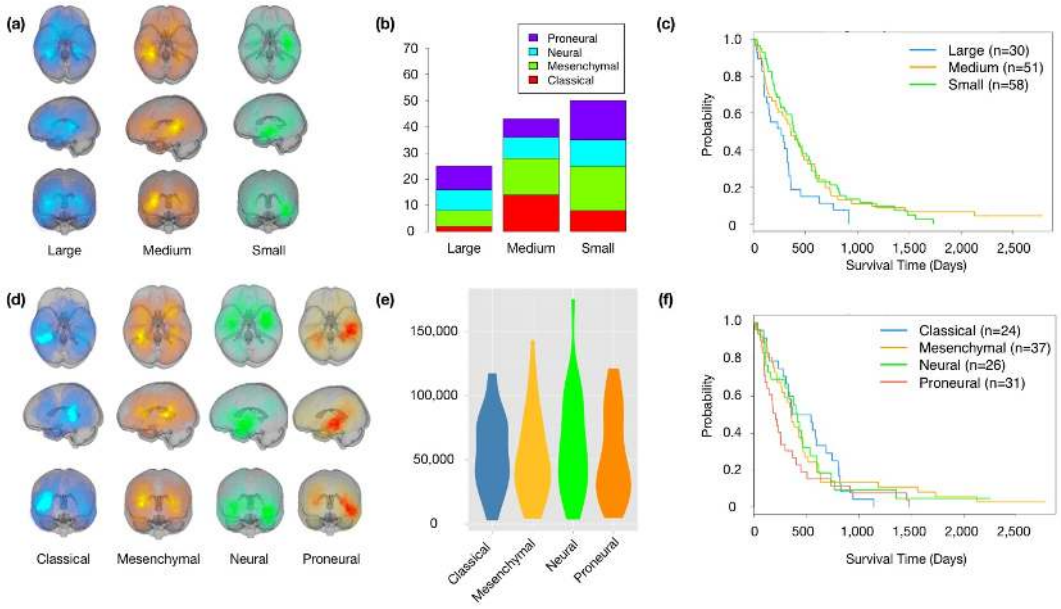

Figure A.9: (a-c) Patients grouped on tumor volume and (d-f) by TCGA defined molecular subtypes for MR image patients. (a) Images of patient tumors grouped by tumor volume (b) molecular subtypes (c) KM survival. (d) Images of patient tumors grouped by molecular subtype, (e) tumor volume per group, (f) KM survival.

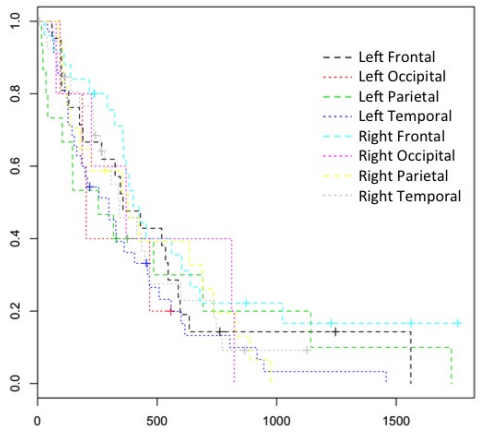

Figure A.10: KM plot of survival when patients are grouped by anatomic location of the tumor. Annotations indicate laterality (right/left) and lobes (parietal, occipital, frontal, temporal).

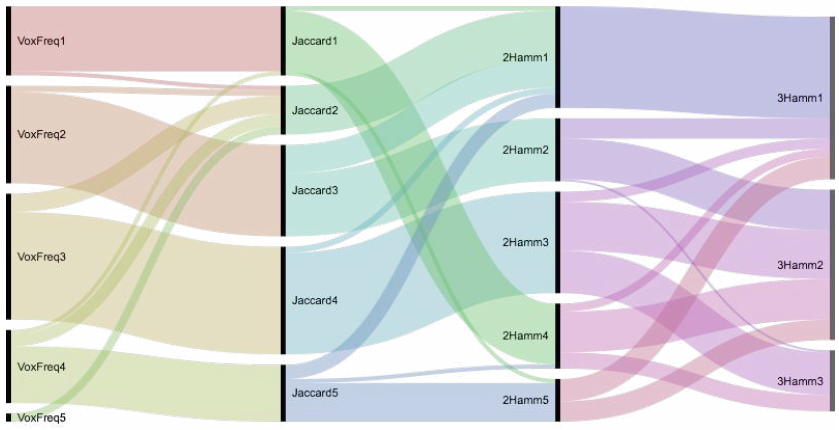

Figure A.11: Alluvial diagram of the different MR image clustering solutions. From right to left, voxel frequency, jaccard, second-order Hamming, third-order Hamming clusters.

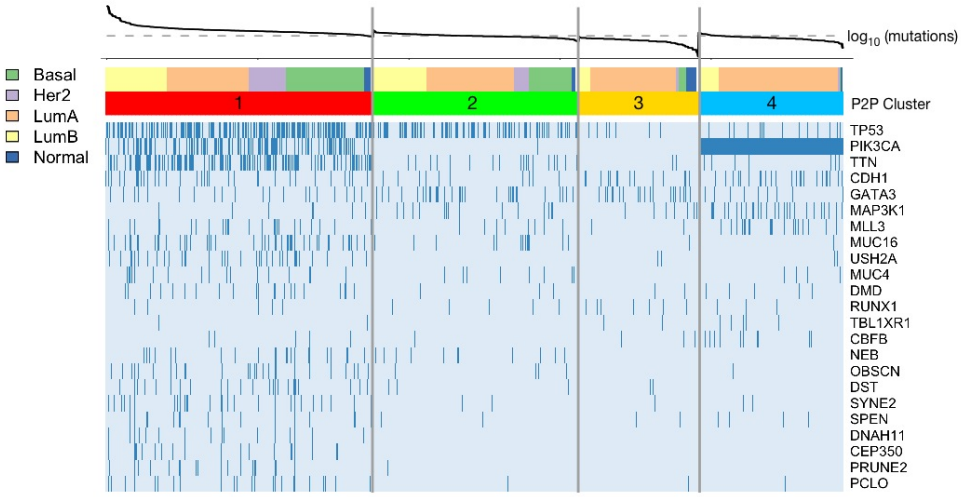

Figure A.12: Oncoprint showing the HOCUS BRCA clusters and associated mutations.

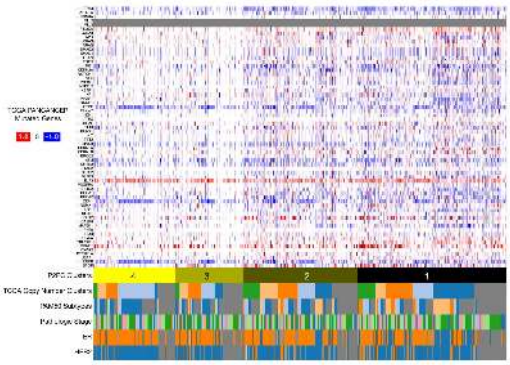

Figure A.13: Visualization of the BRCA copy number clusters and their correlation with the mutation-based subtypes from HOCUS. Heatmap made using the UCSC Cancer Genomics Browser [113], showing TCGA CNV subtypes and CNV alterations in the HOCUS clusters.

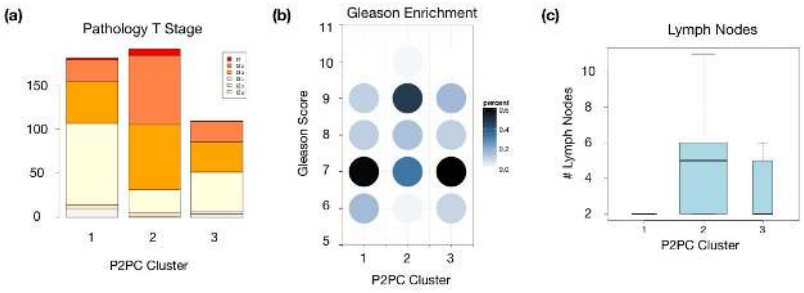

Figure A.14: (a) Pathlogy T stage of the HOCUS copy number clusters. (b) Enrichment of Gleason scores in the HOCUS clusters. Scores are normalized by column and color represents percentage of the cluster with a given combined Gleason score. (c) Boxplot of the number of lymph nodes each cluster's samples have invaded.

Table A.1: Kernel similarity scores between each HOCUS feature space, survival in days, and age.

GBM MRI	pearson	1st order	2nd order	3rd order	4th order
surv diff		0.966	0.814	0.977	0.977
standard					
1st order			0.902	0.992	0.992
2nd order				0.846	0.846
3rd order					1.000
4th order					

OV Mutations	pearson	1st order	2nd order	3rd order
surv diff	0.277	0.978	0.976	0.978
standard		0.265	0.274	0.264
1st order			0.997	1.000
2nd order				0.997
3rd order				

GBM Mutations	pearson	1st order	2nd order	3rd order
surv diff	0.180	0.990	0.985	0.990
standard		0.174	0.184	0.173
1st order			0.997	1.000
2nd order				0.997
3rd order				

BLCA Mutations	pearson	1st order	2nd order	3rd order	4th order
surv diff	0.260	0.981	0.981	0.981	0.981
standard		0.314	0.311	0.314	0.314
1st order			0.997	1.000	1.000
2nd order				0.996	0.996
3rd order					1.000
4th order					

Table A.2: P-values from χ^2 tests between image clusters of all types and clinical covariates.

Clin	First	Second	Third	Jaccard	FreqW	By Vol.
karnofsky score	0.00495	0.255	0.0955	0.0642	0.281	0.0745
race	0.0834	0.0893	0.0363	0.172	0.687	0.863
history lgg dx	0.2	0.646	0.293	0.681	0.624	0.153
days to last followup	0.361	0.483	0.472	0.442	0.52	0.498
history neoadj. treat.	0.384	0.413	0.384	0.112	0.624	0.504
days to birth	0.389	0.389	0.455	0.389	0.401	0.415
patient id	0.44	0.44	0.44	0.44	0.445	0.451
days to death	0.476	0.482	0.486	0.388	0.44	0.409
ethnicity	0.666	0.248	0.635	0.199	0.324	0.385
perf. status timing	0.684	0.839	0.936	0.676	0.827	0.677
hist. diagnosis	0.735	0.377	0.606	0.216	0.602	0.755
initial path dx	0.76	0.446	0.267	0.204	0.272	0.42
gender	0.76	0.661	0.197	0.816	0.867	0.46
vital status	0.85	0.36	0.937	0.757	0.655	0.424
tumor status	0.879	0.455	0.819	0.273	0.332	0.104
age at init. path. diag.	0.894	0.461	0.228	0.3	0.433	0.789
form completion year	0.943	0.107	0.892	0.212	0.358	0.504
init. path. dx year	0.953	0.474	0.693	0.35	0.512	0.507
tissue source site	0.957	0.273	0.94	0.508	0.732	0.355
form compl. month	0.961	0.0429	0.942	0.0965	0.162	0.868

Table A.3: P-values for each similarity metric in a χ^2 test of independence.

Metric	Formula	Surv. Assoc.	Volume Dep.	Molec. Subtype
Hamming	$\sum_{i=1}^{n}(x_i == y_i)$	2.02e-02	4.10e-01	4.19e-01
FreqW	Eq. A.3	1.52e-01	1.58e-01	2.91e-02
Jaccard	$\frac{X \cap Y}{X \cup Y}$	1.98e-02	4.72e-01	2.39e-01

$$\sum_{i=1}^{n} \begin{cases} z_i & \text{if } x_i == y_i \\ 0 & \text{otherwise} \end{cases} \quad \text{s.t. } z_i = \sum_{j=1}^{m} (m_j == 1) \tag{A.3}$$

A.2 Supplement for: Multiview learning

A.2.1 Mutation Issues

Sensitivity cannot be predicted using single genes for several reasons. For example, KRAS and BRAF mutations are common in the CCLE data. For some drugs that should be sensitive in presence of a KRAS mutation, BRAF mutations in the KRAS wildtype patients make them also sensitive, so that overall the drug does not appear correlated with KRAS mutation status. When BRAF mutants are removed, KRAS mutations are predictive of drug sensitivity. The opposite also occurs, where many BRAF mutants share another mutation that is unrelated to a specific drug sensitivity, but because of the shared BRAF mutation the other mutation appears to confer sensitivity. Looking at the mutation status of one gene is not enough to make conclusions about the sensitivity to a drug.

A.2.2 Biological Priors

A.2.2.1 Biological Gene Sets

Metabolic Enzymes: The metabolic enzymes gene set was created by collecting all genes in the CCLE data belonging to the Cytochrome P450 (CYP) family. CYP proteins

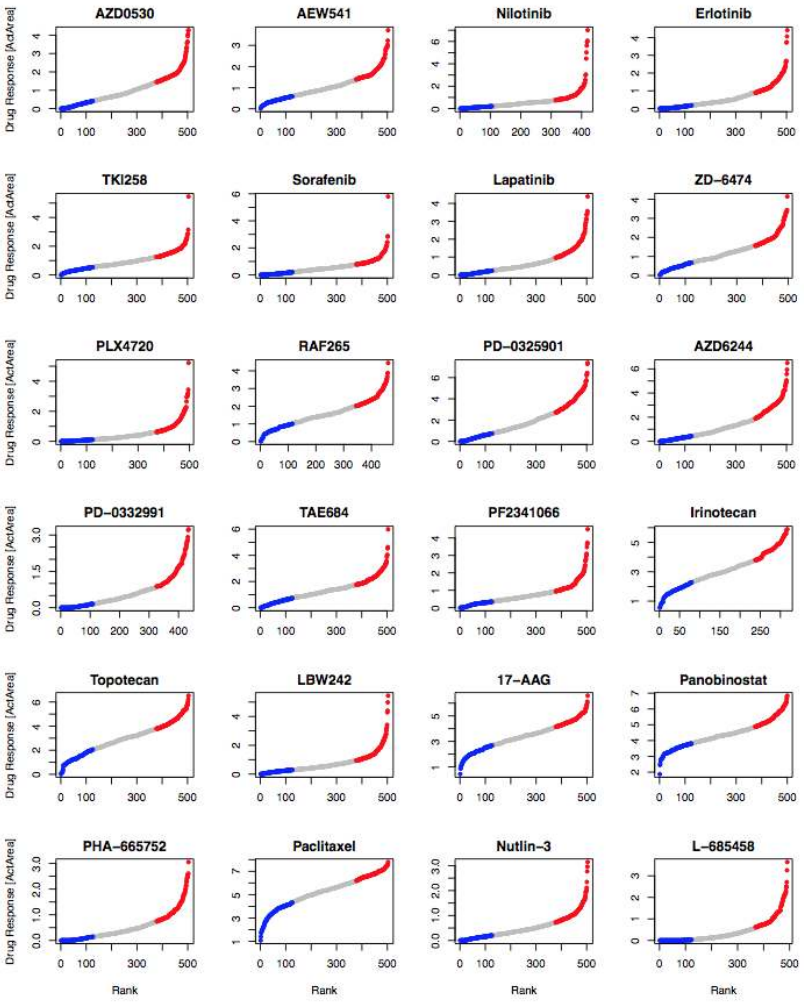

Figure A.15: The ranked ActArea values of each CCLE cell line for the 24 CCLE compounds. Blue dots are cell lines labeled as 'non-sensitive for the correspondent drug, red ones are labeled 'sensitive, gray ones 'intermediate. The number of cell lines in the non-sensitive class corresponds to the bottom 25% of cell lines the drug response was measured for, the sensitive class to the top 25%.

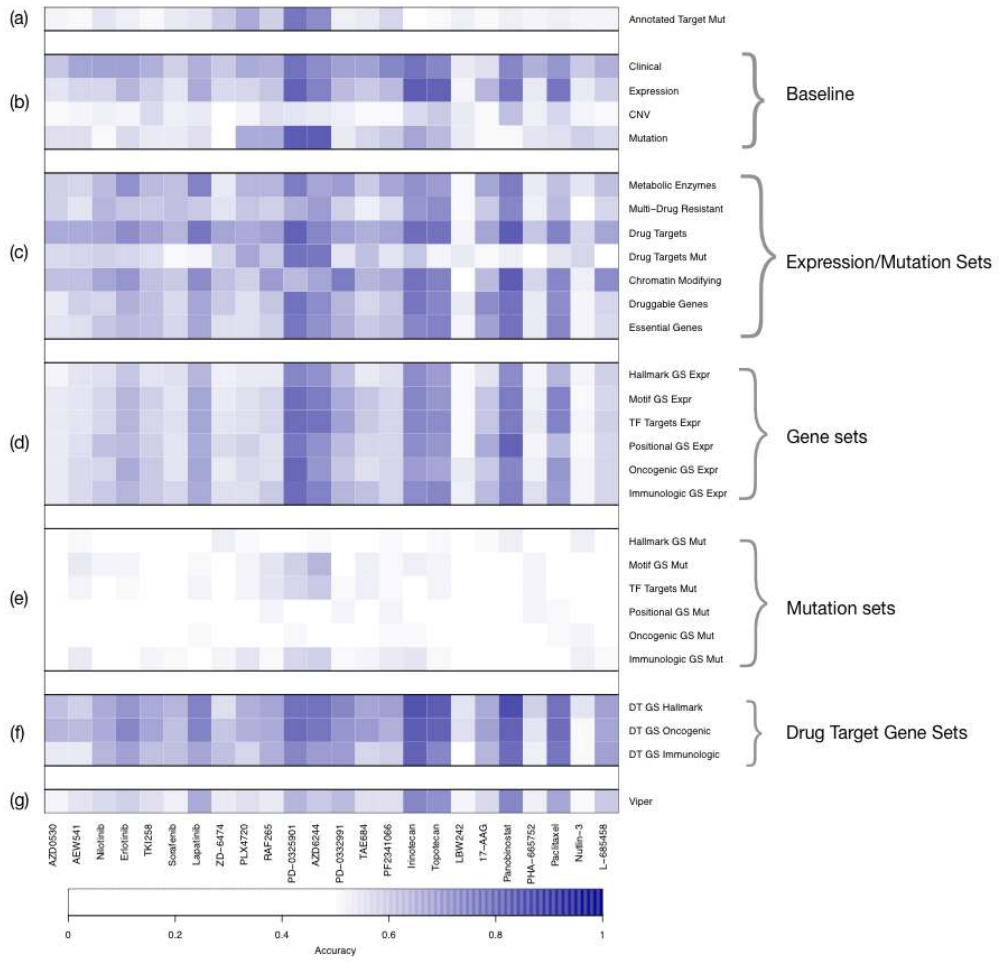

Figure A.16: AUC for each view when predicting sensitivity to each drug in CCLE. Grouped by data type. The cross-validated AUC of single views with their optimized parameter settings. All values ≤ 0.5 (AUC of a random predictor) are shown in white. The simple Annotated Target Mutation predictor (Section 4) is shown in A. The following single views are grouped according to Section A.2.2. GS = Gene Set; DT = Drug Target; Expr = Expression; Mut = Mutation.

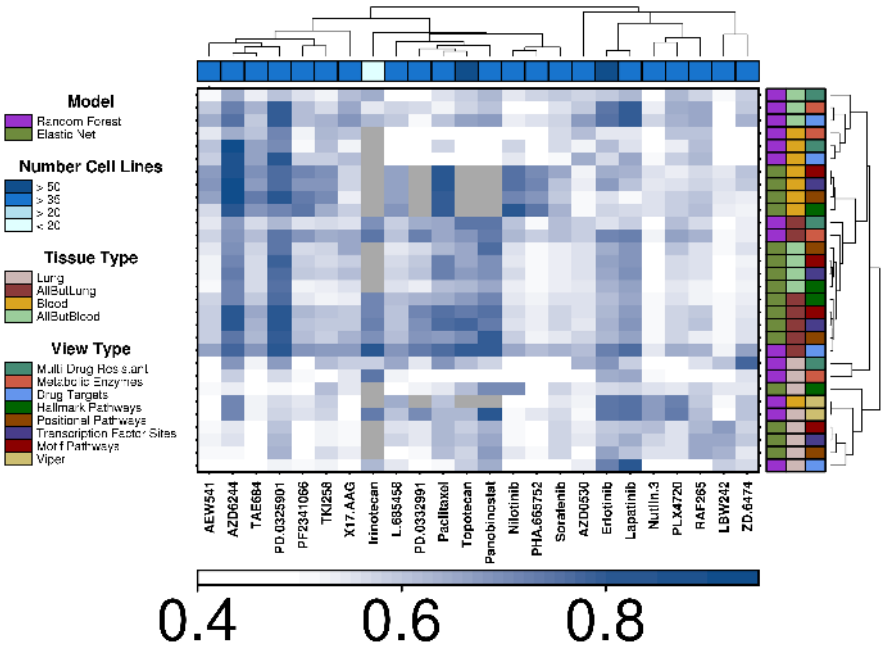

Figure A.17: Tissue-specific run of MVL.

are the key players in drug metabolism; They deactivate or facilitate the excretion of most drugs, but they also transform many drugs into their active form [127]. There are 53 CYP proteins in the CCLE expression data.

Multi-Drug Resistance Proteins: Expression data was subset to a list of multi-drug resistance proteins based on [175]. All 12 defined proteins are present in the data set.

Drug Targets: This view includes all proteins targeted by the 24 anti-cancer compounds in the CCLE data set. The information about drug - protein interactions was collected from DrugBank [203], a recent review of drug targets [278], the Drug Gene

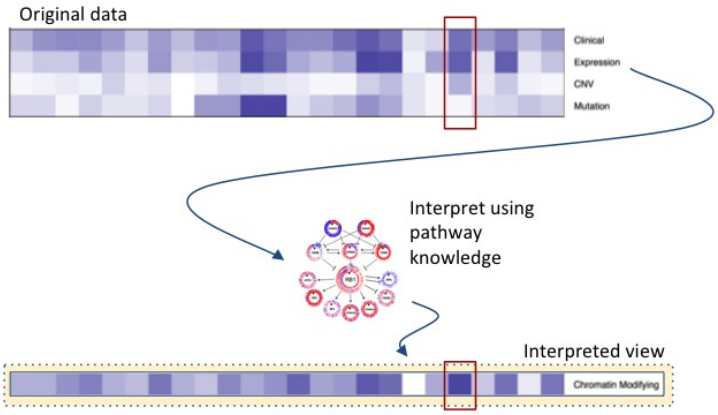

Figure A.18: Interpreted views can be created from any of the baseline data platforms. In this example, mRNA expression data is subset to a list of known chromatin–modifying genes [3]. The new view has higher AUC than the baseline view. One of the largest differences is in Panobinostat, which AUC is highlighted in red.

Interaction Database (DGIdb) [126], and manual literature curation for drugs without an annotated target in the sources named before (see Table A.4). In total, 142 genes found to be drug targets were present in the gene expression data set. In addition to the expression-based view, this view was created using the mutation data, in which 82 of the drug targets are present.

Chromatin-Modifying Enzymes: This gene set includes chromatin-modifying proteins [3, 152]. It contains 65 proteins, of which 56 are present in the gene expression data.

Druggable Genes: The druggable genes view was created from DrugBank [203], a recent drug target review [278], cell surface proteins as defined in [74], membrane proteins and genes on the druggable genome list from DGIdb [126], a manually curated list of kinases (Table A.4), and the Therapeutic Target Database (TTD, [410]). In contrast to the Drug Targets view, the proteins in this set are not limited to the 24 CCLE compounds. Proteins that are not a target of any existing drug, but have the characteristics to serve as one, are included. A total of 4,632 genes from this gene set are present in the gene expression data.

Essential Genes: The information about essential genes in cancer cell lines was retrieved from Project Achilles [71], an effort to identify genes having an effect on cell viability by using short hairpin RNA (shRNA) screens. Two versions of Achilles were merged in order to maximize the overlap with the cell lines in CCLE: Achilles v.2.11

Table A.4: Drug targets manually curated from a literature review.

CCLE Compound	Target	Source
L-685458	PSEN1	[213]
L-685458	PSEN2	[213]
LBW242	XIAP	[92]
Nutlin-3	MDM2	[348]
PHA-665752	MET	[382]
TAE684	ALK	[103]

and v.2.4.3. The 30 most essential genes for each cell line present in both CCLE and Achilles were retrieved. CCLE expression data was subset to the union of these genes resulting in 2,064 features for this view.

A.2.2.2 MSigDB Gene Sets

The Molecular Signatures Database (MSigDB) [316] provides biological gene sets in different collections. Median, variance, and kurtosis values of gene expression in each gene set was calculated and defined as a feature. For using CCLE mutation data with MSigDB gene sets, the enrichment of the mutated genes of a cell line in a gene set was tested using hypergeometric distribution (R function *phyper*). The following MSigDB collections were chosen:

Hallmark Gene Sets: A collection of gene sets created from overlapping gene sets. It features reduced noise and redundancy and contains 50 gene sets.

Motif Gene Sets: 836 gene sets containing genes that share conserved cis-regulatory motif [375].

Transcription Factor Targets: A gene set contains all genes sharing a transcription factor binding site defined by a TRANSFAC record [233]. There are 615 gene sets in this collection.

Positional Gene Sets: Gene sets corresponding to the position of genes on the human genome regarding chromosome and cytogenetic band. The collection holds 326 gene sets.

Oncogenic Signatures: Signatures of 189 cellular pathways which are often disregulated in cancer.

Immunologic Signatures: The 1,910 gene sets represent cell states and perturbations within the immune system.

A.2.2.3 Drug Target Gene Sets

For each drug target defined in Section A.2.2.2, all genes occurring in at least one gene set together with the target gene were unified to build one drug target gene set. The MSigDB collections Hallmark, Oncogenic, and Immunologic were used separately. As before, median, variance, and kurtosis of the expression values were calculated for each drug target gene set and used as features.

A.2.2.4 Regulator Activity by Viper

Virtual Inference of Protein-activity by Enriched Regulon analysis (Viper) is a tool to transform gene expression features into regulator activity [5]. It takes as input

gene expression, a regulon (bipartite regulation network of regulators, e.g. transcription factors), and the genes that are regulated by them. Here, a general regulon called 'multinet' [177] and the CCLE expression data were used. Viper was run in R as part of the Bioconductor project [4].

A.2.3 Data in detail

Mutation CCLE includes mutation data in the form of Single Nucleotide Polymorphism (SNP) and insertion or deletion (Indel) events for 1,651 genes. These were assessed by targeted massively parallel sequencing and later filtered, *e.g.* for presumably neutral variants or common polymorphisms. Additionally, 392 mutations in 33 genes known to be associated with cancer [227] were assessed by mass spectrometric genotyping. SNPs and Indels were combined into a set of non-silent mutations that include all events changing the amino acid composition of the resulting protein, including Indels or missense SNPs in the coding region, splice site, and stop or start codon alterations.

Expression Gene expression data measures expression over 18,900 genes using Affymetrix U133 plus 2.0 arrays, converted to single gene values by Robust Multi-array Average (RMA) and quantile normalization.

CNV Copy Number Variation (CNV) data covers 23,316 genes and was determined using Affymetrix SNP6.0 arrays, and normalizing the values with the most similar HapMap normal samples. In the CNV data, some genes have the same value for each cell line because they lay on the same genome segment varying in copy number. Each of

these gene sets was merged into one feature in order to reduce redundancy in the data, resulting in 20,247 features.

Clinical Sample annotation data for the CCLE cell lines contains the gender of the cancer patient and information about the cancer origin (*i.e.* 24 different tissue types, 21 histology types, and 67 histology subtypes).

Drug Sensitivity There is drug response data to 24 anti-cancer drugs for about 50% of the cell lines in CCLE. A fitted dose-response curve from eight measurements is given, together with the inferred values for EC50, IC50, and Activity Area (ActArea), the area over the dose-response curve (see Fig. 2b from Barretina *et. al.*[17] for definition). ActArea was used for all analyses in this work for three reasons:

1. ActArea captures more information about the dose-response curve than a single point like IC50 or EC50, *i.e.* the angle of the curve and initial points of sensitivity changes.

2. ActArea is always given. EC50 in contrast is set to NA if no sufficient response was measured with the maximal tested dose.

3. ActArea has no artificial values, whereas IC50 is set to the maximal tested dose if no response was measured.

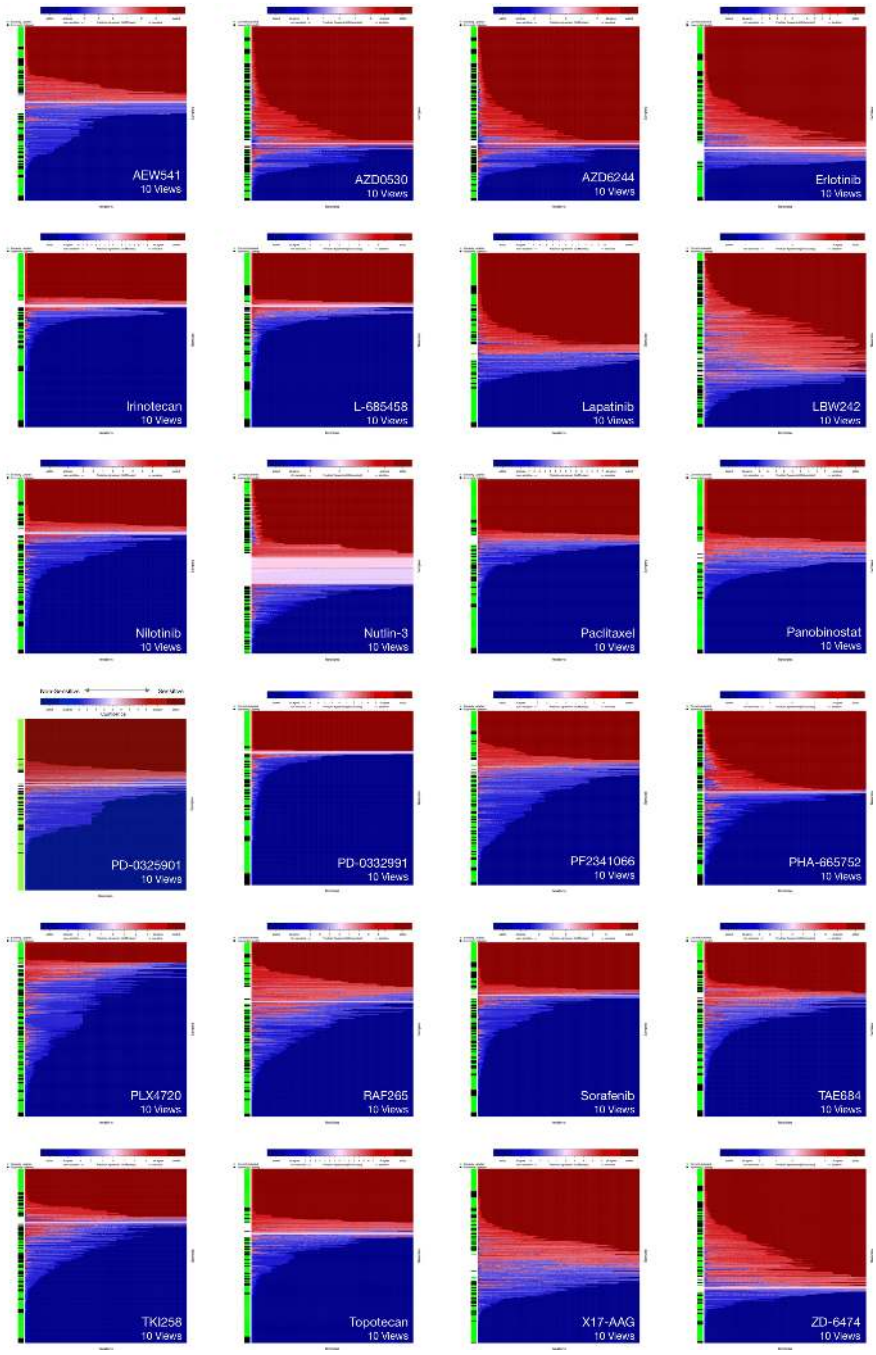

Figure A.19: Label-learning validation for all 24 CCLE drugs. Drug names in bottom right corner of each LLV plot.

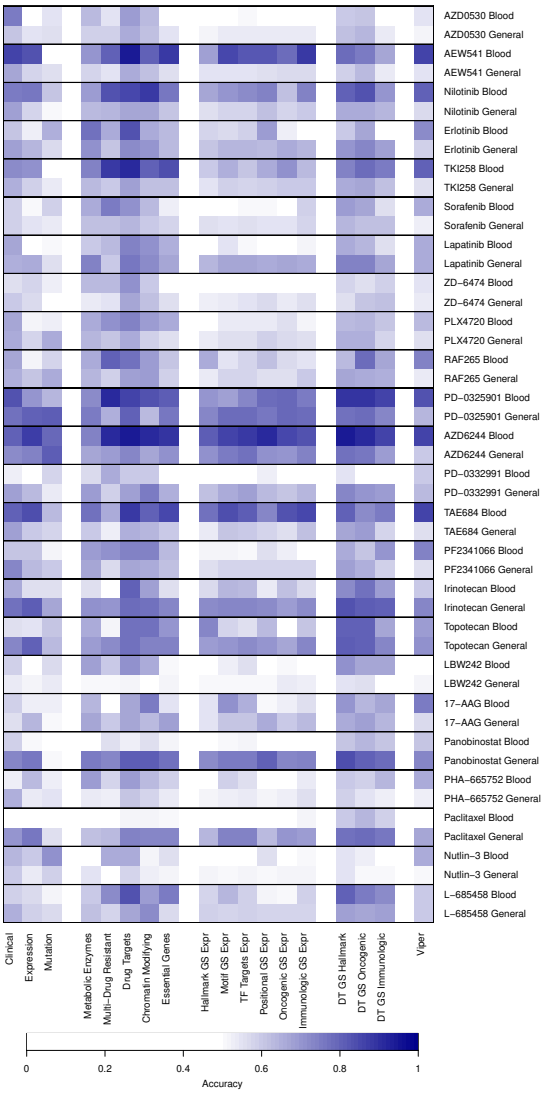

Figure A.20: Cross-validated AUC of single views with their optimized parameter settings. This compares the tissue–specific setting using blood cancer cell lines to the complete CCLE.