COMPUTATIONAL METHODS FOR PERSONALIZED CANCER GENOMICS

BY

JACK PU HOU

DISSERTATION

Submitted in partial fulfillment of the requirement
for the degree of Doctor of Philosophy in Bioengineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2017

Urbana, Illinois

Doctoral Committee:

Professor Olgica Milenkovic, Chair
Associate Professor Jian Ma, Director of Research
Professor Paul Hergenrother
Professor Jun Song

Abstract

In recent years, cancer treatment strategies have moved towards personalized approaches, specifically tailoring cancer treatments on a single-patient basis using molecular profiles from the patients' tumor genomes. Knowledge of a patient's molecular profile can be used to 1) identify the disease mechanisms and underlying cause of a single patient's cancer, 2) assign patients into treatment groups based on the molecular prognosis, and 3) recommend potential treatments for individual patients based on the patient's molecular signature data. However, the bottleneck of the personalized medicine approach lies in the challenge of translating the vast amount of sequencing data to meaningful clinical insights.

This dissertation explores several computational methods that utilize molecular signature data to understand disease mechanisms of cancer, categorize patients into biologically relevant subtypes, and recommend drug treatments to patients. In the dissertation, we present a method, DawnRank, a patient-specific method that determines the potential driving genomic alterations (the drivers) of cancer. We expand on DawnRank's capabilities by using the DawnRank scores in key driver mutations and copy number variants (CNVs) to identify breast cancer subtypes. We found 5 alternative subtypes based on potentially clinically relevant driver genes, each with unique defining target features and pathways. These subtypes correspond to and build upon our previous knowledge of breast cancer subtypes.

We also identify disease mechanisms in identifying key novel cancer pathways in which driver genes interact. We developed a method, C^3, which pinpoints patterns of cancer mutations in a pathway context from a patient population to detect novel cancer pathways that consist of significant driver genes. C^3 improves on current methods in driver pathway detection both on a technical aspect and a results-oriented aspect. C^3 can detect larger and more consistent pathways than previous methods as well as discovering more biologically relevant drivers. Finally, we address the issue of drug recommendation in the wake of molecular signature data. We develop a method, Scattershot, which combines genomic information along with biological insights on cancer disease mechanisms to predict drug response and prioritize drug treatments. Scattershot outperforms previous methods in predicting drug response and Scattershot recommends drugs to cancer patients that are in line with the actual drugs prescribed by the physician.

Acknowledgments

First, I would like to acknowledge my advisor, Dr. Jian Ma, for whose mentorship through the last 5 years has been instrumental to the development a PhD student and in life. Many challenges were overcome during my adventure as a PhD student, and to all the long nights spent toiling to satisfaction of results and success I dedicate to Dr. Jian Ma. I also would like to acknowledge the Ma Lab, most notably Yang Li, Yang Zhang, Ashok Rajaramaran, Yuchuan Wang, Dechao Tian, Yiyi Liu, and (soon) Ben Chidester for supporting me in my endeavors both inside and outside the lab, from discussing complex problems to playing tennis or hanging out with me in Pittsburgh. I wish the Ma Lab good fortune in Carnegie Mellon University.

I acknowledge my other Doctoral Committee members: Dr. Olgica Milenkovic, Dr. Paul Hergenrother, and Dr. Jun Song. The completion of a PhD is a milestone achievement of my life, and I thank my committee for challenging me to be better than myself. To leading thought provoking discussions and projects. Additionally, I thank the Chair, Dr. Milenkovic, for her invitation to bring me on board to finish the C^3 correlation clustering project as well. I thank Dr. Milenkovic and Dr. Hergenrother for advice and help on my other academic research projects as well.

I acknowledge the Dr. Chuck Perou of the University of North Carolina – Chapel Hill. I remember a simple lunch meeting with the Cancer Community at the University of Illinois where I met the researcher who discovered Triple Negative subtype of breast cancer. A simple visit could lead to one of the most fulfilling projects applying the insights of my first paper to determine new subtypes of breast cancer. I would like to acknowledge the Perou Lab, especially Grace Silva and Marni Siegel, for their help that continues to this day. I acknowledge my other research collaborators as well. Dr. Edison Liu, Koichiro Inaki, ChengXiang Zhai and Sam Ng for helpful discussions in DawnRank; the Milenkovic Lab, especially Greg Puleo, Amin Emad, and Minji Kim for guidance and assistance in the C^3 project; and Dr. Christos Faloustos and Hyun Ah Song for their guidance on the Scattershot Project.

I acknowledge the Bioengineering Graduate Program, the College of Medicine, and the Medical Scholars program (MSP). Thank you for providing me with the wonderful opportunity to broaden my horizons, not only to participate in groundbreaking research, but to provide me with the education for me to do the most good to help patients in the bedside to the best of my ability. I acknowledge MSP directors Drs. Jim Hall and Jim Slauch for keeping me on track through the hard times and the good times. I thank the MSP program and my M2 class who join me in my journey. I especially thank my Tuesday lunch buddy Kerim Kaylan for the support he showed my throughout my expedition.

Lastly I acknowledge my support system. I truly could not finish this adventure without the help of my friends and my family. I thank my roommate Dr. Ryan Peddy. From middle school to now, it has been a pleasure to know you. I thank my friends in my past life especially Alex Jones, Charles Li, and Royce Blackledge, for keeping in touch and visiting me throughout my campaign for the PhD. My sanity thanks you the most. I thank the friends I made while in Champaign-Urbana especially Liz Mayerich, Max Candocia, Kai Kobayashi, Jon Cruz, Cameron Mooney, and Michael Vitoux. It has been a pleasure to share my experiences with you. Most of all, I acknowledge my family. My mom, my dad, my sister Emily and my brother Eric. There is no greater family I could have ever asked to be a part of. I am truly blessed.

Table of Contents

Chapter 1: Overview

Cancer is a disease of the genome that is the second-leading cause of death in the United States [1]. Although a "cure" for cancer is a top priority goal for medicine and society alike, cancer is notoriously difficult to treat due to the vast amount of genetic diversity and heterogeneity among tumors [2], [3]. Tumor heterogeneity is the observation that tumors present distinct morphological and phenotypic profiles, and can occur both between two different tumors or even within the same tumor [4]. Tumor heterogeneity can be explained by the clonal evolution model of cancer, in which tumorigenesis occurs when a genomic alteration (a driver) in a cell improves a cell's fitness, allowing it to outcompete in its environment, divide and grow eventually leading to tumorigenesis with the cells of the tumor sharing a common ancestor [5]. Because the driver occurs at the molecular level, understanding the human genome and molecular signature information is a crucial tool in combatting cancer [6].

The ability to utilize vast amounts of molecular signature data has been made possible in the recent years. Advances in the scope and reduction in the cost of next-generation sequencing technologies have provided us with an opportunity to better characterize the molecular signatures of human cancers. Information from sequencing can be used to identify perturbations between cancer cells and normal cells that contribute to tumorigenesis on a single-patient basis, which can be used to classify a patient's cancer as well as recommend potential life-saving cancer treatment. The single-patient precision of NGS data paves the way for personalized treatment strategies in cancer [7]. The ultimate goal of personalized medicine is to integrate genomic information with traditional treatment methods (the patient interview, laboratory testing, and socioeconomic and environment factors) to provide a treatment plan that is tailored to a single patient, revolutionizing the way cancer care is conducted [7], [8].

Although the prospects of a personalized medicine are promising, the challenges of applying the genomic information obtained from the lab bench to the bedside are substantial [9]. Three of the most prominent challenges in realizing the dream of personalized medicine to the clinic are: (1) improving our understanding cancer development and progression by identifying the drivers of cancer [10], [11]; (2) applying our understanding of the genetic basis of cancer to better improve diagnostic capabilities [12]; (3) integrating genomic and diagnostic information to ultimately select and prioritize effective treatments for cancer patients [13]. The sheer volume of complex, multidimensional data that represents the cancer genomic profiles makes it difficult to analyze such molecular signature data [14]. Therefore, new computational and statistical methods are needed to model tumor mechanisms, diagnostic subgroups, and treatment suggestions. The objective of this dissertation is to develop new computational methods that identify important genomic alterations related to tumor mechanisms, comprehensively stratify patient subgroups, and effectively recommend drugs for personalized cancer medicine.

1.1 Personalized approaches in driver identification in cancer

A key question in cancer genomics is focused on identifying the drivers and the driving mechanisms behind the important tumorigenesis pathways related to tumor development and progression. A driver is considered to be a genomic alteration such as a mutation or a copy number change that significantly increases the fitness of the tumor. The functionality and driving pathways of these alterations may vary. Some hallmark examples include constant response to growth signals, no response to anti-growth or apoptosis signals, improved replication potential with telomerase, sustained angiogenesis, and factors that promote invasion and metastasis [10].

Although the alteration rate is high in many cancer cells, only a small number of mutations will lead to tumorigenesis. A key challenge lies in distinguishing “driver” mutations, which contribute to tumorigenesis, from functionally neutral “passenger” mutations [15]. The most basic approach is to categorize mutations based on recurrence, i.e., the most commonly occurring mutations are more likely to be drivers [16], [17], or by comparing mutation rates in individual genes based on an empirically derived background mutation rate, such as MutSig [18] and MuSiC [19]. Machine learning-based approaches use existing knowledge to help identify drivers. For example, CHASM utilizes random forest to classify driver mutations using alterations trained from known cancer-causing somatic missense mutations [20] and CONEXIC was developed to integrate copy number change and gene expression data to identify potential driver genes located in regions that are amplified or deleted in tumors [21]. One very promising class of driver detection methods models the interaction a driver might have with associated genes in a cancer pathway. Network and pathway-based approaches are one of the most promising methods to understand drivers due to their ability to model gene-gene interactions by aggregating small effect sizes from individual genes. Examples of network-based driver models include PARADIGM-Shift [22], which was developed to utilize pathway-level information along with other features (such as expression, methylation, copy number) to infer gain and loss of function for mutations; DriverNet [23], which classifies driver mutations as mutations that propagate outlying downstream differential expression in the transcriptional regulatory network [23]; and

MAXDRIVER, [24] which was proposed to identify driver genes by integrating multiple omics data and heterogeneous networks. Our method DawnRank, addresses several of the shortcomings of previous methods by providing a truly patient-specific model that does not require population-level information to make an inference on driver genes. Since DawnRank, multiple new ensemble methods have been developed to build consensus drivers that are found by multiple types of previously-established methods, including DawnRank. Ensemble methods combine insights from recurrence-based information, sequence information, and network information. Two recent ensemble methods: EC [25] and MADGiC [26] incorporate this information in a machine learning framework to predict drivers. DriverDBv2 [27] is an ensemble method that detects drivers from multiple established sources, including results from in DriverNet and DawnRank.

Driver identification software has contributed tremendously to our understanding of how alterations in genes may impact cancer. However, the narrative of tumorigenesis does not necessarily begin and end with a single alteration in a driver gene. An altered gene may have many downstream effects, leading to effects on several pathways that drive cancer [28]. Discovery of driver pathways provides insight on how mechanisms of tumorigenesis. Several methods have been proposed to model potential driver subnetworks and pathways. One method, MEMo [29], found closely related driver groups, called modules, that contribute to tumorigenesis using principles of mutual exclusivity. The mutual exclusivity in cancer pathways is supported by the observations in which one mutated gene suffices to perturb the function of its corresponding pathway. Multiple mutations require significantly higher energy investments on the part of cancer cells, and are hence selected against. Zhang et al. [30] expanded the ideas behind the concept of MEMo with iMCMC, and provided a framework to integrate mutation data, copy number, and expression information into cancer network weights which they used to identify modules. Dendrix [31] was developed to identify potential driver subnetworks using mutual exclusivity and coverage over a patient cohort, without relying on known network information. It has the potential to facilitate the discovery of new modules. MDPFinder [32] expanded on the overall framework of Dendrix by incorporating gene expression information to ensure that genes in discovered mutually exclusive pathways were also co-expressed. Multi-Dendrix [33] and CoMDP [34] address the limitations of Dendrix and MDPFinder, respectively, by allowing their algorithms to find multiple co-occurring modules. More recently, CoMEt [35] was proposed to address an inherent bias in Dendrix and Multi-Dendrix that resulted in high frequency mutations being significantly more likely to be included in mutually exclusive modules. The previous methods are not without limitations. The most prominent limitations are the size of the modules and the inability to integrate biological insights such as gene expression and gene network interaction in determining driver pathways. Even the most recent method, CoMEt cannot efficiently identify modules consisting of more than 10 genes. Incorporation of biological insights and the ability to identify expansive functional pathways in cancer are needed to improve our understanding of the driving pathways in cancer.

1.2 Personalized approaches to discovering diagnostic cancer subgroups

One of the most useful diagnostic tools in cancer care is the identification of clinically relevant patient subgroups. These molecular subgroups, or subtypes, account for tumor heterogeneity by stratifying patients with different prognoses, variable first sites of metastasis, differential response to targeted therapeutics, and different rates of survival [36]. The end goal of subtyping is to divide patients into strata that will likely respond to tailored cancer treatments. Molecular subtypes in breast cancer have been crucial to our understanding in both the clinical features and treatments in cancer. Subtypes from both DNA/RNAseq and Microarray data serve as prognostic markers that can be used to both predict survival times, relapse times, and other clinical features as well as define genetic markers that can serve as therapeutic drug targets [37].

Breast Cancer (BRCA) has one of the most well-studied molecular subtypes [38]. The earliest molecular subtyping for BRCA used the major hormone receptors in the tumor: Estrogen Receptor (ER) and Progesterone (PR), and the growth factor receptor, Her2 (Her2) [39]. By testing for the receptor presence in these three subtypes, clinicians prescribe treatments that selectively target these receptors and its corresponding signaling pathway. Nevertheless, 18% of BRCA patients do not test positive for any of the three receptors. These patients, called Triple Negative Breast Cancer (TNBC) patients, are associated with poor prognosis, poor survival, and poor response to traditional BRCA therapeutics due to lack of available drug receptor targets [40], [41].

The advent of next-generation sequencing has made it possible to categorize BRCA subtypes through genomic and molecular signature data with the hope to finding new genomic markers that guide novel drug development, especially for the sorely needed TNBC patients. The most prominent of these methods is PAM50 [36], [42]. PAM50 illustrates a list of fifty gene markers whose gene expression serves as features in a median-centered hierarchical clustering, which ended up with five major BRCA subtypes: “Luminal A”, “Luminal B”, “Her2”, “Basal”, and

“Normal-like”. In addition to the PAM50 subtypes, the authors of the Molecular Taxonomy in Breast Cancer International Consortium (METABRIC) found that the PAM50 “Luminal A” and “Luminal B” subtypes could be further divided in significant subtypes using Item Cluster Analysis (iClust) [43]. METABRIC found 10 significant BRCA subtypes. The METABRIC clusters further break down several of the original PAM50 clusters, especially differentiating several types of Luminal A and Her2 clusters. Other methods have defined BRCA subtypes using other types of method such as The Cancer Genome Atlas’s (TCGA) BRCA landmark paper which found five significant clusters using copy number calls through a Non-Negative Matrix Factorization (NMF) [44]. This model used copy number calls exclusively with no gene filter and no incorporation of others molecular signature information such as gene expression. One promising new approach in BRCA subtype detection lies in using driver genes as features to further stratify BRCA subtypes. Specific mutations and copy number alterations have used as factors to identify specific subgroups within Luminal breast cancers [45], and one potential future direction in breast cancer research lies in using driver genes as features in identifying alternate BRCA subtypes which may result in new molecular markers targets that can be used to diagnose and treat BRCA patient populations.

1.3 Personalized drug response prediction in cancer

The computational identification of novel driver genes and pathways and the integration of genomic data to discover alternative subtypes have set the stage to accurately portray the molecular and clinical profiles for a cancer patient. The next step in personalized medicine is to accurately predict a cancer therapy for individual patients in the context of the newfound genomic and clinical diagnostic tools [13]. This is crucial to guide clinicians to assign the most effective therapeutic treatments individual cancer patients to ultimately combat the cancer and improve the quality of life [46][47]. One of the most promising personalized treatment strategies available to physicians is the prescription of drugs that target the driver genes of the patient [11]. For example, a lung cancer patient with an aberrant epidermal growth factor *EGFR* may respond well to a tyrosine kinase inhibitor which inhibits *EGFR* [48] while a breast cancer patient with an aberrant *Her2/ERBB2* receptor may respond well to a monoclonal antibody, Trastuzumab, that targets the *Her2/ERBB2* receptor [49]. The goal of computational methods that recommend drug treatment is to provide a framework which assigns the right targeted therapy to the right patient based off of the patient's molecular signature information.

Modeling the effect of cancer drugs is ripe with many major challenges. On the treatment side, cancer treatments work under a variety of drug mechanisms, each with unique indications, and contraindications which add many confounding variables to the precision and reliability of the prediction of the response of the targeted therapy [50]. Even in targeted therapies, cancer drugs have complex interactions with cell lines in which the interaction between the drugs and the targeted pathways are not well understood in many cases [51]. On the disease side, Cancers are multifactorial genetic diseases that are heterogeneous and operate under different disease mechanisms from patient to patient [52].

The majority of data available for drug response analysis comes from cell lines compendiums such as The Genomics of Drug Sensitivity of Cancer (GDSC) [53] and the Cancer Cell Line Encyclopedia (CCLE) [54]. Cell line information has spawned many of the landmark studies in drug response research. One example comes from the National Cancer Institute's (NCI) DREAM7 project, where contestants predicted the drug response of "hidden" BRCA cell lines using RNA-seq data from training cell lines. Other cell line studies include Dong et al.'s machine learning model which uses Support Vector Machines (SVM) predicting the drug response of GDSC cell lines [55]; CancerDP, a drug prioritization method based on SVM with F-stepping feature selection [56]; A linear model study that calculated the drug response or Lymphoblastic cell lines using a linear model [57]. Most recently, a flagship study modeled drug response predictions through an ensemble method by identifying functionally impactful and unique Cancer Functional Events (CFEs) [58]. However, results from cell line studies is not without drawbacks. Experimental procedure differences between the major cell line compendiums have shown inconsistent drug response when the same drug is treated with the same cell line [59] Additionally, cancer tumors do not reside in a closed system. Tumors react closely with normal cells and the patient's environment [60].

Ideally, a drug treatment model built on real patient data and histories such as one utilizing TCGA data would accommodate these factors; however, such a model would need to be able to handle the added complexity and separate out the important features. Some methods have been developed to model drug effectiveness in drug response. The authors of [61] utilized a linear Ridge-Regression model to bridge the gap using *in vitro* gene expression models to make predictions in cancer patients. While gene expression models have shown a degree of success, gene expression models alone have been found to be insufficient in predicting drug response in some cancers [62]. The IntOGen platform has also built a drug recommendation model based on the proximity of the driving cancer gene to the drug target [63]. The identification of targetable genes was expanded using the EMD model, which identified a list of candidate drivers using integrated gene expression, mutation, copy number and network information with potential drug targets for the drivers [64]. Another method GOPredict [63] integrates both genomic and pathway data to provide a ranked drug list of potential targets [65]. Most recently, Zhang et al. [66] developed a method ElasticNet Regression machine learning method that predicts the clinical response of a drug directly from TCGA molecular signature data using mRNA expression, mRNA expression, methylation, or copy number individually. However, this method has been hampered by several significant limitations. No model presented in their paper predicted drug response with a higher AUC of 0.7 when compared to the actual prediction. Additionally, the lack of a filter for curated cancer genes has led to overfitting due to the incorporation of low-information and redundant variables in the model. In this dissertation, we describe a novel method in drug response prediction and recommendation which addresses the limitations of previous methods by only using high-impact biological features to prevent overfitting as well as integration of multiple types of genomic features to increase the reliability of the model.

1.4 Contribution of the dissertation

While the aforementioned computational methods in driver detection have greatly contributed to our understanding of cancer progression from both an individual gene and a driver pathway perspective. The goals in driver detection addressed in the dissertation are two-fold: 1) the ability to precisely detect individual rare drivers that are potentially obscured by conventional methods and 2) the ability to describe in a biological context the interaction of multiple driver genes working together. The work in this dissertation also seeks to explore potential clinical application. We use the insights highlighted by our driver detection methods as a guide to the identification of novel clinical subtypes using driver genes as important features for classification. Additionally, we use the integrated knowledge from drivers and other genomic sources to develop a new approach which prioritizes and predicts the response of cancer drugs.

In **Chapter 2**, we introduce a method called DawnRank [67] that detects driver genes using data from a single patient sample. By only using data from an individual patient sample rather than a large cohort, we identify personalized, patient-specific drivers. The single patient approach detects drivers regardless of mutation frequency, thereby allowing us to focus on potential rare (infrequent) drivers. DawnRank ranks potential driver genes based on their impact on the overall differential expression of its downstream genes in the molecular interaction network. Mutated genes with a higher ranking are more likely to be drivers. DawnRank has been shown to outperform previous methods in detecting known, biologically-verified driver genes, while also proposing potential novel and rare driver genes. In **Chapter 2**, we explore the biological significance of the DawnRank driver genes, by using the DawnRank scores as the basis for a diagnostic tool to identify subtypes in breast cancer. In this analysis, we performed a consensus clustering on the DawnRank score on genes with mutation and copy number alterations to identify breast cancer subtypes. This method is novel in its application as it clusters BRCA over an integrated dataset of both mutation drivers and copy number drivers simultaneously. Our framework identified five alternative BRCA clusters which we compared to existing BRCA clinical subtypes as well as the established PAM50 gene expression subtype, and we identified potential driver genes that may serve as molecular markers for each of these subtypes.

With respect to driver pathways, while current methods such as Dendrix, Multi-Dendrix and CoMEt all have the ability to identify driver subnetworks/pathways involving multiple driver genes, the aforementioned methods are typically inefficient when applied to large-scale datasets with large values of their relevant parameters. Some of these methods are randomized in nature and no guarantees exist that multiple runs of the methods will produce compatible results. Almost all methods are only able to identify a small number of modules of limited size as cluster sizes are critical algorithmic parameters from the perspective of computational tractability. Most importantly, they have to be redesigned or restructured whenever new biological information is included in the discovery process. **Chapter 3** introduces a novel method called Cancer Correlation Clustering C^3 [68] which addresses the shortcomings of the existing methods. C^3 uses a new agnostic optimization framework specifically developed and rigorously analyzed for the driver discovery task that allows for the integration for flexible biological data from multiple sources such as coverage, mutual exclusivity, expression data and network pathway information. C^3 has low computational cost compared to previous methods, and it allows for adding relevant problem constraints while retaining good theoretical performance guarantees.

Chapter 4 of the dissertation introduces a novel method, Scattershot. Scattershot addresses several of the limitations of previous methods in order to develop a drug prioritization tool that assigns the right drug to the right patient. The data that Scattershot uses is from real patient, TCGA data, rather than the closed-system *in vivo* cell line studies. Scattershot models the drug recommendation problem as a multilabel machine learning problem [69] in which we develop ensemble classifiers from multiple genomic sources such as mutation, expression, copy number, and pathway-level information as well as clinical variables. Scattershot uses the multilabel framework to build binary classifiers that predict the drug response of an individual drug while at the same time, aggregating multiple pairwise binary classifiers comparing pairs of drugs in a drug list to prioritize the drug rankings. We compared Scattershot's treatment predictions in cancer patients to the treatments actually assigned to the patient by physicians and we found Scattershot's predictions are mostly in line with the physician recommendation, outperforming the previous models.

Chapter 2: Integrated Mutation and Copy Number Driver Analysis Identifies Molecular Subtypes of Breast Cancer[1]

2.1 Introduction

Breast cancer remains the second leading cause of cancer related death in women each year [1]. Breast cancer is a heterogeneous disease with multiple subgroups. Patients in different subgroups have different prognoses, variable first sites of metastasis, and differential response to targeted therapeutics. Currently, the estrogen-independent breast cancers have the worst prognosis, fewest therapeutic options, and no currently approved targeted therapies. Standard of care includes chemotherapy and radiation therapy [38]. Identification of targetable drivers in breast cancer could provide novel therapeutic targets.

The discovery of the driving events in cancer has been the subject of years of research in personalized medicine [2], [19], [21]–[24]. However, these methods, while providing a starting point in identifying common drivers, are often challenged by limitations in identifying rare, patient-specific drivers. Most of the methods listed above select drivers based on categorize mutations based on recurrence, i.e., the most commonly occurring mutations are more likely to be drivers [16], [17], and thus are disadvantaged due to the fact that they require a large number

[1] The description of the DawnRank method in this chapter is based on a published paper in *Genome Medicine* and is referred to in the dissertation as "J. P. Hou and J. Ma, "DawnRank: discovering personalized driver genes in cancer," *Genome Med.*, vol. 6, no. 7, p. 56, 2014"

of patient samples to generate reliable results and lack the ability to discover rare and patient-specific drivers. Other methods such as PARADIGM-Shift are designed to determine drivers in small pathways and often require detailed previous knowledge of specific pathways and focus genes to operate effectively. New methods are needed to identify novel and rare drivers when we do not have much prior knowledge of the tumor.

It is now acknowledged that individual tumors of the same type are highly heterogeneous and have diverse genomic alterations [70], [71]. This stems from the "long-tail phenomenon" which states that cancer mutations are characterized by a small number of frequently mutated genes and a large number of infrequently mutated genes [72], [73]. Discovering rare drivers in the long tail of genetic alterations remains difficult. Therefore, we urgently need methods to assess the impact of patient-specific and rare mutations from individual tumor samples in order to elucidate personalized molecular drivers.

Large efforts to identify the genetic underpinnings causing breast cancer have led to unprecedented amounts of both DNA and RNA genomic data. However, the significance of these alterations often is not well understood. Copy number alterations (CNA), are known to be an early, common, and critical factor in the development of breast cancer. It is much more common across the TCGA cohort and has been shown to be an early event in the development from normal breast to pre-invasive cancer to invasive and metastatic tumors. CNAs, in conjunction with mutation-based alterations have been used to define and distinguish cancer subtypes in the past. Mutation and Copy Number alteration markers have used as factors to identify specific subgroups within Luminal breast cancers [45]. These driver genes present a unique perspective in stratifying breast cancer subgroups [74]. Patient subgroups can be treated using targeted therapy directly aimed at the driver genes that that define the subgroup [75]. The implications of this are especially important in treating Triple Negative Breast Cancers (TNBC) which are defined by their lack of targeted therapeutic targets and poor overall prognosis [76]. Therefore, an integrated approach that identifies patient subgroups based on their driver genes may provide alternative targets in breast cancer targeted therapy.

The identification of driver-defined subtypes requires a reliable method to identify the driver genes in a given cancer patient. One method that identifies personalized driver alterations in cancer is DawnRank [67]. DawnRank detects driver genes using data from a single patient sample. By only using data from an individual patient sample rather than a large cohort, we identify drivers in a personalized fashion. DawnRank allows for the integration of mRNA gene expression, DNA mutations, and DNA copy number data. The proportion of drivers from CNA as compared to mutation is not well known. Additionally, it is largely unknown if drivers on an individual tumor level are consistent within and across subtypes or private to a tumor. By applying DawnRank to TCGA breast cancer data, an understanding of the biology driving breast cancer can be explored with the hopes of identifying alternative, tractable therapeutic targets especially in estrogen-negative breast cancer.

2.2 Results

We applied DawnRank to the TCGA datasets. For evaluation purposes, we applied DawnRank to 512 glioblastoma multiforme (GBM) samples, 504 breast cancer (BRCA) samples, and 572 ovarian cancer (OV) samples in TCGA. The datasets we used in this work include gene expression and coding-region mutation data for three cancer types generated by TCGA [44], [77], [78]. The data was accessed on May 20, 2013. The mutation data we used included non-synonymous point mutations and insertions and deletions (indels) in coding regions. We first showed that DawnRank outperforms two pathway-based methods DriverNet and PARADIGM-Shift. We then used the results of DawnRank to determine both potential novel drivers (new genes mutated frequently), and more importantly, potential rare and personalized drivers that previously could not be assessed by other methods. The discussion of potential novel and rare driver alterations as well as an in-depth comparison of DawnRank to other methods can be found in the DawnRank paper [67].

We then applied DawnRank to discover BRCA subtypes. We developed a framework for an integrative analysis of somatic mutations and copy number alterations that identified five breast cancer molecular subtypes within a TCGA breast cancer cohort of 351 patients weighted to be representative of a BRCA population using known subtypes. An overview of our method is shown in Figure 2. For this study, we utilized the Cancer Genome Atlas [44] as the discovery dataset and the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) [43] as a validation set. DawnRank was used to identify candidate driver genes. ConsensusClusterPlus [79] was then applied to discover subtypes of breast cancer based on driver alterations with a classification to nearest centroids classifier (ClaNC) [80]. Consensus Cluster Plus identifies stable clusters by assigning a patient into a cluster through a thousand runs of randomized patient sample, while ClaNC provides a feature compact method that predicts the class of a sample using the fewest features possible. Association with PAM50 [36] subtype and clinical characteristics were assessed. Finally, we defined subtype-specific candidate drivers with both an Analysis of Variance (ANOVA) and significance analysis of microarray (SAM) [81]. The resulting alteration-based clusters yielded five potential BRCA subgroups with strong associations to the PAM50 gene expression clusters and the ER/PR/Her2 clinical classifications. We identified multiple sub-chromosomal altered hotspot regions encompassing candidate and subtype-specific, potentially new breast cancer drivers.

2.2.1 Comparison of DawnRank to previous methods

We evaluated the performance of DawnRank's ability to identify known drivers and compared it with DriverNet and PARADIGM-Shift. As mentioned above, we utilized CGC as an approximate benchmark of known drivers. Here, we implicitly assume that all non-synonymous mutations in driver genes are potential driver mutations if they are selected by a method. We performed two separate comparisons. (1) We compared DawnRank to DriverNet over a large network in order to evaluate the performance of the two methods using a large human interaction network (which PARADIGM-Shift is not able to work with practically). (2) We also compared DawnRank to PARADIGM-Shift and DriverNet over a smaller, but well-annotated gene network based on KEGG in order to determine the effectiveness of the three algorithms in smaller networks. The network used in the first comparison was the same network described earlier. The network used in the second comparison was a smaller network built from the aggregation of the KEGG cancer pathways with 1,492 gene nodes and 8,070 edges. We ran DriverNet version 1.0.0, defining a differentially expressed gene using their default settings of 2 standard deviations (http://www.bioconductor.org/packages/release/bioc/html/DriverNet.html), and we ran

PARADIGM-Shift version 0.1.9 using the suggested global-rank transformation for expression data (http://sysbio.soe.ucsc.edu/paradigm/tutorial/). To facilitate the comparison, we applied the Condorcet rank aggregation (see Methods) for the DawnRank scores based on individual patient samples to provide the consensus population-level driver scores. For each comparison, we used the following three measures (Precision, Recall, and F1-Score).

Precision, recall and F1 scores were based on the top *N* genes. We first evaluated the performance between DawnRank and DriverNet. In general, DawnRank outperforms DriverNet in all three cancer datasets with respect to CGC (Figure 3). Although DriverNet performs comparably in ranking the top genes in GBM, it has poorer performance in OV and BRCA. A potential explanation of the difference may lie in the total number of mutations in the three cancer datasets. GBM had 5,478 mutations over 599 genes, while OV had 13,520 mutations over 4,968 genes and BRCA had 11,900 mutations over 5,205 genes. The numbers indicate that there may be more passenger mutations in BRCA and OV and DawnRank is less affected by noise than DriverNet. An illustration of this is DriverNet's ranking of the gene *TTN* as a top 5 driver in both BRCA and OV. *TTN* is the longest gene in the human genome and recent TCGA analysis has suggested that that higher mutation rate in *TTN* is likely to be artifacts [78]. *TTN* was not ranked among the top 60 genes in any cancer according to DawnRank. We then evaluated the performance of DawnRank, PARADIGM-Shift, and DriverNet using the smaller KEGG network. Overall, DawnRank outperforms both DriverNet and PARADIGM-Shift in terms of precision, recall, and F1 scores using CGC as a standard (Figure 4) or the Pan-Cancer results as a standard. In BRCA, although some known drivers such as *TP53* and *ATM* were detected by multiple methods, DawnRank detected important known driver genes in the top 10 such as

CDH1, *PIK3R1*, and *BRCA1* in breast cancer which were not detected by either PARADIGM-Shift or DriverNet as top ranking drivers.

2.2.2 Identification of driver-based subtypes

In order to define the genetic drivers of breast cancer through an integrated analysis of gene expression, copy number, and mutation, we applied DawnRank to BRCA using a custom balanced cohort in TCGA, which samples a cohort where the proportion of each PAM50 subtype matches that of the population. Genes with cohort-wide DawnRank p-values ≤ 0.05 were considered for clustering to define driver-subtypes. 65 copy number altered genes and 38 mutated genes were significant across the cohort (Figure 5A). These genes are selected by DawnRank to maximize pathway impact and driving potential. The copy number altered genes cluster along 1q gains, 8q gains, 11p loss, and 16p loss (Figure 5C-F). Running ConsensusClusterPlus on TCGA tumors with 1000 iterations of ConsensusClusterPlus with 80% resampling of genes and samples, we identified five as the ideal number of clusters by observing the maximum cophenetic correlation when testing $k = 2$ to $k = 10$. We compared the clusters after 25 different runs of ConsensusClusterPlus and observed consistent clustering results with a pairwise Rand Index of 0.97. Centroids for each subtype were built with ClaNC classifier with the feature parameter of 11, causing the least amount of misclassification in TCGA.

2.2.3 Driver-subtypes reflect the genomic heterogeneity of breast cancer

To classify the driver-subtypes in the context of previously defined clinical predictors and molecular taxonomy, we examined the correlation with PAM50 subtype and known clinical predictors. Plotting Pearson residual for each driver-subgroup, we observe subgroups highly correlated to the PAM50 molecular subtypes (Figure 5B). The first cluster (red) is weakly positively correlated to both Luminal A and B PAM50 subtypes. The second subgroup demonstrates strong association with Basal-like and weakly positive correlation with Her2. The third cluster demonstrates an association with Luminal B subtype and a weak association with Her2-enriched PAM50 subtype. The fourth alteration cluster is consistent with previous work demonstrating shared genomic features between Luminal B and HER2-enriched PAM50 subtypes [82]. The two remaining alteration clusters, “Luminal A1” and “Luminal A2”, are both strongly associated with the Luminal A PAM50 subtype.

Since the mutation data is used for clustering, it is unsurprising that multiple mutation markers have strong associations with the alteration clusters which we confirmed using the Chi-square test for association. We tested mutation status of *TP53, PIK3CA, GATA3, MAP2K4*, and *MAP3K1* due to their previously defined significance as drives by MutSig [44]. *TP53* mutations are highly correlated with the Basal/Her2 and Luminal B/Her2 alteration clusters (p-value < 2e-16). PIK3CA and GATA3 mutations are highly associated with the Luminal B and Luminal A2 alteration clusters (p-value < 2e-16 and p-value = 0.043) [83]. These results confirm previously identified mutational and PAM50 subtype associations [44].

2.2.4 Driver-subtypes correlate with Estrogen Receptor status, Progesterone status, and tumor stage

We next examined the correlation of each driver-subgroup to known clinically predicted values including: estrogen receptor (ER) status, progesterone receptor (PR) status, and Her2 receptor (Her2) status, tumor stage (T), node status (N), metastasis status (M). Clusters demonstrated significant association with 3 of the 6 tested clinical and mutation features, including ER, PR, and T. Interestingly, Her2 did not significantly associated with any subgroup. As expected, the three Luminal alteration clusters (Luminal A1, Luminal A2, and Luminal) demonstrated positive signals with ER and PR (Chi-squared test for association p-value < 2e-16). Tumor Stage associated with Luminal A2 group p-value = 0.001. These results further validate our classification scheme to recapitulate known clinical markers and biological subtypes.

2.2.5 Subtype-defining drivers

To define subtype-specific drivers, each driver (68 CNAs and 38 mutations) are tested by ANOVA for overall variation among the subgroups, and one class against all others to define the driver-subgroup specific to the driver with the significance analysis for microarray (SAM). For large, focal, CNAs we limited candidate drivers to the top two gene markers to represent the focal length alteration. ANOVA identified 11 copy number altered genes and 8 somatically mutated genes significantly associated with one subtype (Figure 6A). SAM analysis identified 37 significantly altered genes within the driver-subtypes under a false discovery rate of 0.05 (Figure 6B). Comparing results across both statistical analyses, we found five significant driver genes (*ARF1, AKT3, PIK3CA, ATM* and *BCAR1*) across both the ANOVA and SAM analyses.

For the Luminal driver-subtypes, we identified consistent drivers gained on chr1q across all three subgroups; however, subtype-specific candidate drivers are also present. *BIRC3* (chr11) copy number loss is specific to the Luminal alteration cluster. 77.6% of the Luminal subtype has a copy number loss at *BIRC3*, and 66.0% of all *BIRC3* alterations occur in the Luminal subtype. We visualized the network impact of *BIRC3* within the Luminal driver-subtype (Figure 6C) compared to the network in the other driver-subgroups (Figure 6D). We identified a large downstream down-regulation of several genes within the network specific to the Luminal driver-subtype (Figure 4C). In particular, *PAK1* is the most distinct downstream differentially expression gene within the *BIRC* network for the Luminal subgroup but not in the other subgroups. *PAK1* is two degrees of separation from BIRC3 and is a known oncogene that activates *MAPK* and *MET* signaling in cancer [84]. Our results suggest that *PAK1* is highly overexpressed in Luminal subtype, which may be due to the deletion of upstream *BIRC3*.

2.2.6 Subtype-specific driver CNAs in four chromosomal hotspots

Specific regions of the genome are known to be commonly gained and lost within breast cancer. The drivers at these locations, however, are not well understood. Whether subtype-specific differences within each region of the genome selects for different driver genes is not known. We explored four known regions with a high prevalence of copy number alterations in BRCA to define subtype-specific drivers within each region including: 1q amplification, 8q amplification, 11q deletion, and 16q deletion (see Figure 5C-F).

The chr1q amplification is one of most frequently occurring CNAs in breast cancer [85] [86]. Using a Chi-squared test, chr1q amplification is significant in the Luminal, Luminal A1, and Basal/Her2 clusters (p-value < 2e-16). In each of these clusters, chr1q copy number gains occurred in more than half of the samples with 76.2% alteration rate in samples in these three clusters as compared to only 12.2% alteration rate in the other two clusters. The DawnRank drivers in chr1q significant in these three driver-subgroups include *AKT3* (25/335, 7.1%) and *NCSTN* (23/335, 7.1%). AKT3 is an integral member of PIK3CA signaling pathway, responsible for many vital cell functions such as growth and apoptosis [87]. *NCSTN*, a recently identified candidate target for altered Notch signaling activity within Basal-like breast cancers, provides the structural support for Notch signaling and is required for GSC cleavage of Notch receptor [88]

Chr8q is a frequently occurring copy number gain with subtype defining features [89]. Chr8q amplification is significant in the Luminal B/Her2 (95.9% alteration rate) and Basal/Her2 (67.0% alteration rate) but not in the Luminal A related subtypes (20.2% alteration rate; Chi square p < 2e-16). LumB/Her2 is significant for MYC (driver in 4.6% samples, SAM p-value=5.7e-8) and NCOA3 (driver in 2.8% samples, SAM p-value=1.47e-7). *MYC* is a key regulator of cell growth, proliferation, metabolism, differentiation, and apoptosis [90]. *NCOA3* is a nuclear receptor that is known to be overexpressed in breast cancer and involved in estrogen-mediated cancer cell proliferation [91].

Chr11q loss is primarily defined by the Luminal alteration cluster. Subtype defining driver genes in chr11q include *BIRC2* and *BIRC3* (drivers in 5.4% samples). These genes function by inhibiting apoptosis by binding to tumor necrosis factor receptor-associated factors *TRAF1* and *TRAF2* [92].

The fourth major hotspot is chr16q loss. Previously associated with Luminal A breast cancer [93], chr16q loss is consistently associated with our three Luminal driver-subgroups. This region is marked by *CDH1* (driver in 16.5% samples) and *TRADD* (driver in 4.0% samples). *CDH1* has prominent role in epithelial differentiation and may play a role in tumor differentiation and metastasis [94] . *TRADD* codes for an adaptor molecule that interacts with *TNFRSF1A/TNFR1* and mediates programmed cell death signaling and *NF*-kappaB activation [95]. *TRADD* also interacts with key drivers *TRAF* and *CASP3* genes.

2.2.7 Validation using METABRIC dataset

Utilizing gene expression, recently published mutation data, and copy number data from the METABRIC dataset (n = 339 patients), we calculated the DawnRank scores for each tumor. We then applied ClaNC classifier on the 103 driver genes identified in the TCGA cohort to classify METABRIC samples into the 5 subtypes. We associated the 5 METABRIC driver-subtypes with PAM50 subtypes (Figure 7A). Similar to the TCGA alteration clusters, four of the METABRIC subgroups significantly associate with PAM50 subtypes (Chi-squared test for p-value=1.2561e-10). The fifth subtype is associated with both Her2 and Luminal B. A Chi-squared based Goodness-of-Fit test confirmed that the distribution of between the TCGA and METABRIC results share the same distribution as compared to the PAM50 results (p-value=0.2128). These results from the ClaNC classification of METABRIC driver-subgroups validate the TCGA driver-subgroups.

Using the METABRIC dataset, we compared the overall survival time from the alterations clusters from METABRIC subtypes (Figure 7B) to that of PAM50 (Figure 7C). Both the alteration clusters and the PAM50 subtype demonstrate significant differences in survival within each group (p-value = 0.0251 and p-value = 0.0186), with the Basal-associated subgroups showing worsened overall survival.

2.3 Discussion

It is now acknowledged that individual tumors of the same type are highly heterogeneous and have diverse genomic alterations. Therefore, we urgently need novel methods to identify patient-specific and rare drivers from individual tumor samples in order to elucidate personalized molecular mechanisms in different types of cancer. The goal of DawnRank is to integrate mutation data, gene expression, and network information to discover drivers in a personalized manner. We applied DawnRank to a large number of TCGA samples. By comparing to previous studies, our results demonstrated the effectiveness of DawnRank: (1) Despite its single-patient scope, DawnRank detects common and known drivers with as much or more precision than existing methods. (2) DawnRank can identify rare and novel genes that are potential drivers to specific patients. We believe this method will complement existing driver identification methods and will help us discover potential personalized drivers. The application to breast cancer subtypes further demonstrates that the rare drivers predicted by DawnRank provides new insights into the molecular explanations of cancer subtypes with higher tumor heterogeneity.

Using DawnRank, we present a new and different classification of breast cancer and subtype-specific driver analysis of both copy number and mutation data. Utilizing both TCGA as a test set and METABRIC as the validation set, we demonstrate five robust driver-subtypes. Three subtypes correlate highly with the Luminal A subtypes, one with Basal/Her2, and the final with LumB/Her2. Additionally, the subgroups correlate with known clinical markers such as the estrogen and progesterone receptors with the Luminal subtypes, TP53 mutation in the Basal/Her2 subtypes, and worsened overall survival in the Basal/Her2 subtype.

Known hotspots of copy number alteration in breast cancer, including 1q amplification, 8q amplification, 11q loss, and 16q loss, demonstrate subtype-specific differences. Chromosome 11 loss is specific to Luminal subtype including *BIRC3* and *CBL* loss. *BIRC3* network analysis demonstrates loss of *BIRC3* and a strong up-regulation of *PAK1*, a known oncogene downstream of *BIRC3*. A second interesting result is the loss of *CBL*, an E3 ubiquitin protein ligase which recognizes known oncogenes including *FGFR2, KIT,* and *PDGFRA*. *CBL* loss has not been previously described in the context of Luminal breast cancer. Targeting of *FGFR* family members with dovitinb has been showing to be effective in a small cohort of breast cancer patients in Phase 2 trial [96]. *CBL* loss could be a second marker for *FGFR* sensitivity in patients who lack FGFR amplification but still may be dependent on this pathway.

Integrating gene expression to evaluate the impact of a genomic alteration allows for novel driver identification such as the loss of *BIRC3* and *CBL* playing major roles in defining the Luminal A subtype. Novel therapeutic targets are desperately needed for breast cancer patients, especially triple negative (TNBC) patients who lack estrogen receptor (ER), progesterone receptor (PR) overexpression or amplification of the human epidermal growth factor (HER2). In the metastatic setting, TNBC patients often do not benefit from the addition of systemic therapy. The paucity of systemic, targeted anti-cancer therapies in these patients begs for new treatment options. Improving our understanding of the underpinning molecular drivers of this subgroup are necessary to develop better targeted and more effective therapies.

Future *in vitro* and *in vivo* confirmation will be needed to confirm our findings. We are also limited by the biases in the curated pathway used to evaluate the networks. Finally, assessment of these drivers through both therapeutic selection (comparing pre-treatment and post-treatment samples) and the selection of these drivers through the metastatic process are needed. Metastases are the leading cause of cancer related deaths, and often a small percentage clone in the primary causes seeding of distant metastases. Thus, drivers identified in the primary may not be the main causes of metastasis or the genes that need to be targeted to halt metastatic progression. Future studies on large cohorts of matched primaries and metastases will soon answer these questions.

The heterogeneity of breast cancer has long been described and understood from a clinical, histopathologic, and molecular lens. Through DawnRank, we were able to capture this heterogeneity and assess novel molecular drivers for each breast cancer subtype. Future functional studies confirming the role of these drivers in a subtype-specific manner are needed in order to lead to novel therapeutic development. Incorporation of mutations, copy number alterations, and gene expression confirm the importance of evaluating not only mutations but also copy number variations in understanding the underlying biology driving breast cancer.

2.4 Methods

2.4.1 DawnRank algorithm

Our method ranks genes according to their impact on the perturbation of downstream genes, i.e., a gene will be ranked higher if it causes many downstream genes, directly or indirectly in the interaction network, to be differentially expressed. DawnRank views the gene network as a directed graph. We adopted the random walk approach used in PageRank [97], [98] to model this process iteratively.

In DawnRank, a gene will possess a higher impact score (i.e., rank) if the gene is highly connected to differentially expressed downstream genes (directly and indirectly connected). Driver genes tend to display a high-degree of connectivity within the gene network [99], [100]. For example, using the number of outgoing edges alone, known driver genes as classified by the Cancer Gene Census (CGC) [101] have a mean and median of 31.45 and 12 outgoing edges, respectively, whereas genes not typically classified as drivers (not in CGC) have a mean and median of 17.73 and 3 outgoing edges, respectively. The higher number of outgoing connectivity of known driver genes suggests that the PageRank model would be appropriate to prioritize driver genes based on their impact in the gene interaction network. PageRank has had several adaptations in genomics. GeneRank utilized PageRank to rank the importance of genes in a molecular network [102]. PageRank derivatives (such as SPIA [103]) have also been used to analyze pathway-level importance. More recently, it was utilized to predict clinical outcome of cancer patients based on gene expression [104] and to assist subtype identification [105]. Such approaches also show similarity in nature to modeling network impact as a heat diffusion process as used in HotNet [106] and TieDIE [107]. DawnRank builds on the original PageRank algorithm by providing a way to model a network's directionality with more stable rankings by utilizing dynamic damping factors (see below).

DawnRank views the gene network as a directed graph. Let *N* be the number of nodes (in our case, genes) in the directed graph, and *A* be the adjacency matrix representation of the graph, a 0-1 matrix (if node *i* links to *j*, then $A_{ij} = 1$). Note that the current 0-1 adjacency matrix can be naturally extended to consider weighted edges to further distinguish different gene-gene interactions.

We define the rank of each gene iteratively:

$$r_j^{t+1} = (1 - d_j)f_j + d_j \sum_{i=1}^{N} \frac{A_{ji} r_i^t}{deg_i}, 1 \leq j \leq N \tag{2.1}$$

r^t is the rank in the t^{th} iteration. The output of the rank describes a gene's overall impact on the network: the higher the rank, the higher the impact of the gene. d is the damping factor, a parameter representing the extent to which the ranking depends on the structure of the graph. In DawnRank, the damping factor is individualized based on gene connectivity (discussed below). f is the prior probability of the gene which we set to the absolute differential expression. The absolute differential expression is the absolute value of the difference of the log scale tumor and normal expression values. The $deg_i = \sum_{j=1}^{N} A_{ji}$ is the in-degree of i, or the number of incoming nodes to i. This differs from the original PageRank definition of deg_i, which was the out-degree of i. A webpage's PageRank is dependent on the rank of the webpages that link to it (incoming edges), whereas our gene's rank is dependent on the rank of the genes that it links to (outgoing edges).

The zero-one gap problem refers to the potential pitfall that assigns biased ranks to some nodes [108]. When trying to rank nodes with 0 incoming edges, known as "dangling nodes", the deg_i will be 0, arising to a divide-by-zero error. In our real gene network data, 15.5% of all genes do not have any incoming edges. The initial PageRank algorithm attempts to handle the problem by setting the damping factor to be 0 for such genes, while using the damping factor 0.85 for all other nodes. If we use this approach, the ranks of genes with no incoming edges will be based solely on its differential expression (and not the network structure). However, this correction in itself causes a large gap in the damping factor for genes with 0 and 1 incoming edge This large gap in the damping factor can cause a drastic change in the ranking of the gene when an incoming edge is added to the gene which in turn may cause unstable rankings [108]. An unstable ranking system is especially concerning to gene network data, as it is still not a complete representation of all interactions among genes [109]. Therefore, small modifications and additions to certain gene interactions may significantly alter the rankings of potential drivers. To address this problem, we utilize dynamic damping factors [108], where each gene possesses an individualized damping factor based on the number of incoming edges to that gene (Eq. 2). As the number of incoming edges increases, the damping factor gradually rises to incorporate more connectivity information into the ranking of the gene, therefore no large gap is observed from 0 in-degree and 1 in-degree.

$$d_i = \frac{deg_i}{deg_i + \mu} \tag{2.2}$$

The parameter μ follows a Dirichlet prior trained from maximizing the values of μ over 100 random samples. We selected the μ value of 3 because it had the highest average DawnRank scores for known drivers in CGC. Overall, the dynamic damping factor mitigates the large change in the damping factor in nodes with 0 and 1 incoming edges by gradually increasing the damping factor as the gene's in-degree increases, thereby creating more reliable and more stable rankings. We also show that DawnRank performs more reliably with a dynamic damping factor than a static damping factor on the TCGA datasets.

In addition to the iterative version of DawnRank, the method can also be presented in matrix form:

$$r_{t+1} = (1 - d)f + dM\times r_t \tag{2.3}$$

where r_t, d, and f are $N\times1$ matrices to represent the rank, gene-specific damping factor, and the gene expression, respectively, and M is the transition matrix defined by:

$$\boldsymbol{M} = \begin{bmatrix} \frac{A_{1,1}}{deg_1} & \cdots & \frac{A_{1,n}}{deg_n} \\ \vdots & \ddots & \vdots \\ \frac{A_{n,1}}{deg_1} & \cdots & \frac{A_{n,n}}{deg_n} \end{bmatrix} \tag{2.4}$$

DawnRank converges when there is no longer a significant update in the ranks. This is when the magnitude of the difference of the ranks between time $t + 1$ and the previous time point t falls below a small ε, which we set to 0.001, the same value suggested by [108]. DawnRank also stops when no solution is present after a maximum number of iterations, which we set at 100. In practice, DawnRank always converges for any reasonable μ between 0.01 and 20 within 20 iterations. Nonetheless, there are corner cases at low damping factors ($\mu < 10^{-10}$) where DawnRank either does not converge or converges very slowly.

2.4.2 Rank aggregation for population rankings of drivers

To aggregate the rankings of genes from individual patient samples to determine the most impactful drivers in a population (e.g., known drivers for the same type of cancer or a specific sub-type), DawnRank applies a modified version of the Condorcet method [110]. The Condorcet method is a voting scheme in which “voters” vote for the best “candidate” by submitting a rank-ordered list of candidate preferences. The list of preferences is allowed to be either partial or full. The Condorcet method then selects a winning candidate by comparing every possible pair of candidates *A* and *B* and determining a “winner” by comparing the number of voters that preferred *A* to *B* and vice-versa. We applied the Condorcet method to the personalized rankings of genes to determine aggregate ranking of genes in a patient population.

Although the Condorcet method is built to handle partial voting lists, one difficulty of implementing the Condorcet method is the lack of patient samples that possess the commonly mutated genes. Many pairwise comparisons are missing for many gene combinations due to the lack of patients that have mutations in both genes simultaneously. However, since DawnRank can output a ranking as an impact score for all genes regardless if a gene is mutated, we evaluated pairwise comparisons of two genes based on patients with a mutation in at least one of the two genes. This approach avoids the use of non-mutated gene comparisons to calculate the aggregate score of genes, as the objective of DawnRank is to determine the altered genes that are the most impactful. However, since mutation recurrence is an important factor in detecting common drivers, we also implemented a penalty heuristic, δ, a number between 0 and 1 in our approach to lower the ranking of a gene in a pairwise comparison that is not mutated. This penalty allows us to rank aggregate frequent drivers based on both impact and frequency.

$$\text{PairwiseWinner}(A, B) = \begin{cases} A & \text{if } \delta(A) \times Rank(A) > \delta(B) \times Rank(B) \\ B & \text{otherwise} \end{cases} \tag{2.5}$$

where

$$\delta(A) = \begin{cases} \delta & \text{if } A \text{ is NOT mutated} \\ 1 & \text{if } A \text{ is mutated} \end{cases} \tag{2.6}$$

We used the output from DawnRank, which we converted to percentile rank format, to represent the ranking of the gene. The penalty heuristic lowers the value of a non-mutated gene when comparing it against a mutated gene. This heuristic serves as both a means to prevent a rare mutation that is impactful in one patient from winning all pairwise comparisons (akin to a candidate winning just because one and only voter that voted for it ranked it higher than any other candidate) and to prevent a low impact, high frequency mutation from winning a pairwise comparison against high-impact genes that are not frequently mutated (akin to an unpopular candidate winning just because many voters had a low-preference vote for that candidate). We selected δ by running DawnRank over 100 random patient samples for various instances of δ between 0 and 1 and calculating the precision with respect to CGC genes. We found δ to be 0.85.

2.4.3 Patient sample selection for BRCA subtype analysis

We selected gene expression and copy number data from 871 TCGA samples and 1,992 METABRIC samples. The gene expression is converted into a Z-score, and segmented CNAs are converted into a discrete copy number gene matrix. Significant copy number altered segments with segment means greater than 0 are assigned 1 while significant segments with segment means less than 0 are assigned -1, while all other regions are assigned 0. Using the hg19 gene annotation, genes that are completely encompassed within a segment (based on genomic location) are such that the segment's discrete copy number value and all other genes are assigned 0. DawnRank mutation scores are further distinguished with mutations in oncogenes represented as positive values and tumor suppressors as negative values. A gene mutation matrix is created by assigning -1 to mutations in known tumor suppressors and 1 to mutations in known oncogenes (based on publically available OncodriveRole data) and the value 0 is assigned to all others [111]. Overall survival data is calculated up to 10 years and plotted using a Kaplan-Meier survival curve. Patient samples with greater than 10-year survival are censored at the 10-year mark. An ANOVA analysis is performed to test for significant difference in survival within a patient group. We selected 500 samples from each TCGA and METABRIC. To keep the relative distribution of PAM50 subtypes consistent between the two datasets, we randomly selected TCGA samples based on the average distribution of PAM50 subtypes within the METABRIC cohort. The composition of samples based on the PAM50 molecular subtypes: 19.3% Basal, 10.9% Her2, 39.5% Luminal A, and 30.3% Luminal B (Normal-like breast cancer not included).

2.4.4 Alteration based subtype classification using consensus clustering

During clustering, ConsensusClustersPlus was used to partition the samples and features (driver genes), and builds an unsupervised hierarchical cluster from that particular data subset. Through iterations, a final agglomerative hierarchical consensus clustering using distance of 1-consensus values is completed and pruned to *k* groups. ConsensusClusterPlus is run on $k = 2$ to $k = 10$ groups with sample distances calculated using the Pearson distance over 1,000 iterations. To ensure that ConsensusClusterPlus rarely samples a subset where a patient has no driver alterations (which makes the Pearson distance calculation yield undefined numbers), we trimmed the TCGA and METABRIC datasets to only include samples with at least 5 driver alterations (TCGA n=351 and METABRIC n=339). Since ConsensusClusterPlus is not deterministic, we used 1,000 iterations to minimize the misclassification rate between different runs of ConsensusClusterPlus to less than 10%. Each iteration sampled 80% of samples of the total dataset and the corresponding pairwise misclassification rate of only two single iterations of the sample was 22%. We also sampled 80% of all features (all common drivers in TCGA and METABRIC) compared to only TCGA drivers (53/65=81%).

2.4.5 Validation of the classifier

ClaNC is a custom implementation of Linear Discriminant Analysis (LDA) that selects for features using regular *t*-statistics to account for class difference given a number of features and classes. Using a 5-fold cross-validation approach on misclassification, ClaNC calculates both the number of classes and the number of transformed features. We used ClaNC in the METABRIC validation section, and we used TCGA DawnRank alteration clusters through

ConsensusClusterPlus results as training with METABRIC data as testing. We found that ClaNC works optimally at reducing misclassification when *k*=5 at 11 transformed features with the misclassification rate of 0.223. In addition to the optimal parameter setup of the supervised ClaNC classifier, we also determined that 5 alteration subtypes were optimal in reducing misclassification, and thus 5 classes were selected.

FIGURES

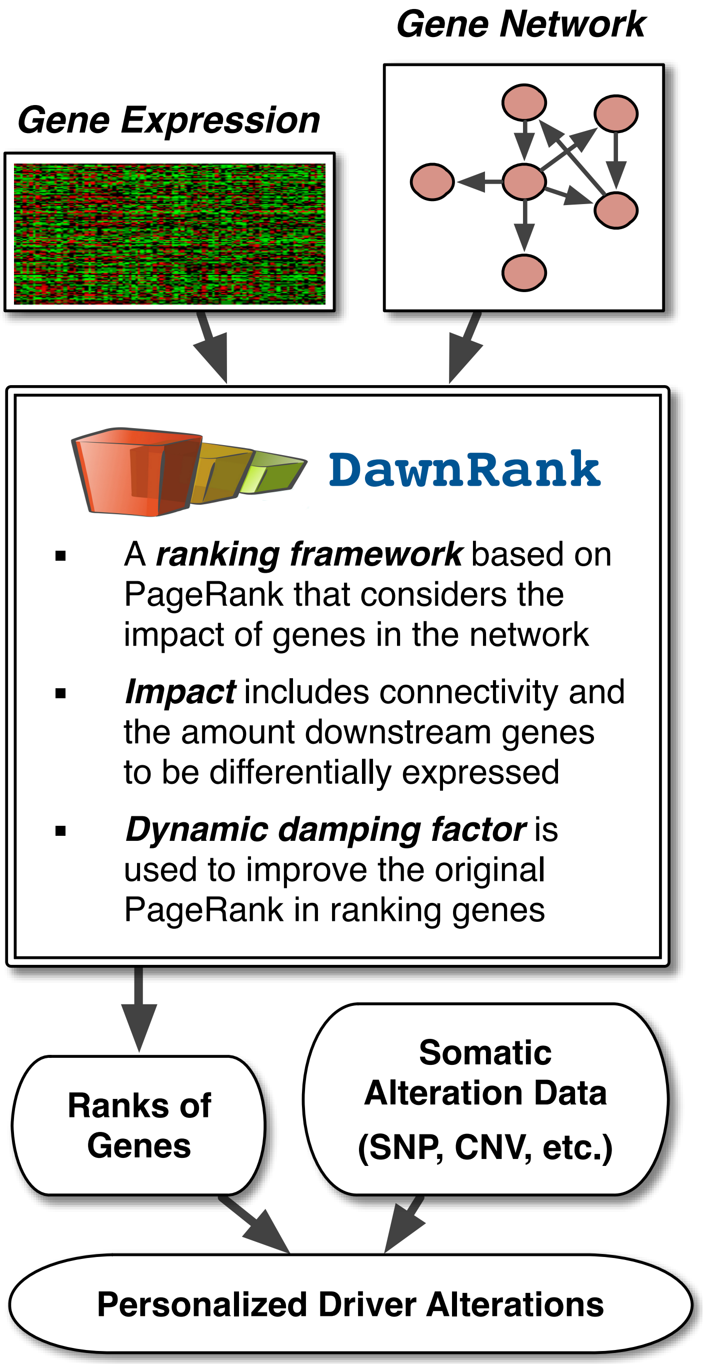

Gene Network
Gene Expression
DawnRank
A *ranking framework* based on PageRank that considers the impact of genes in the network
Impact includes connectivity and the amount downstream genes to be differentially expressed
Dynamic damping factor is used to improve the original PageRank in ranking genes
Ranks of Genes
Somatic Alteration Data (SNP, CNV, etc.)
Personalized Driver Alterations

Figure 1: Overview of the DawnRank Method

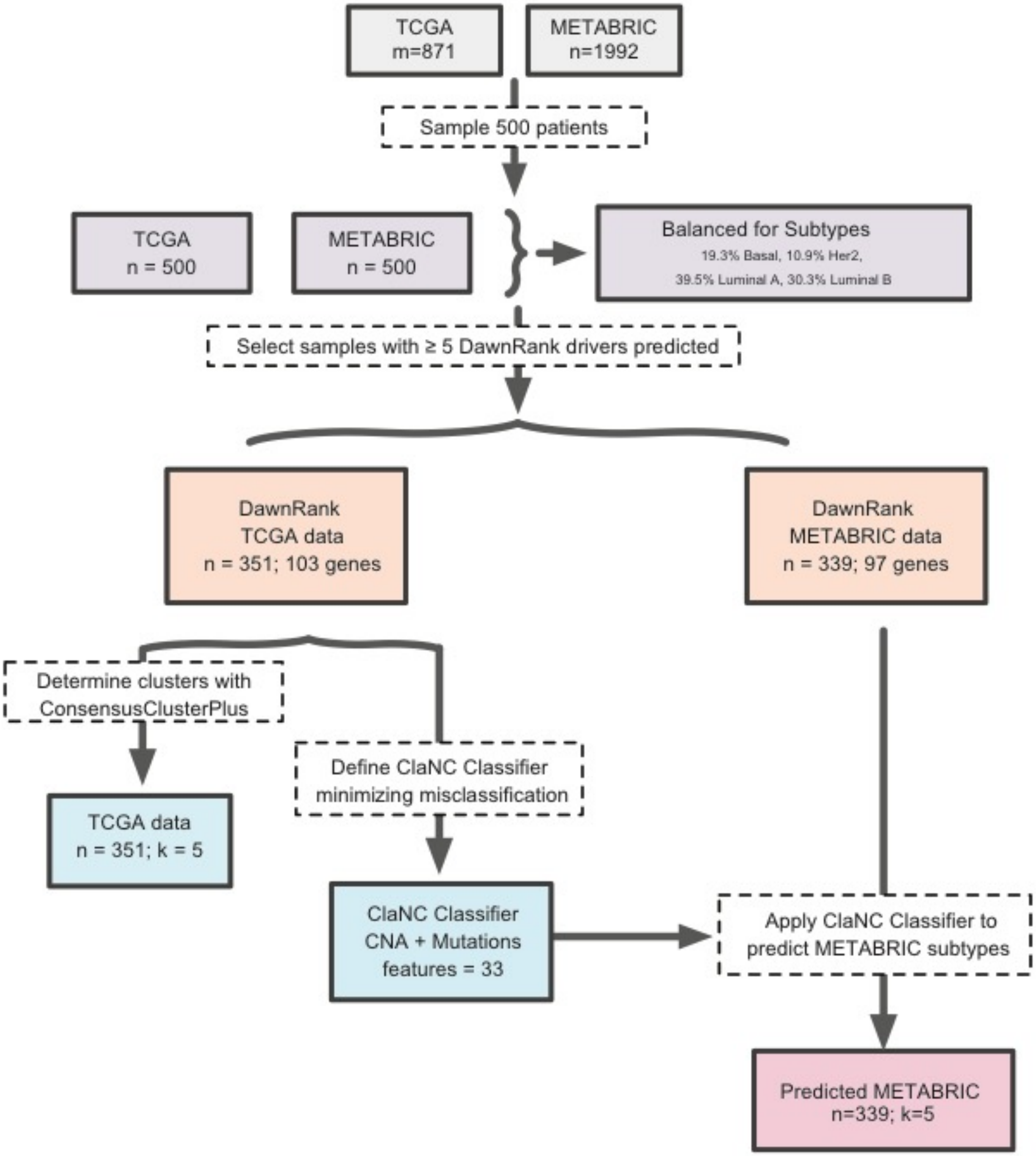


Figure 2: A schematic diagram detailing the overall workflow in this work.

FIGURES

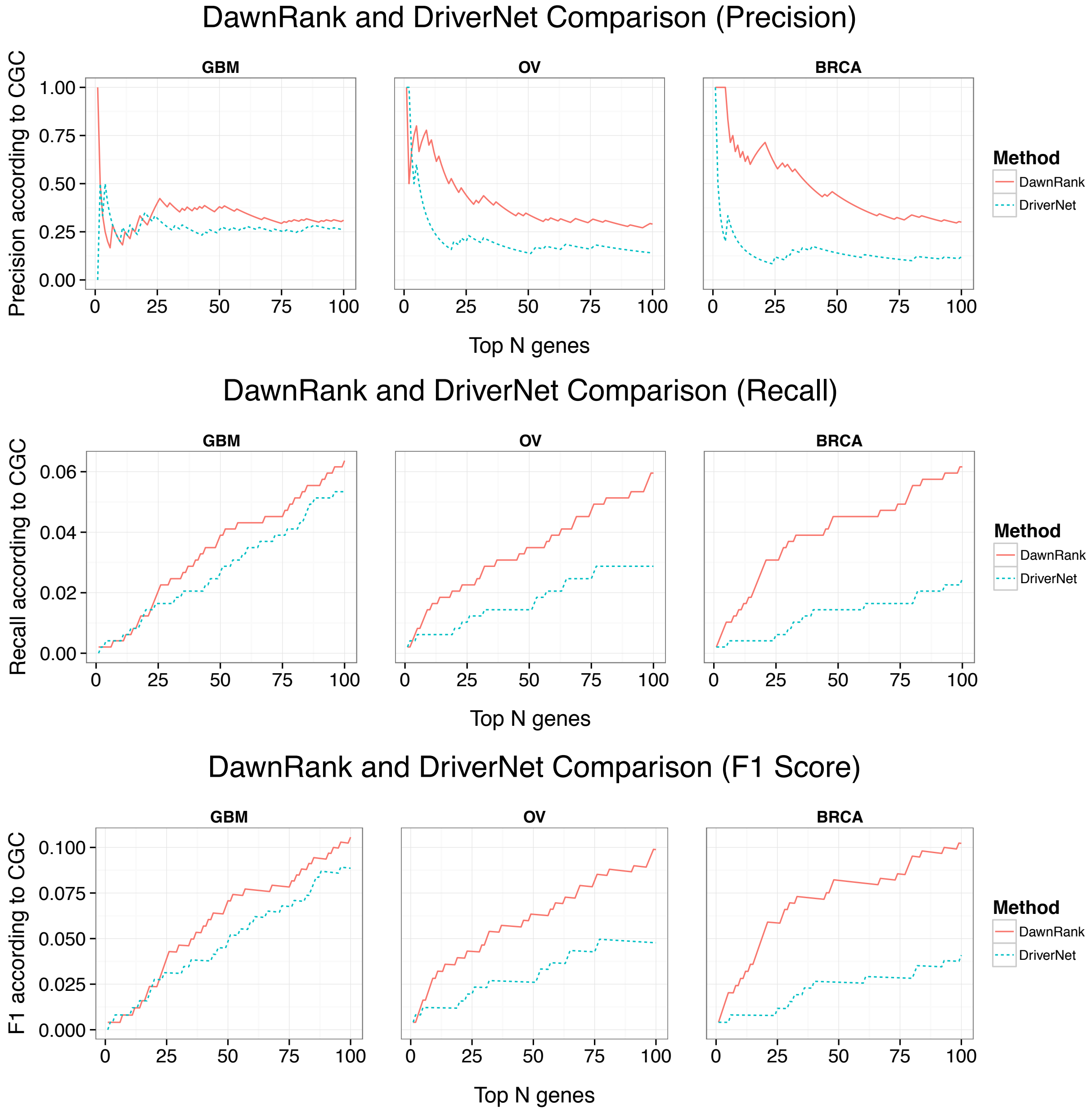


Figure 3: A comparison of the precision, recall, and F1-scores for the top ranking genes in DawnRank and DriverNet. The X-axis represents the number of top ranking genes involved in the precision, recall, and F1 score calculation. The Y-axis represents the score of the given metric.

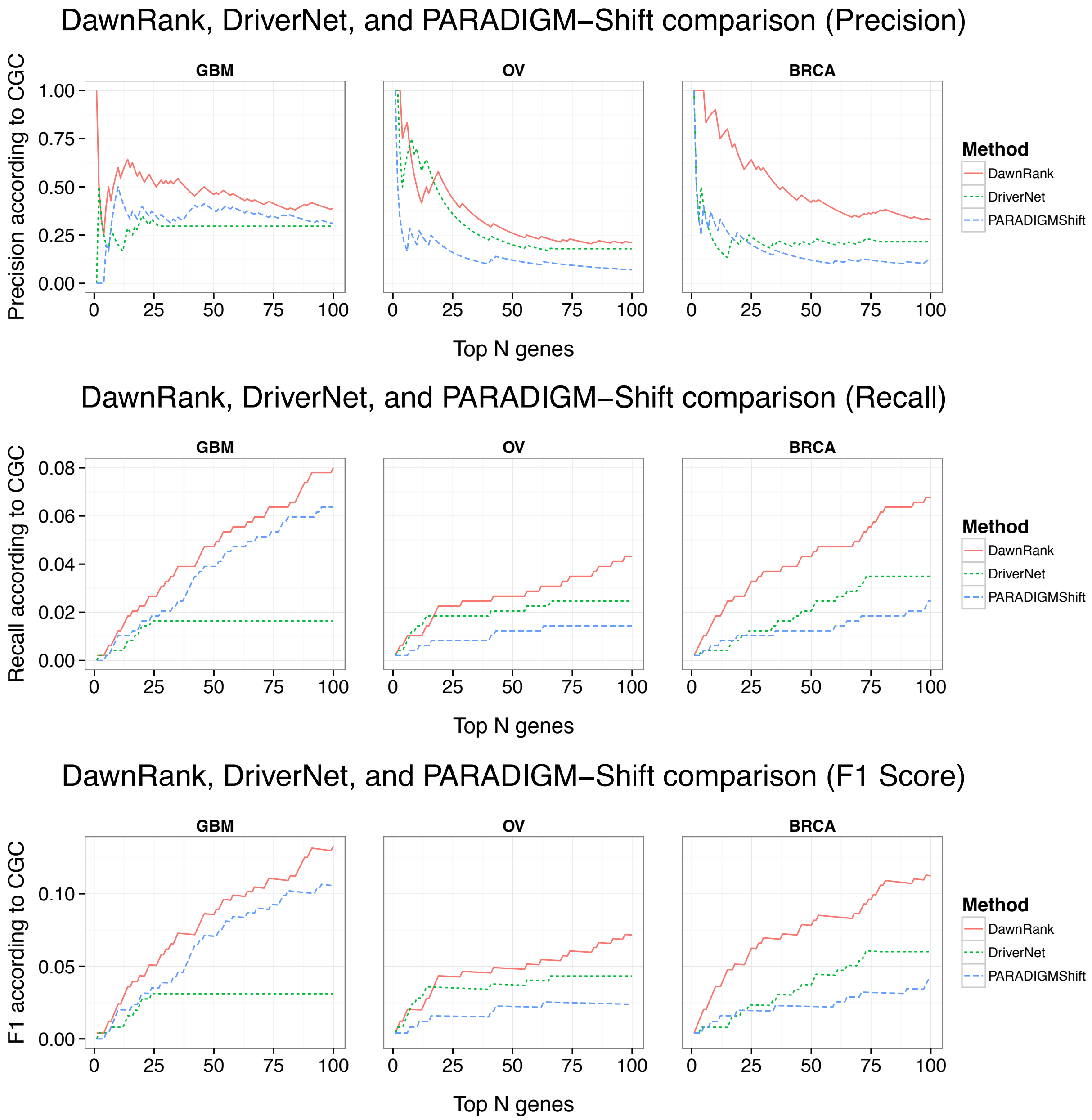


Figure 4: A comparison of the precision, recall, and F1 scores for the top ranking genes in DawnRank, DriverNet, and PARADIGM-Shift on a small network (defined from the KEGG database). The X-axis represents the number of top ranking genes involved in the precision, recall, and F1 score calculation. The Y-axis represents the score of the given metric.

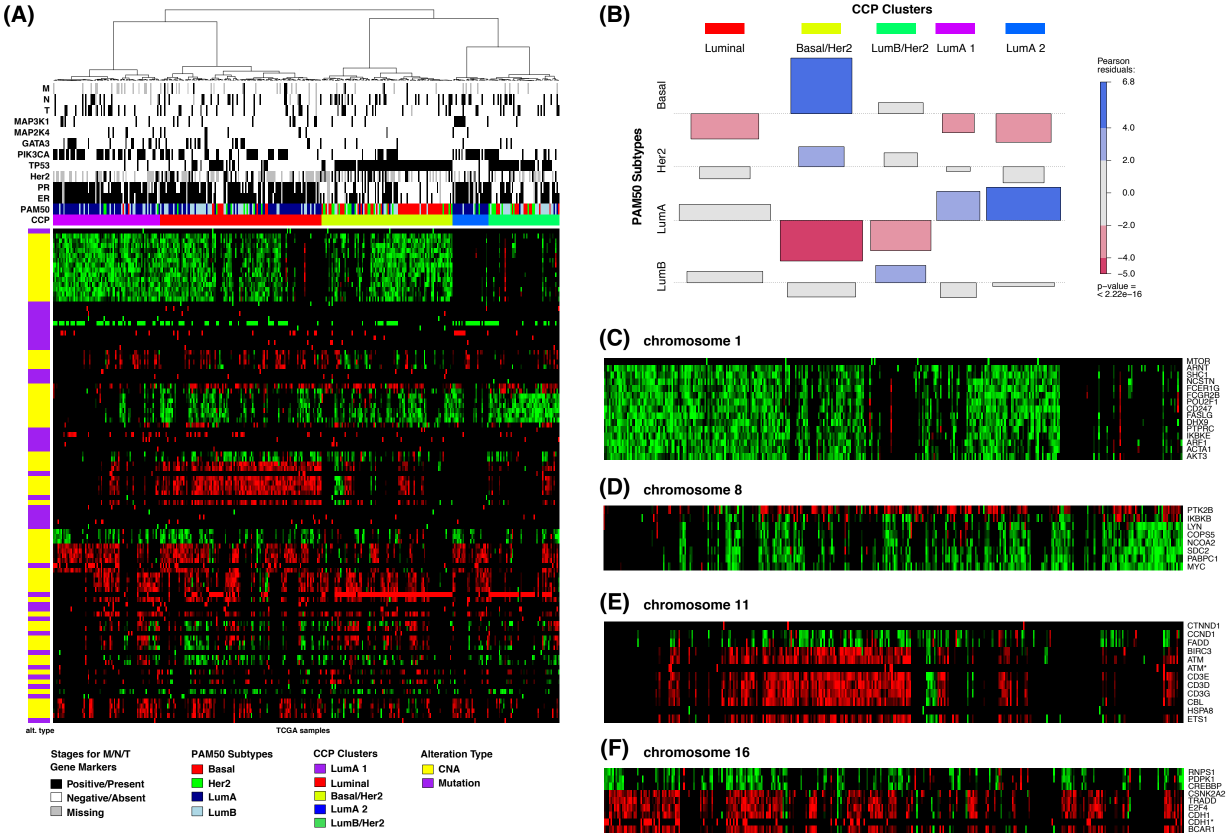


Figure 5: Clustering result based on driver genomic alterations using TCGA data. **(A)** A landscape plot detailing the subtypes defined by driver genomic alterations in TCGA breast cancer samples. The columns represent the samples and the rows represent the DawnRank-selected genes used in the clustering. The green entries represent copy num© gain (for CNA) and oncogene mutations (for point mutations). The red entries represent copy ©ber loss (for CNA) and tumor suppressor mutations (for point mutations). The intensity of the color reflects the DawnRank score. The tracks above the heatmap shows clustering results as well as comparison to other information such as PAM50 subtype, tumor stage, and ER/PR/Her2 status. **(B)** Correlation result between the ConsensusClusterPlus (CCP) and PAM50 subtypes. Positive associations are in blue and negative correlations are in red. The p-value at the bottom of the legend shows the p-value of the Chi-squared association test that determines the difference between the clusters. **(C)-(F)** The zoom-in view of the clusters with focal CNAs on chromosomes 1, 8, 11, and 16, as well as the key genes involved.

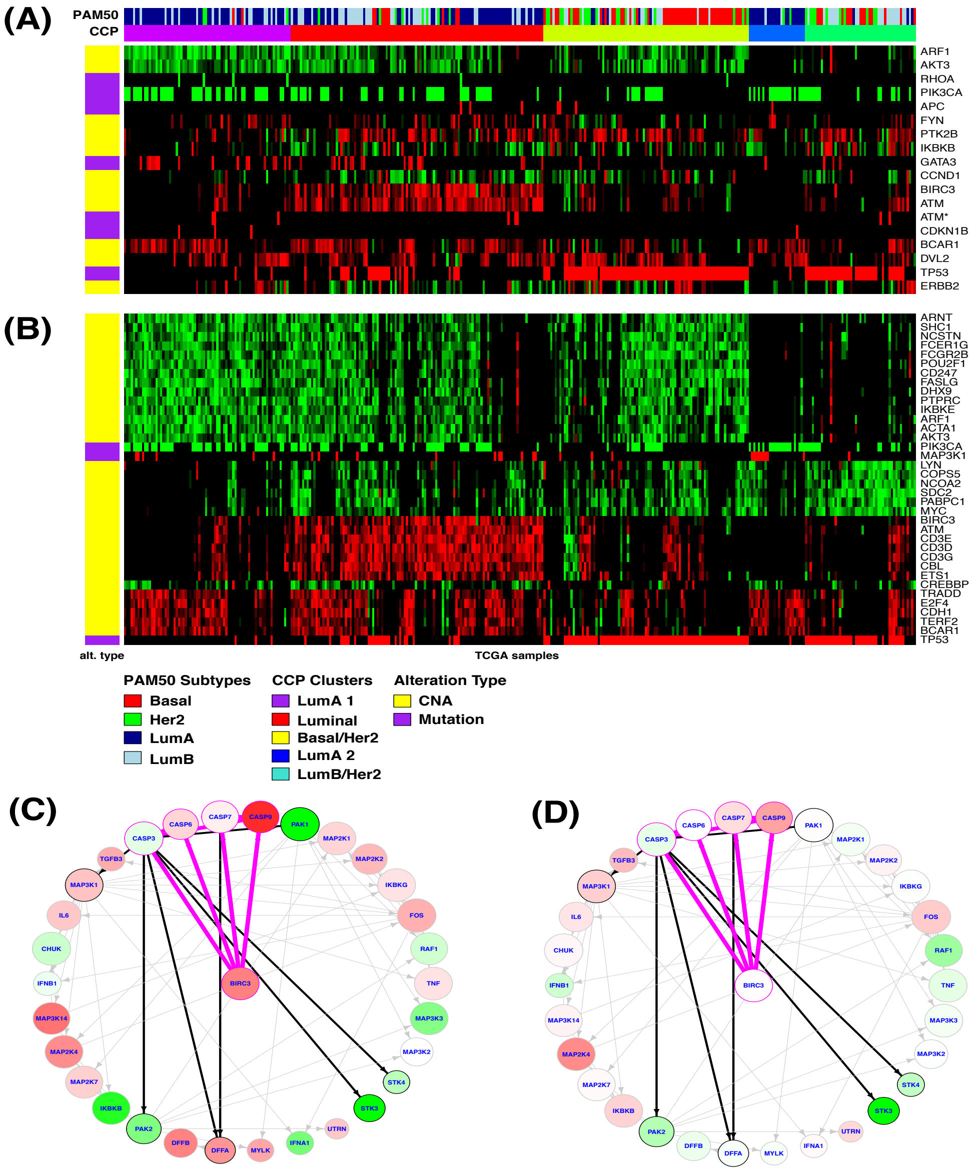


Figure 6: Subtype defining genes. **(A)** Genes selected from ANOVA analysis. **(B)** Gene selected from SAM analysis. **(C)** A network view in Luminal subtype detailing the gene interactions between BIRC3 and nearby genes in the network up to two levels downstream. Red nodes represent downregulation and green nodes represent overexpression. The intensity of the node represents the magnitude of gene expression. Edge thickness and color represent the distance between the gene in question and BIRC3. Magenta edges represent 1 degree of separation from BIRC3, black represents 2, and gray represents 3. **(D)** A network view of BIRC3 in non-Luminal subtypes.

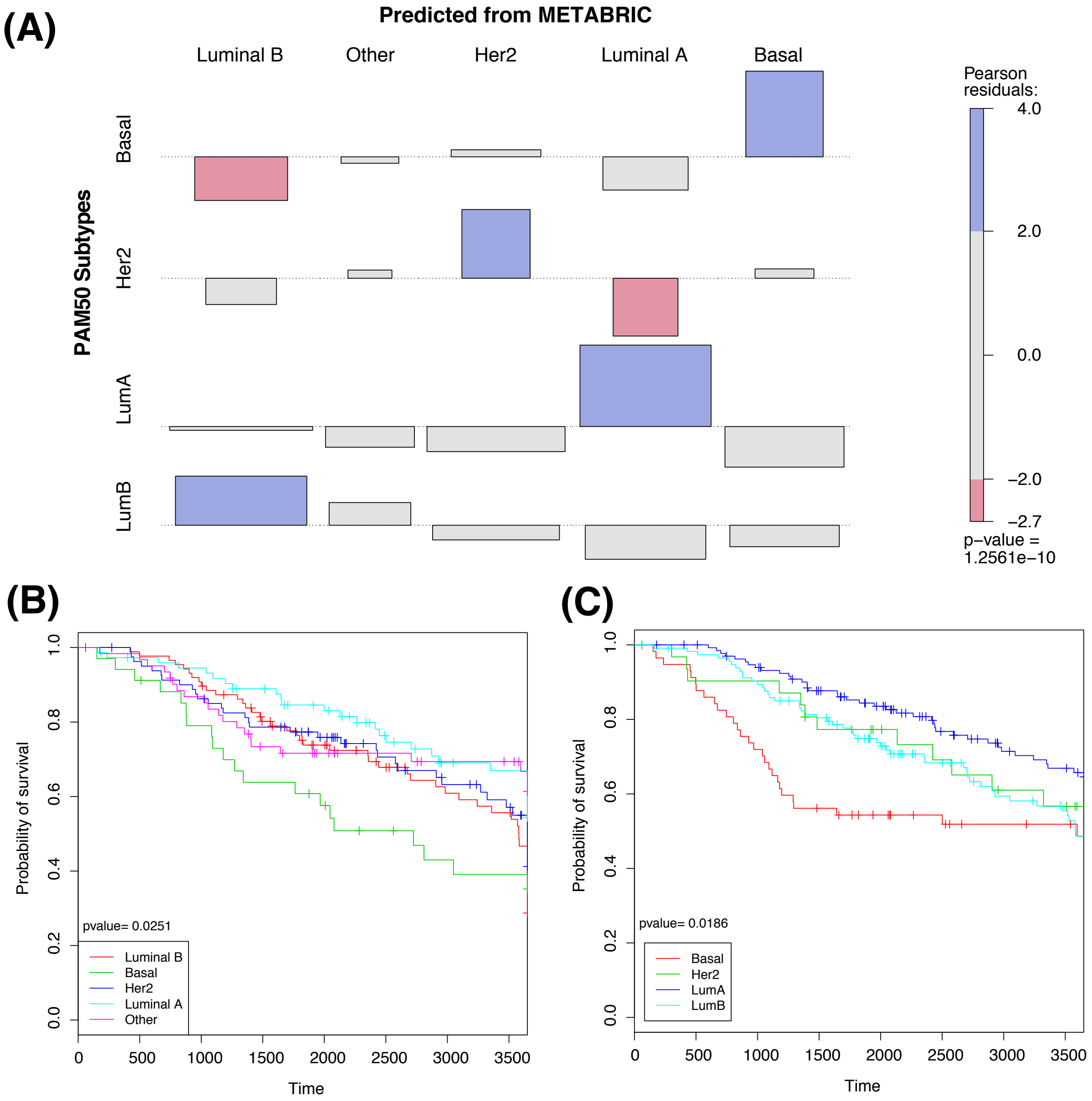


Figure 7: Comparison with the results in METABRIC dataset (based on classifier trained from TCGA clusters). **(A)** Comparison between the METABRIC predicted subtypes and PAM50 subtypes. **(B)** The K-M plot of the METABRIC predicted subtypes. **(C)** The K-M plot of the PAM50 subtypes.

CHAPTER 3: A new correlation clustering method for cancer mutation analysis[2]

3.1 Introduction

Rapid advances in high-throughput sequencing technologies have provided unique opportunities for analyzing large numbers of cancer genomes. However, the complexity of genomic alterations in cancer causes significant analytical and computational challenges that have to be overcome in order to fully characterize the functional roles of various mutations. In particular, as cancer genomes tend to contain a large number of diverse mutations (e.g., point mutations or copy number changes) most of which are neutral, one problem of significant importance is to identify a small set of mutations that perturb key biological pathways and have significant impact on tumorigenesis [10]. Hence, a central question in cancer genomics is how to distinguish “driver" mutations, which contribute to tumorigenesis, from functionally neutral “passenger" mutations.

Many computational methods have been developed to facilitate the discovery of driver genes [19], [112]–[115], most of which rely on mutation counts. Due to the high level of inter-tumor heterogeneity, two patients with the same cancer may have vastly different drivers and as a result many cancer mutations occur with low frequency in the patient population. Therefore, approaches relying on simple estimates of recurrence or frequency of mutations usually do not work well in practice. To mitigate this problem, several recent approaches have integrated frequency analysis with pathway-based and network-based models in order to ensure high

[2] This chapter appeared in its entirety in *Bioinformatics* and is referred to in the dissertation as “J. P. Hou, A. Emad, G. J. Puleo, J. Ma, and O. Milenkovic, ‘A new correlation clustering method for cancer mutation analysis.,’ *Bioinformatics*, p. btw546, Aug. 2016.”

accuracy of common driver mutation discovery [23], [67], [71], [107], [116]. Such methods have an advantage in so far that in addition to mutation analysis, they take into account gene interactions as an added source of prior knowledge.

In parallel, methods have been proposed to identify driver pathways, i.e., groups of genes that may interact together in combinatorial patterns to promote tumorigenesis. [29] described a method called MEMo, and subsequently used it to show that mutually exclusive modules based on known networks can aid in determining groups of genes that contribute to tumorigenesis. These gene groups, or modules, are jointly highly recurrent, have similar pathway impact in terms of biological processes, and their corresponding mutations tend to be mutually exclusive, meaning that very often only one gene in each gene group is mutated at a given time in any given patient. This mutual exclusivity rule in cancer pathways is supported by the observations that, in general, one mutated gene suffices to perturb the function of its corresponding pathway. Multiple mutations would require significantly higher energy investments on the part of cancer cells, and are hence selected against. [30] expanded the ideas behind the concept of MEMo with iMCMC, and provided a framework to integrate mutation data, copy number, and expression information into cancer network weights which they used to identify modules; they also performed multiple types of integrative cancer perturbation data analysis. Dendrix [31] was developed to identify driver pathways *de novo* using mutual exclusivity and coverage (patient coverage) principles, without relying on known network information that has the potential to improve the discovery process of new modules. MDPFinder [32] expanded on the overall framework of Dendrix by incorporating gene expression information to ensure that genes in discovered mutually exclusive pathways were also co-expressed. Multi-Dendrix [33] and CoMDP [34] improved on the limitations of Dendrix and MDPFinder, respectively, by allowing their algorithms to find multiple co-occurring modules. More recently, CoMEt [35] was proposed to address an inherent bias in Dendrix and Multi-Dendrix that resulted in high frequency mutations being significantly more likely to be included in mutually exclusive modules.

However, while methods such as Dendrix, Multi-Dendrix and CoMEt all have the ability to identify mutually exclusive modules *de novo*, they still have significant limitations.
The aforementioned methods are typically inefficient when applied to large-scale datasets with large values of their relevant parameters. Also, some of these methods are randomized in nature and no guarantees exist that multiple runs of the methods will produce compatible results. Furthermore, almost all methods are able to identify only a small number of modules of limited size as cluster sizes are critical algorithmic parameters from the perspective of computational tractability. Most importantly, they have to be redesigned or restructured whenever new biological information is included in the discovery process.

To overcome these and other shortcomings of existing methods, we introduce a novel method called Cancer Correlation Clustering C^3 to directly tackle the problems of integrating diverse sources of evidence regarding driver pattern behavior and eliminating computational bottlenecks associated with large cluster sizes or cluster numbers. The C^3 method uses a new agnostic optimization framework specifically developed and rigorously analyzed for the driver discovery task, in which patient data is converted into a simple set of weights used in the objective function that do not require the algorithm to change upon incorporation of new data sources. In addition to this flexibility, C^3 has low computational cost, and it allows for adding relevant problem constraints while retaining good theoretical performance guarantees. Furthermore, the algorithm outperforms CoMEt in three out of four evaluation criteria, where the three criteria depend on which weights are "emphasized" in the optimization problem: tuning the weights allows one to select which features to improve or emphasize. What the relevant constraints features are may be chosen by the user, although our analysis included coverage, mutual exclusivity, expression data and network pathway information. We also point out that the weights may be chosen so as to cater to the need of many other computational biology problems that involve optimization on graphs.

To test C^3, we ran extensive simulations for several cancer types (including breast cancer, kidney cancer, ovarian cancer, glioblastoma, etc). Unfortunately, the patient sample set sizes for all except two cancers -- breast cancer and glioblastoma -- did not allow for accurate and statistically significant driver identifications for any of the used methods. We hence report results for these two cancers only, although a pan-cancer study is easy to conduct once sufficiently many samples become available.

The chapter is organized as follows. Section Results contains the main results of our analysis, a comparison of the performance of C^3 and CoMEt on breast cancer and glioblastoma data. A discussion of our findings and concluding remarks are given in Discussion. Section Methods contains a basic introduction of the principles of correlation clustering and the evaluation criteria used to compare C^3 and CoMEt.

3.2 Methods

3.2.1 C^3 approach

The basic idea behind C^3 approach is correlation, an agnostic learning technique first proposed in Bansal et al. [117]. In the most basic form of the clustering model, one is given a set of objects and, for all or some pairs of objects, one is also given an assessment as to whether the objects are “similar” or “dissimilar”. This information is described using a complete graph with labeled edges: each object is represented by a vertex of the graph, and the assessments are represented by edges labeled with either a “+” symbol, for similar objects, or a “-” symbol, for dissimilar objects. The goal is to partition the objects into clusters so that the edges within clusters are mostly positive and the edges between clusters are mostly negative. Unlike in many other clustering models, such as *k*-means [118], the number of clusters is not fixed ahead of time and finding the optimal number of clusters is part of the problem. Furthermore, the assignment of positive and negative edges does not have to be mutually consistent: for example, if the graph contains a triangle with two positive edges and one negative edge, then we must either group the endpoints of the negative edge together, erroneously putting a negative edge inside a cluster, resulting in a “negative error” or else we must group them separately, forcing one of the positive edges to erroneously go between clusters, resulting in a “positive error”. When a perfect clustering is not possible, we seek an optimal clustering: one that minimizes the total number of “error”. This form of correlation clustering is known to be NP-hard, but depending on the graph topology, various constant or logarithmic approximation guarantees exist. Bansal et al [117] also proposed a weighted version of the correlation-clustering problem. A more general weighted formulation was introduced in Chakiar et al [119] [120], and this is the formulation we subsequently generalize. In this model, each edge *e* is assigned two nonnegative weights, $w_e^+$and w_e^-. A clustering incurs cost, w_e^+ if *e* is placed between clusters, and incurs cost w_e^- if *e* is placed within a cluster.

If no restrictions are placed on the weights w_e^+ and w_e^- then it is possible to have edges with $w_e^+ = w_e^- = 0$; these edges are effectively absent from the graph, so there is no loss of generality in assuming that the graph is a complete graph. Nevertheless, in order to arrive at problems that have efficient constant approximation algorithms, one needs to place certain restrictions on w_e^+ and w_e^-. The probability constraints give a natural restriction on the edge weights $w_e^+ = w_e^- = 1$ for every edge *e.* Another restriction involves the *triangle inequality*, and one requires that $w_{uw}^- = w_{uv}^- + w_{vw}^-$ for all distinct vertices u, v and w.

The analytic approach pursued in this work operates on the following model: genes which show sufficiently large mutation prevalence in cancer patients represent vertices of a *complete connected graph* whose vertices are to be clustered according to similarity criteria and weights to be described in detail in the next section. Note that we only use the top 5% of mutated genes in cancer patients, ordered by mutation frequency, as vertices. The reasoning behind our approach is as follows: First, low-frequency mutations require *specialized statistical and network analysis methods* which have to be developed in parallel and for which not sufficiently many patient samples are yet available [121], [122]; Second, even when restricting our attention to the most frequently mutated genes we outperform all known methods, which illustrates that one can significantly scale down the set of genes under consideration and at the same time improve identification performance. The low-frequency trimming approach results in 170 genes in glioblastoma (GBM) and 130 genes in breast cancer (BRCA). Although these numbers may appear prohibitively small given that more than a hundred cancer driver genes are reported, usually only a very small number of driver genes are needed to initiate the process of tumorigenesis (For example, in [123], it was shown that only three driver gene mutations are required for the development of lung and colorectal cancers.)

The weights w_e^+ and w_e^- assigned to an edge *e* connecting two genes *u* and *v* are weighted sums of weights capturing driver gene features, such as mutual exclusivity, coverage strength, network distance and expression similarity. More precisely, the negative weights w_e^- are chosen to be relatively small if the endpoint genes describing the edge are deemed to be mutually exclusive in cancer patients. A small negative weight encourages placing mutually exclusive genes *within the same cluster*, as the penalty paid for placement in the same cluster is small. The positive weights jointly depend on the coverage, network distance and expression correlation of the endpoint genes: The larger the joint coverage, co-expression and inverse of the network distance of the endpoint genes, the larger the positive weight and the more likely the genes will end up in the same cluster so as to avoid paying a large cross-cluster cost. For a detailed and rigorous discussion of the exact method for determining clustering weights with respect to the expression, coverage, network and mutual exclusivity, refer to the main paper [68].

To control the size of the resulting clusters so as to discourage uninformative singleton and giant clusters, we developed two new correlation clustering algorithms that use cluster sizes as problem parameters that may be chosen by the users. These cluster size bounds also allow for more accurate comparison with other methods which operate with inherent cluster size constraints. Furthermore, as pointed out in [31], driver pathways obeying mutual exclusivity and coverage constraints are usually smaller than most pathways annotated in the literature. This observation provides another reason for using bounded cluster sizes as well. Note that unlike in the aforementioned known methods, the cluster sizes have no bearing on the complexity of our algorithm nor on their overall approximation quality, and they may be completely removed by the user if so desired.

The driver discovery approaches closest to C^3 are Multi-Dendrix [35] and CoMEt [124] Multi-Dendrix is an integer linear programming clustering algorithm that ensures that the genes within a cluster have mutation patterns that satisfy mutual exclusivity and coverage: In a nutshell, for any two genes in a cluster, the number of patients in which these genes are mutated at the same time is relatively small; in addition, a large portion of the patients has at least one mutation in each cluster. CoMEt uses a statistical score for mutation exclusivity that is conditioned on the frequency of each alteration, alleviating the inherent bias caused by frequently mutated genes.

Compared to Multi-Dendrix and CoMEt, C^3 uses *weighted* linear programming relaxation instead of an integer linear program which significantly improves the versatility and running time of the algorithm. Furthermore, the weights allow for straightforward incorporation of heterogeneous sources of evidence into the clustering method and the algorithm itself remains unchanged with the addition of new data. On the other hand, Multi-Dendrix cannot be easily adapted to new problem constraints. This flexibility comes at the cost of C^3 providing only an approximate solution, but the approximate solutions exhibit large overlap with the exact solutions for a number of tested smaller synthetic networks. In addition, given the inherently approximate nature of optimization criteria, the weight selection and parametrization of both algorithms, this does not appear to be a significant shortcoming. Also, empirical evaluations on real data suggest that the approximation algorithms produce results very close to the optimal solution.

3.2.2 Clustering algorithms

The classical formulation of correlation clustering does not include cluster size restrictions. On the other hand, all known driver identification methods operate with de facto cluster size bounds, as the cluster sizes govern the computational complexity of the method. For example, comprehensive testing of CoMEt reveals that the algorithm fails to operate beyond cluster sizes of 10-12. In order to perform a fair comparison, we introduce a cluster size constraint in our algorithm, by assuming that all clusters are of size *K*. Clearly, setting *K* equal to the number of vertices (genes) removes the cluster size constraint, hence our algorithm has a large flexibility in cluster size selection. An additional reason for choosing a restricted cluster size is that we expect driver genes of specific cancer types to be grouped together within clusters, and as already remarked, a number of recent results suggest that only a few drivers are actually present in any cancer type. Making the clusters excessively large would potentially lead to inclusions of multiple cancer type drivers in the same cluster, thereby obscuring the fine partition of the drivers. Nevertheless, the user of the method may choose *K* according to her/his own requirements. Yet another reason for introducing cluster sizes is to avoid the shortcomings of many known clustering algorithms which tend to produce non-informative “giant clusters” and singleton clusters.

The bounded cluster size correlation clustering problem for driver gene inference may be formulated as follows. As already described, let *K* be a “hard” bound on the size of the driver clusters, and let the positive w^+ and negative weights w^- be chosen according to a desired combination of datasets, as explained in the previous section. The optimum clustering may be found by solving the integer linear program (ILP) below.

$$\underset{x}{\text{Min}} \sum_{e \in E(G)} (w_e^+ - w_e^-(1 - x_e)) \tag{3.1}$$

$$\text{Subject to } x_{uv} \leq x_{uz} + x_{zv} \text{ (for all distinct u, z, v } \in \text{V(G))} \tag{3.2}$$

$$\sum_{u \neq v} (1 - x_{uv}) \leq K \text{ for all u} \in \text{V(G)} \tag{3.3}$$

$$\text{x}_\text{e} \in \{0,1\} \text{ for all } e \in E(G) \tag{3.4}$$

In this formulation, and for a fixed edge $e = uv, x_{uv} = 1$ implies that u and $v$should belong to different clusters and $x_{uv} = 0$ implies that the two vertices should belong to the same cluster. Note that the triangle inequality (3.2) ensures that if *u* and *z* are in the same cluster and *z* and *v* are in the same cluster, then *u* and *v* are also in the same cluster. Any clustering of the vertices can be described using the variables x_e. For a fixed clustering, the objective function is the cost associated with that clustering.

Solving the ILP is NP-hard. We hence relax the problem by changing the integer constraint $x_e \in \{0,1\}$ to an interval constraint $x_e \in \{0,1\}$. This relaxation leads to a classical linear program (LP), the solution of which may be fractional. To obtain a valid clustering, the fractional solutions have to be subsequently rounded to produce integer solutions. Unfortunately, known rounding algorithms we previously developed in [125] tend to produce very small clusters, often as small as single-vertex clusters that are not meaningful. For our study, we hence slightly modify the algorithm by moving the cluster size constraint (3.3) from the LP to the rounding procedure (See original paper, Algorithm 1). Hence, the clustering algorithm involves solving (3.1) without the constraint $\sum_{u \neq v} (1 - x_{uv}) \leq K$ and then applying the rounding procedure of the rounding procedure.

The rounding procedure is closely based on the rounding algorithm described in [119], [120] The idea behind the rounding algorithm is to pivot on one vertex, examine its closest neighbors, where closeness is governed by the value of the output variables $x_e$of the LP, and partition large neighborhoods if needed to get clusters of size at most $K + 1$. Given that the parameter α is set to 2/7 and given that the weights obey the following constraints:

- $w_e^+ \leq 1$ For each edge *e*
- $w_e^+ + w_e^- \leq 1$ For each edge *e*

The above inequalities were addressed as described in the previous section, and we remind the reader that they were imposed on the weights through proper normalization. Note that we only used high frequency mutations for our clustering problem, and hence did not encounter any computational issues with the LP solvers. On the other hand, if one were touse all 25,000 genes in the analysis, the LP solver implemented in Gurobi (https://www.gurobi.com/) would inevitably break down due to the large number of constraints, which is quadratic in the number of genes. In this case, a much simpler scalable solution is to use approximate LP solvers, akin to those described in [126].The approximate solver is guaranteed to produce a solution that does not exceed the LP solution by more than a factor $1+\varepsilon$, for some small value of ε, by using gradient descent methods that are highly scalable.

3.2.3 Evaluation methods

We evaluated the performance of both C^3 and CoMEt in terms of their ability to detect *mutually exclusive, high-coverage, and biologically relevant gene clusters*. At this point, it is important to observe that the inference and evaluation strategies may appear to involve circular arguments: Mutual exclusivity, coverage and network distance, used to predict the clusters, are also used to evaluate the performance of the clustering method. But this is clearly not the case, as mutual exclusivity, coverage and network distance are *optimization constraints*, and one always needs to test the quality of a (approximate) solution to an optimization problem based on how well the constraints are accounted for. Other driver discovery tools, such as CoMEt, use the same constraint modeling and evaluation criteria. Furthermore, we added one more evaluation criteria, related to biological significance and pathway enrichment analysis, which is independent on the optimization criteria. As will be shown in the subsequent section, this evaluation criterion confirms the quality of the C^3 analysis for cancer driver gene inference and its improvements over CoMEt.

We ran both the C^3 and CoMEt methods using mutation and CNA data collected from TCGA, pertaining to breast cancer (BRCA) [44] and glioblastoma (GBM) [77]. In addition to GBM and BRCA, we also considered kidney cancer (KIRC) and ovarian cancer (OV), but the available patient data appeared limited at this stage to allow for statistically significant and comprehensive results. We accessed the TCGA provisional data using the cBioPortal platform [127] on August 14, 2015.

We ran both methods using the same alteration dataset. We evaluated both point mutations and indels, and for CNAs, we used the GISTIC thresholds [128] of -1 and 3 as our cut-offs (as already pointed out in the previous section). To focus on mutations with high frequency, we only selected genes in the top 95 percentile of alteration frequencies, thereby obtaining 130 genes spanning 959 patient samples in BRCA and 170 genes spanning 291 patient samples in GBM.

To test the effects of cluster sizes and the quality of our results, we ran both C^3 and CoMEt to find clusters of sizes upper bounded by 5, 6, 7, 10, and 15. As already pointed out, larger cluster sizes are easily accommodated for C^3, but since CoMEt failed to produce solutions for clusters of sizes roughly greater than ten, we restricted our attention to the aforementioned range of values. Due to the fact that correlation clustering and CoMEt will cluster all genes in a dataset, and hence produce a partition of the gene set, a large number of clusters will contain neutral mutations only and will hence have no biological significance. This is why we only compared the top ten most mutually exclusive gene sets generated by C^3 with those of CoMEt.

We ran CoMEt with 1,000 iterations each and 3 initialization points to ensure both timely and consistent runs. For C^3, we ran the C^3 clustering method for all combinations of weights $w_1, w_2, w_3 \in \{0,0.25,0.5,0.75,1\}$ that satisfy $w_1 + w_2 + w_3 = 1$ but selected to report only results for the weight parameters $w_1 = 0.167$ (coverage), $w_2 = 0.333$ (network information) and

$w_3 = 0.5$ (expression data). Our choice is governed by the fact that coverage seems to be a biologically much less important criterion then network information or expression. Hence, high weights for expression and network information increase the ability of the C^3 algorithm to detect biologically significant clusters. Furthermore, the patient coverage criteria appear to be less relevant than pathway coverage and some other coverage properties that have not been explicitly investigated in the literature. Nevertheless, we observe that the choice of the weights may be completely governed by the user, and that the increase in one weight may produce better results in one performance category while reducing the performance in another category.

We used four statistical methods to assess the performance of the algorithms which reflect both the statistical and biological significance of the clusters found.

Mutual Exclusivity: To evaluate the degree of mutual exclusivity in a cluster' we performed a Fisher's exact tests [129] for each pair of genes in 'he cluster. The Fisher's exact test uses a hypergeometric distribution to calculate the probability of observing a 2 × 2 contingency table of a total of *n* samples, with *a* samples that have an alteration in two genes (say, g_i and g_j), b samples with an alteration in gene g_i only, and c samples with an alteration in gene g_j only. If *d* is the number of samples with no alteration in either gene, then the probability of co-mutation is evaluated according to

$$P(g_i, g_j) = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} \tag{3.5}$$

We also evaluated the overall exclusivity of a cluster as the median value of each pairwise exclusivity test, for each pair of genes g_i, g_j in the netwo'k. The pairwise Fisher's method has also been used by the Mutex suite to establish mutual exclusivity [130]. However, because the context that the Fisher's exact test is used as an evaluation rather than as a discovery tool, we used the median pairwise p-value rather than the maximum p-value to get a better sense of the overall exclusivity of genes within a cluster. It is also important to note that while CoMEt has a built-in method that generalizes the exclusivity test to a *2^k* contingency table for a cluster size $k \geq 2$ the exponential size of their test set makes evaluation for large cluster sizes computationally impractical. An alternative test for overall mutual exclusivity is a permutation test, as implemented by MEMo, which compares the exclusivity of a gene set by sampling random gene sets and patients with multiple alterations.

Coverage: To compare and evaluate the overall coverage of a cluster found by C^3 or CoMEt, we calculated and reported the proportion of patients with at least one alteration in a gene belonging to the given cluster.

Network Clustering: We performed an additional pathway analysis for the potential cancer gene drivers. As pointed out in the previous section, driver genes tend to be, on average, closer to each other in a pathway compared to randomly selected genes. Our tests involved assessing the shortest network distance of genes within the discovered clusters. We remind the readers that the distances were e'aluated using Dijkstra's Algorithm on 8,726 genes from [29].

Biological Significance: In addition to testing the quality of the algorithm in terms of optimizing mutual exclusivity and coverage, we also investigated the biological significance of the C^3 and CoMEt methods from the perspective of gene discovery and pathway analysis. Although there is no overarching gold standard to determine biological significance, a commonly accepted metric employed by MEMo, Dendrix, Mutex, CoMEt and other similar tools is to count the number of known driver genes found within *the best clusters* according to the given criteria. These clusters usually contain known driver genes. To determine the driver gene-based biological significance, we calculated the proportion of drivers found in the ten most mutually-exclusive $C^3$and CoMEt clusters using a comprehensive, curated list of known drivers from the CGC.

It is important to point out that while the four test benchmarks we introduced are a reliable way to test the optimization quality and performance of CoMEt and C^3, no perfect benchmark exists for detecting mutually exclusive and biologically significant genes clusters. The hope is that multiple evaluation methods taken together may provide a better understanding of which methods outperform others in a given parameter and criteria setting.

3.3 Results

In what follows, we demonstrate that C^3outperforms CoMEt in almost all of the aforementioned benchmarking criteria, or more precisely, for three out of the four chosen criteria. This is achieved without any special parameter tuning or optimization. As a rule of thumb, C^3 can be made to outperform CoMEt *in any chosen single, pair of triple of criteria* by adjusting the weights. This observation may be explained by the fact that the weights trade off the strengths of different modeling assumptions. We supplement our statistical analysis with a discussion of the biological relevance of our findings, and explore the role of the new potential drivers found by C^3 within their driver gene communities. In particular, we discuss the significance of large mutually exclusive clusters that cannot be recovered by other methods. Recall that we restrict our attention to the ten best performing clusters according to mutual exclusivity, as this approach was used in the original evaluation process of the CoMEt algorithm.

3.3.1 Performance evaluation

The results of our extensive comparison between C^3 and CoMEt, regarding mutual exclusivity, coverage, driver identification, and pathway-level evaluation, are shown in Figure 10. Both algorithms were tested on the same server with a 256GB RAM memory. Both methods ran uninterruptedly when the cluster sizes were constrained $k = 5, 6, 7, 10$. CoMEt reported segfault memory errors for $k = 15$, and for this case, only C^3 was benchmarked.

To assess the biological significance of the two methods in terms of their ability to cluster high-impact drivers from the CGC repository together, we compared the results of C^3 and CoMEt both to each other and to a "baseline" value equal to the average proportion of drivers in the ten most mutually-exclusive clusters found, in this case 0.067, using uniform random sampling of genes (see Figure 10A).

In BRCA, we found that C^3 detected a median driver proportion of 0.160 and CoMEt detected a median driver proportion of 0.117 in the top ten clusters. C^3 outperformed CoMEt for each cluster size. We also used a Mann-Whitney Rank Sum test [131] to compare the overall performance of the algorithms with respect to mutual exclusivity, for all cluster sizes. We chose a rank-sum test because it is unclear that the drivers are following a normal distribution due to the small amount of data points available. The results show that C^3 outperforms CoMEt (p-value of 0.0079) in terms of amount of drivers in clusters. C^3 also outperforms CoMEt on GBM, with a median proportion of drivers per cluster equal to 0.170, compared to a 0.12 proportion of drivers per cluster found by CoMEt. This finding holds for every cluster size, with a rank-sum test p-value of 0.0361. Both methods succeed in finding biologically significant drivers within clusters exhibiting high mutual exclusivity, and both methods significantly outperform the expected number of drivers per cluster in the random setting p-value 1.594e-5 and p-value 1.312e-3 for C^3 and CoMEt, respectively).

We next tested the clusters found by each method based on their mutual exclusivity (see Figure 10B). To do so, we used the previously described pairwise Fisher's exact test to obtain a p-value for each of the top ten clusters of the two methods. For better visualization, we performed a negative log transform on the p-values, and plotted the transformed p-value distribution. Hence, in this system, larger values indicate more mutually exclusivity.

We again used a Mann-Whitney rank-sum test to evaluate the performance of C^3 and CoMEt. For BRCA, one can see that while both methods have significant median exclusivity values (p = 7.541e-6 for C^3 and p = 3.337e-4 for CoMEt, C^3 has an overall more significant p-values for each cluster size. The median p-value of C^3 for each cluster size is lower than its CoMEt counterpart except for the case k=10. However, C^3 does have superior performance overall with a rank-sum p-value of p = 4.020e-4. For GBM, the median exclusivity results are not as strong as for the BRCA set, for both the C^3 and CoMEt method. C^3 has a median p-value of 0.3095 as opposed to CoMEt's 0.5022. The general drop in significance may be attributed to a lower c'nfidence of the Fisher's test due to a small number of samples available; recall that the GBM set involved 291 samples, compared to 959 BRCA samples. This indicates that one should look at individual significant clusters to evaluate mutual exclusivity. Even for the reduced median p-value regime, C^3 outperforms CoMEt in significance, having lower median p-values for each cluster size. Overall, the C^3 p-values are consistently and significantly lower than those produced by CoMEt for mutual exclusivity (the rank-sum test p-value equals 0.04401).

The results of the coverage tests are depicted in Figure 10C. In the coverage benchmark, CoMEt outperforms C^3 for GBM, but neither method outperforms the other for BRCA. In BRCA, both methods show comparable performance, with a median result for the fraction of samples covered equal to 0.5505 for C^3, and 0.5662 for CoMEt. This rather poor performance of both methods is observed for all values of *k,* with no p-value based on Student's T-test [132] being less than 0.05. The largest difference in coverage recorded for the two methods is present for $k = 6$. In conclusion, there appears to be no statistical difference between C^3 and CoMEt in terms of BRCA coverage percentage (p-value of 0.5127). In GBM, the median p-value for coverage difference is more pronounced. The median coverage of C^3 is 0.632 and the median coverage of CoMEt is 0.696. CoMEt finds significantly higher-coverage cluste's according to Student's T-test, with p-value 0.0345, and the most pronounced coverage percentage differences exist for small values of *k* (0.3745 vs. 0.6495 for $k = 5$ C^3 and CoMEt, respectively).

It is also important to note the wide distribution of coverage score values produced by C^3 for small *k*; the IQR (Interquartile range) value is roughly 0.35 for $k = 6$. The most likely reason behind this result is that our test weights were chosen to boost the relevance of mutual-exclusivity and biological significance rather than coverage. Mutual exclusivity accounts for 100% of the negative weights of edges, while coverage accounts for only 16.7% of the positive weights. We justify this weight choice by the fact that it leads to multiple significant cluster discovery and with our assumption that coverage is a less significant driver property compared to mutual exclusivity. We also point out that it appears that a biologically more relevant coverage constraint is pathway coverage, rather than patient sample coverage. Another setting in which we analyzed C^3 and CoMEt involves pairwise distances of drivers in the network (see Figure 10D). Here, we calculated the average pairwise distance between all pairs of genes clustered together. We then used Student's T-test to determine the statistical significance of this value. We also compared the values for both algorithms based on 1000 randomly selected genes by using a permutation test. For BRCA, we found no significant performance difference between the two methods in terms of the average pairwise distance: 3.110 for C^3 and 3.070 for CoMEt, with a p-value of 0.9330. In GBM, C^3 showed a smaller average pairwise distance of 2.908 compared to CoMEt's 3.097. This difference is statistically significant, with a p-value of 0.0379. The small average network distance results of C^3 for GBM, coupled with the low coverage, leads to the conclusion that C^3 favors niche, exclusive clusters in biologically relevant cancer pathways. Hence, the method may be useful for discovering specific molecular cancer subtypes. Both methods had an average pairwise distance well below the permutation benchmark of 3.903: the p-values of both C^3 and CoMEt were less than 2e-16 for both cancers.

In conclusion, from our detailed evaluation we conclude that although C^3 does not simultaneously outperform CoMEt with respect to all four evaluation criteria, but only three of them (which already represents a significant advantage), the C^3 performance indicates a strong overall propensity to select biologically more relevant and more mutually exclusive clusters, with a higher degree of flexibility compared to CoMEt.

3.3.2 Discovering potential driver pathways

We examine next the potential of the C^3 algorithm to detect clusters whose genes may be new candidate cancer drivers. We focus our search on clusters that contain biologically significant driver genes and known biological network interactions, and exhibit high mutual exclusivity and coverage. At the same time, we only consider the large cluster size regime, as results in this domain have not been previously reported in the literature and as they offer many new interesting insights. Two examples of our analysis are shown in Figure 11 and Figure 12.

In BRCA, one candidate cluster with several potential novel driver genes is the cluster containing *PTEN, HUWE1, CNTNAP2, GRID2, CACNA1B, CYSLTR2, MYH1* depicted in Figure 11. The genes in the candidate cluster are mutually exclusive (p-value 0.0084). The genomic landscape of this cluster is dominated primarily by mutations in *PTEN* and *HUWE1,* and secondarily by homozygous deletions in *PTEN* and *CYSLTR2*. The most frequently altered gene in this set is a common driver gene *PTEN*, a tumor suppressor gene that negatively regulates the AKT/PKB apoptosis pathway [133]. The remaining six genes in the cluster are potential driver candidates *HUWE1* is a part of the Mule multidomain complex of the HECT domain family of E3 ubiquitin ligases responsible for apoptosis suppression, DNA damage repair, and transcriptional regulation [134]. *CNTNAP2* is a neurexin protein with functions in cell-to-cell adhesion and an epidermal growth factor and was found to be hypomethylated in breast cancer cell lines [135]. Hypomethylation and the association with epidermal growth factors, coupled with a large number of amplifications in the alteration landscape of *CNTNAP2* suggest potential oncogenic functions of the gene. *GRID2* is an ionotropic glutamate receptor that is frequently deleted in lymphoma [136]. *CACNA1B* codes for a N-type calcium channel which is responsible for calcium influx. Defects in the calcium influx channel can lead to alteration in the apoptosis, proliferation, migration and invasion pathways of breast cancer [137]. *CYSLTR2* is a proinflammatory cysteinyl leukotriene receptor that plays a role in cancer cell differentiation and is associated with breast cancer survival rates [138]. *MYH1* is a myosin heavy chain protein that plays a role in cell signaling and pro-apotosis pathways.

Perhaps more important than the propensity of each individual gene to be a driver is the collective interaction pattern of the seven genes in the cluster in a cancer pathway. From Figure 11, it is clear that the each gene in the cluster interacts with each other in a tightly-connected community with no gene more than three nodes away when plotted in the network, using the cBioPortal visualization tool [127]. The seven genes in the cluster *PTEN, HUWE1, CNTNAP2, GRID2, CACNA1B, CYSLTR2, MYH1* are strong candidates to define a novel driver pathway.

This conclusion is reinforced by the presence of high impact common drivers *TP53, MYC, AKT, and PIK3R1* which define several important cancer pathways such as apoptosis, DNA repair, and cell cycle arrest [139], [140].We also examined a cluster containing potential cancer drivers relevant for GBM. In GBM, we found a cluster of size 10 with four known drivers and many potential drivers. The cluster includes *GLI1, WNT2, BRAF, PLCG1, FAS, CREBBP, BRCA2, GLI2, PIK3R5, VAMP3* (see Figure 12). This large cluster has a p-value of 0.0901 in terms of mutual exclusivity, which is actually low as compared to other GBM clusters. The cluster also contains several important driver genes such as *WNT2, BRAF, BRCA2* and *CREBBP* which encompass pathways such as sonic hedgehog signaling, cell fate determination, cell growth and apoptosis, checkpoint activation, and DNA repair. Additionally, six out of the ten members are within the same compact network community *GLI1, PLCG1, FAS, CREBBP, BRCA2, PIK3R5*. Of these six genes, *GLI1* and *GLI2* are hedgehog signaling genes that are common and first isolated in glioblastoma. These genes are responsible for cell differentiation and stem cell self-renewal [141]. *PLCG1* is involved in intracellular transduction of receptor-mediated tyrosine kinase activators, and it has been classified as a biomarker in GBM [142]. *FAS* is a cell surface receptor that mediates apoptosis. *FAS* is known as a histological hallmark of GBM, affecting both apoptosis and necrosis factors [143]. Finally, *PIK3R5* is a subunit of phosphatidylinositol 3-kinases who together have important effects on cell growth, proliferation, differentiation, motility, survival and intracellular trafficking.

3.4 Discussion

We described a novel method, termed C^3, which has the potential to precisely and efficiently identify clusters of gene modules with mutually exclusive mutation patterns. The C^3 algorithm uses large-scale cancer genomics datasets which are pre-processed to yield parameters governing novel constrained correlation clustering techniques. The optimization criteria used in clustering include patterns of mutual exclusivity of mutations, patient sample coverage, and network driver concentration.

There are several major advancements of our method when compared to previously known approaches. Unlike methods that use randomized approaches without the guarantee that multiple runs of the methods on the same data will produce compatible results (such as CoMEt), C^3 is "consistent" in so far that by running the same LP solver, the same results will be generated. Also, C^3 has computational complexity that does not depend on the chosen cluster sizes, and is hence much more appropriate for large cluster problems than other methods. Furthermore, it partitions the gene set and hence creates clusters covering all genes used in the analysis, although it may also be adapted to accommodate overlapping clusters. This is in contrast with the results produced by other methods that tend to identify only a small number of modules with limited number of genes.

None of the previous methods were able to identify clusters utilizing different sources of information via a weighting mechanism. This is important because it gives us flexibility to focus more on certain aspects based on the analysis. For example, we can focus more on mutual exclusivity instead of coverage to identify clusters specific to a group of samples which may facilitate the discovery of subtype-specific modules.

By addressing the above challenges, we believe our new method C^3 represents a unique tool to efficiently and reliably identify mutation patterns and driver pathways in large-scale cancer genomics studies.

FIGURES

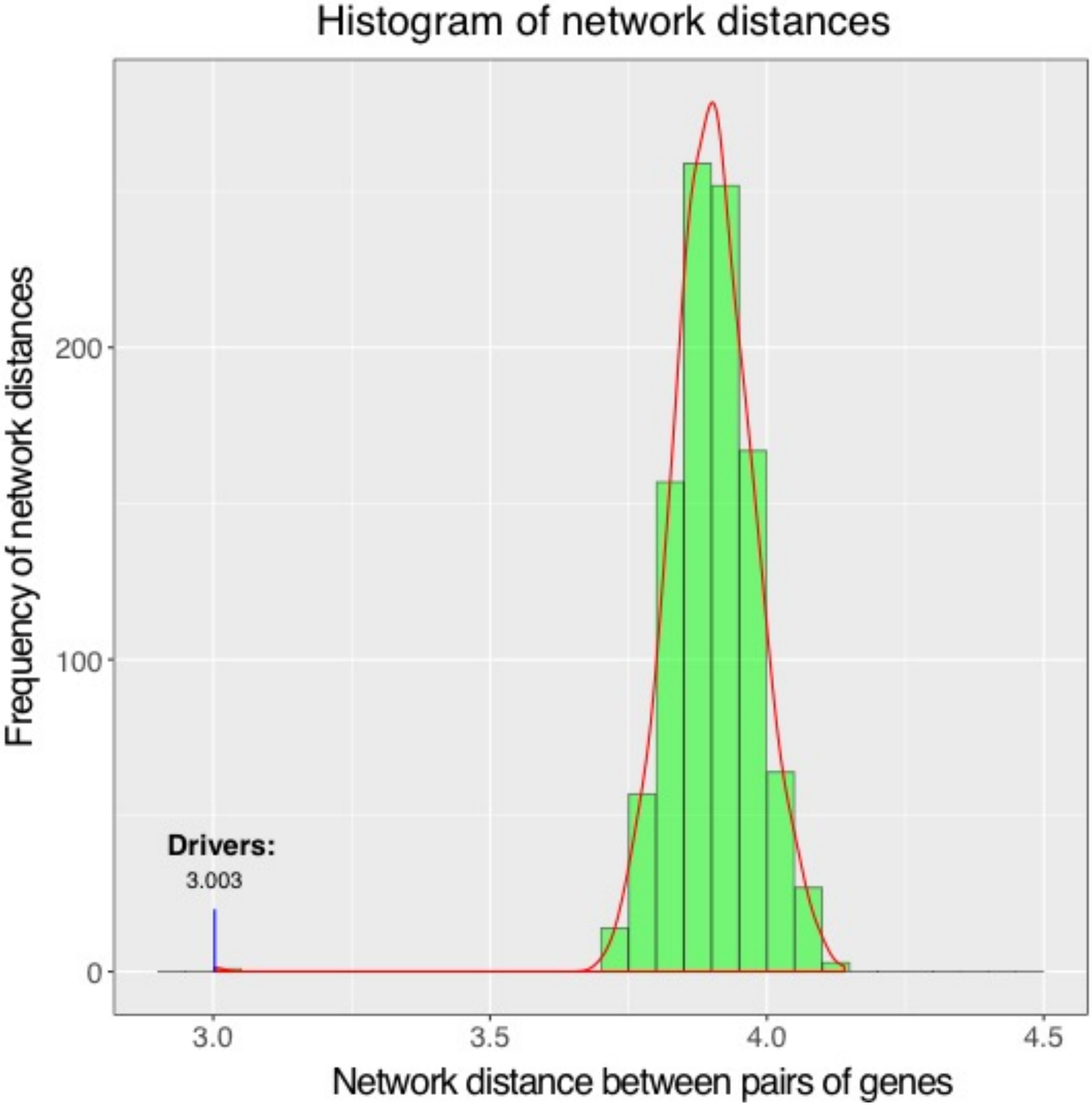


Figure 8: Histogram of shortest distances between randomly selected genes and driver genes in the network.

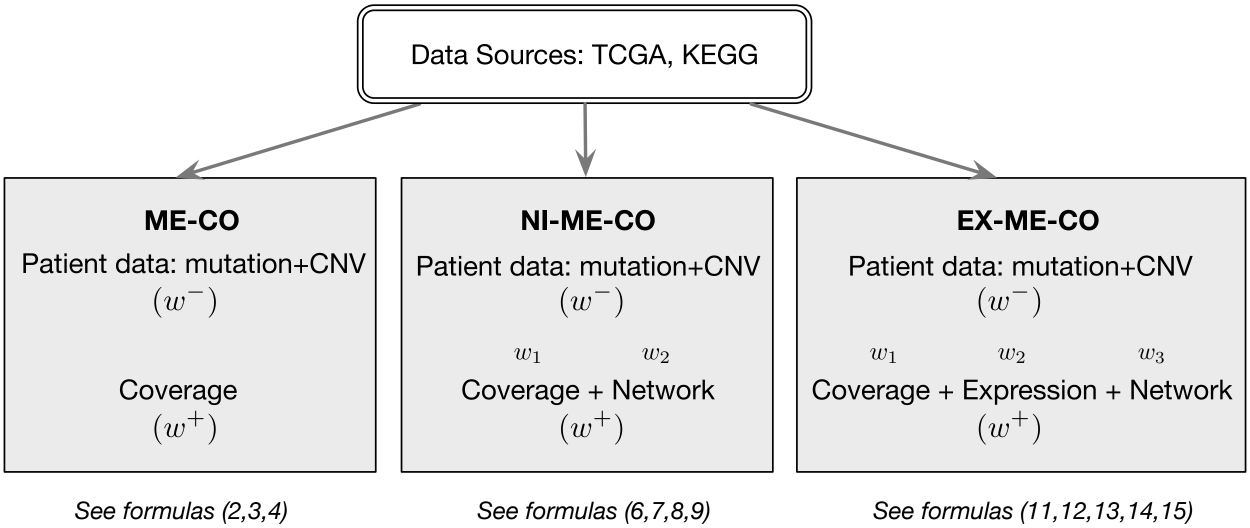


Figure 9: A workflow of C^3 displaying heterogeneous data sources converted into different clustering weights.

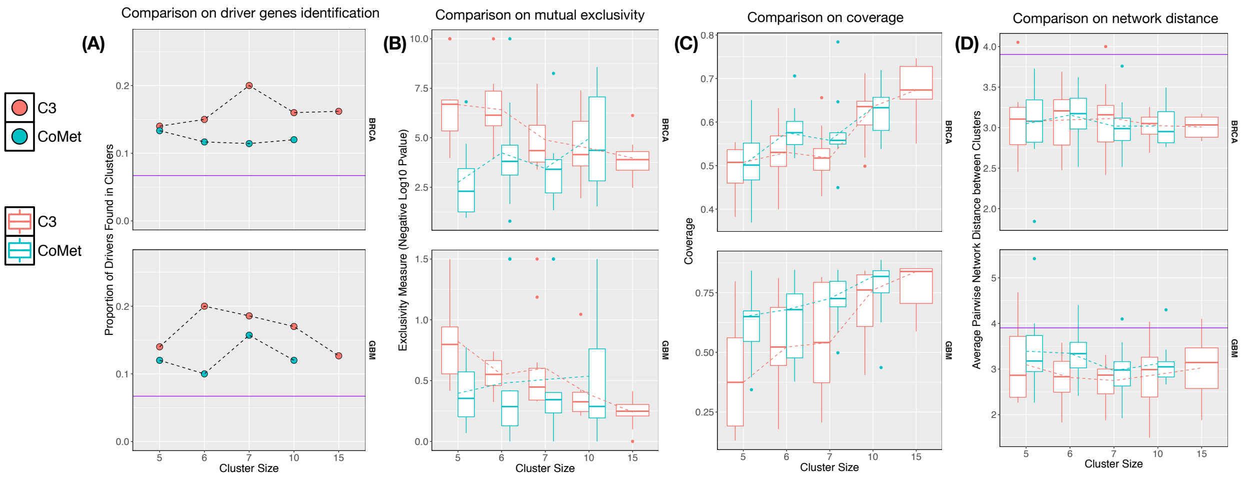


Figure 10: A comparative analysis of C^3 (Red) and CoMEt (Blue) based on four evaluation criteria. We used five cluster sizes (5,6,7,10, and 15) that index the x-axis in each benchmark test. **(A)** depicts the results based on the driver gene evaluation criteria. The y-axis represents the proportion of drivers found by each method, contained within the best ten clusters found. The purple line represents the expected value of drivers detected if clusters are randomly selected. **(B)** shows the pairwise mutual exclusivity of each run. The y-axis represents the negative log transform of the mutual exclusive p-value such that larger values are more mutually exclusive than smaller ones. The boxplots illustrate the distribution of exclusivity results concerning each of the top ten individual clusters for C^3 and CoMEt. **(C)** shows the distribution of coverage, measured by proportion of samples with at least one alteration in a given cluster (the y-axis). The boxplot illustrates the distribution of coverage results for individual top ten cluster results. **(D)** includes the network connectivity results of C^3 and CoMEt. The y-axis measures the average pairwise network distance between all genes in a cluster, and the distribution of each cluster is shown in the boxplot. The purple line represents the average pairwise distance of random clusters.

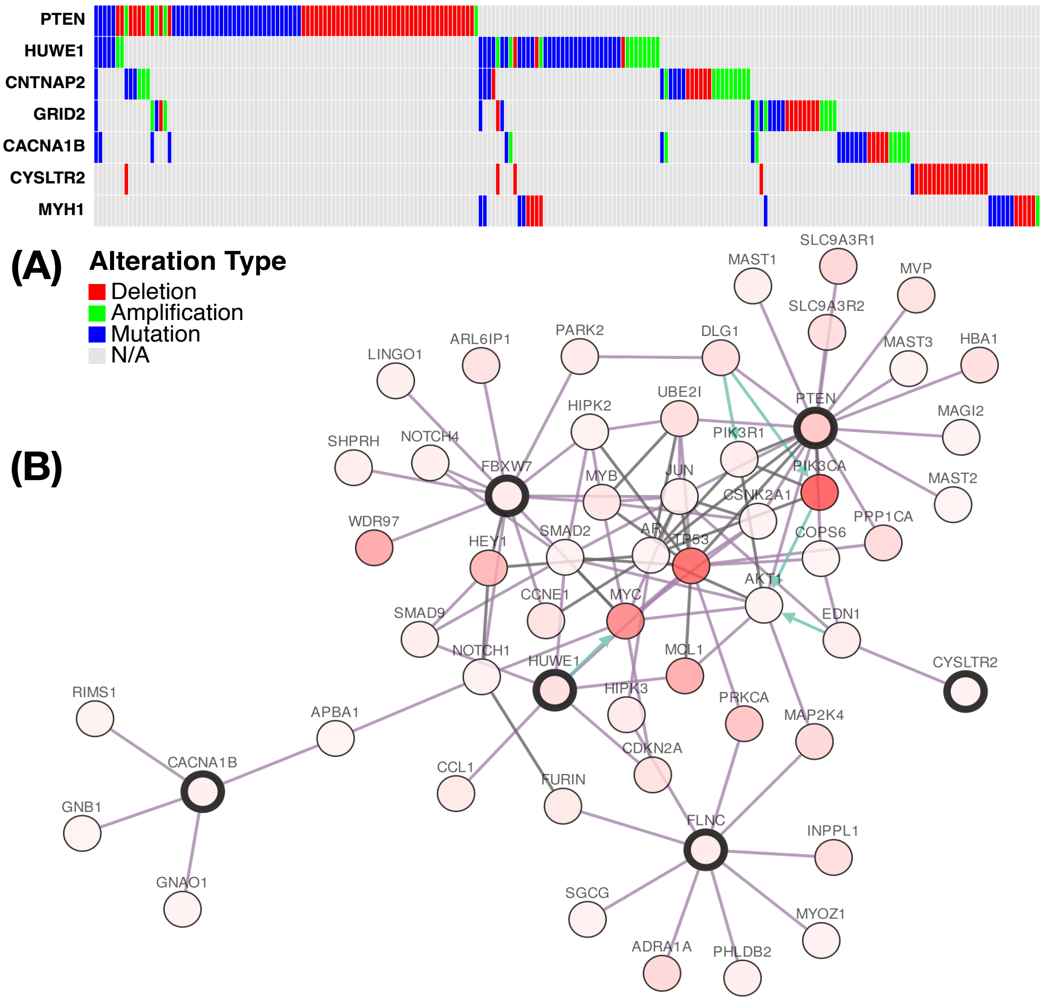


Figure 11: A cluster of potential driver genes inferred from BRCA. **(A)** shows the alteration landscape of the cluster, with blue representing mutation events, red representing copy number deletions, and green representing copy number amplifications. **(B)** represents a *known* subnetwork which contains 6 genes (out of 7) in **(A)**. The more intense the red, the higher the alteration frequency of the gene. Nodes highlighted in black represent driver candidates identified by C^3 within a small subnetwork. Edges are depicted in black if there exists a direct interaction between two genes. Green edges represent an interaction that undergoes a protein state change. Purple edges are other interactions.

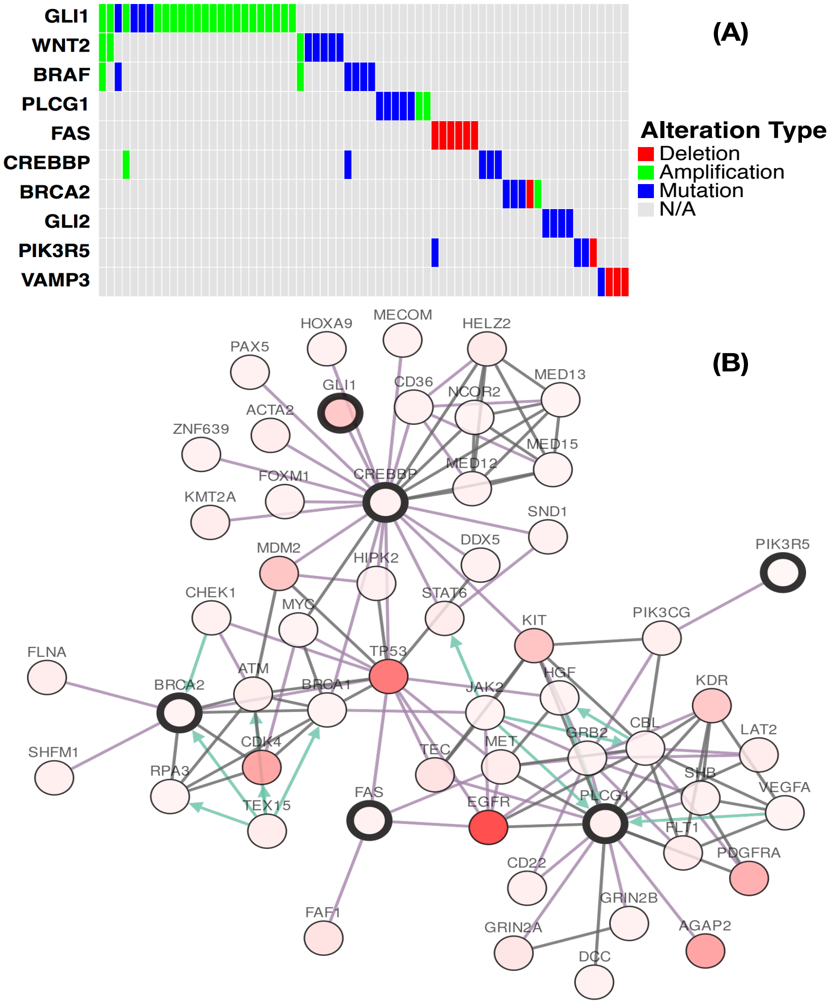


Figure 12: A cluster of potential driver genes inferred from BRCA. **(A)** shows the alteration landscape of the cluster, with blue representing mutation events, red representing copy number deletions, and green representing copy number amplifications. **(B)** represents a *known* subnetwork which contains 6 genes (out of 7) in **(A)**. The more intense the red, the higher the alteration frequency of the gene. Nodes highlighted in black represent driver candidates identified by C^3 within a small subnetwork. Edges are depicted in black if there exists a direct interaction between two genes. Green edges represent an interaction that undergoes a protein state change. Purple edges are other interactions.

CHAPTER 4: Scattershot: Personalized cancer drug recommendation

4.1 Introduction

The practice of oncology continually faces the challenge of matching cancer patients with an optimal treatment regimen. The challenge is especially daunting in cancer chemotherapy, where the success rate of cancer compounds meeting FDA approval for effectiveness and safety is a mere 13.4% [144]. The marginal success rate of cancer therapeutics is likely due to the enormous complexity of the disease mechanism of cancer coupled with an inability to properly match the drug to the patients where it would have the largest positive impact [145]. Cancer is a disease of the genome is driven by unique, patient specific, alterations that affect major pathways in growth, survival, and division [52]. One strategy that can be employed by physicians is to target the genome by prescribing targeted therapies in which treatments are tailor-made to individual patients that specifically target perturbations in the patient's genome [146]. In recent years, computational methods have been utilized to define and process the enormous swaths of data needed to identify the patient's genomic perturbations and predict the drug targets that work best for the patient.

The problem of developing computational tools to model drug treatment presents a set of major challenges. The interaction between cancer drugs and cancer cell lines is complex and not well not well understood in many cases [51]. Even though databases such as the Drug Gene Interaction Database (DGIdb) [147] have mapped out many of the interactions between drugs and the genome, the database is far from complete and many drug interactions with the genome and drug interactions with other drugs are unknown [148]. The context of the data is also imperfect. The majority of data available for drug response analysis comes from cell lines compendiums such as The Genomics of Drug Sensitivity of Cancer [53] and the Cancer Cell Line Encyclopedia [54]. However, experimental procedure differences between the major cell line compendiums have shown inconsistent drug response when the same drug is treated with the same cell line [59] Additionally, cancer tumors do not reside in a closed system. Tumors react closely with normal cells and the patient's environment [60]. This may limit the scope of many cell line-based studies of drug response.

Many pioneering studies concerning drug response have been made using cell lines. One of them was the NCI's DREAM7 initiative. [149]. The DREAM7 project was a community driven project where teams would predict the drug response of "hidden" BRCA cell lines using RNA-seq data from training cell lines. The winning methods in DREAM7 were a Bayesian kernel multitask model and an integrated Random Forest method. Since DREAM7, several contemporary methods have been developed predict the drug response in cancer cell lines. Such studies include the machine learning methods using GDSC and CCLE datasets such as SVM with Recursive Feature Elimination binary prediction approach in calculating acute drug response [55] and CancerDP, another drug prioritization method based on SVM with F-stepping feature selection [56]. The authors of [57] implemented a linear method which calculated the drug response of Lymphoblastic cell lines (LCL). An ensemble method utilizing the integration of multiple machine learning methods, PGM, was especially unique in that it simultaneously modeled chemical and cell line information together to make a prediction [150]. Most recently, a comprehensive study unconverted a list of features corresponding to Cancer Functional Events

(CFEs) and used those features to accurately predict the drug response of over 1000 human cell lines [58].

One important limitation of the cell line studies is that the extent that application of *in vitro* studies extends to conclusions of the treatment paradigm in real patients is still unknown [151]. Therefore, several studies in drug response have shifted focus from cell line data models to models based on real patient data. Many of these methods utilize patient data from The Cancer Genome Atlas [152]. The authors of [61] utilized a linear Ridge-Regression model to bridge the gap using *in vitro* gene expression models to make predictions *in vivo.* While gene expression models have shown a degree of success, gene expression models alone have been found to be insufficient in predicting drug response in some cancers [62]. The IntOGen platform also has a tool that assigns drugs to patients based on their proximity to the driver gene in a cancer network [63]. The identification of targetable genes was expanded using the EMD model, which identified a list of candidate drivers using integrated gene expression, mutation, copy number and network information with potential drug targets for the drivers [64]. Another method GOPredict [63] integrates both genomic and pathway data to provide a ranked drug list of potential targets [65]. While these methods provide a starting point in computational drug prediction, none of these the methods evaluate their approach using recorded actual drug response or the actual drugs that were prescribed to the patient. Rather, these methods rely on indirect comparisons of potential drug targets, or they only look at evaluating a few select patients, drug target, and drug response combinations.

Recently, the authors of [66] have presented a method that predicts the clinical response of a drug directly from *in vivo* molecular signature data. The authors of [66] used an ElasticNet Regression classifier to predict a physician-coded drug response on cancer patients using data from one type of feature ranging from mRNA or miRNA expression, methylation, or copy number. However, while the approach was new, it was also hindered by several limitations. Most drug-specific models in [66] exhibited poor performance due to lack of sophisticated feature selection and filtering coupled with the limitation of a small *n* large *p* (large number of features compared to a small number of samples) and the inability to build models using multiple types of features.

To address the limitations of previous methods, we developed a novel drug response prediction and drug prioritization algorithm called Scattershot. Scattershot models the problem of drug response and drug recommendation as a multilabel machine learning problem in which multiple response variables (labels) are predicted simultaneously while accounting for the interactions among the labels [69] where we develop ensemble classifiers from multiple genomic sources such as mutation, expression, copy number, and pathway-level information as well as clinical variables. Scattershot uses the multilabel framework to build binary classifiers that predict the drug response of an individual drug while at the same time, aggregating the results of multiple pairwise binary classifiers comparing pairs of drugs in a drug list to prioritize the drug rankings. Scattershot's integrated approach has outperformed previous methods in predicting drug response in actual patient data, and its novel recommender has consistently ranked actual prescribed drugs high in a large majority of patients.

4.2 Results

4.2.1 Method overview

Here we provide an overview of the Scattershot algorithm. Detailed method description is in the Methods section. Figure 13 provides an overview of the whole method. Scattershot is a multilabel machine learning algorithm that predicts multiple responses (a list of drugs) with two modes: 1) single drug mode (SDM) and 2) pairwise recommendation mode (PRM). Single drug mode uses a classifier to predict the clinical response of a single drug in a group of test patients. Pairwise recommendation mode is a multilabel classification method that ranks a list of drugs in a test patient ordered from most to least likely to respond by simplifying the multilabel problem into a combination of binary label classifiers and rank-aggregating the binary classifiers to provide the final rank list. The first step of Scattershot is feature selection. Scattershot assembles features from multiple different sources. These sources include genomic features from expression, mutation and copy number information, drug target interaction data in a human gene pathway context as well as other user input features such as clinical information. The feature selection step for genomic information includes only genomic features that have been proven to have a biological and clinical significance to drive cancer. The significant cancer gene filter was assembled from 3 sources: 1) DawnRank [67] 2) Cancer Functional Events (CFEs) [58] and 3) the cancer gene census [101]. The second step in the Scattershot process is the machine learning classifier. This was done as a binary classifier using Random Forest with recursive feature elimination, RFE to further whittle down extraneous features. The response variable for the Random Forest differs in the single drug mode and pairwise recommendation mode. In SDM, the binary classifier calculates effectiveness (did the drug work?) for a test patient while in pairwise recommendation mode, the binary classifier predicts a preference between any two drugs (which of the two drugs would work better?). For Scattershot in PRM, multiple binary classifiers comparing all combinations of any two drugs were rank-aggregated in a FAS-pivot algorithm to identify a preference list using the multi-label learning framework.

We ran Scattershot on two TCGA drug response datasets: a Pan-Cancer dataset consisting of 1508 samples, and a breast cancer dataset consisting of 647 samples. Within each dataset, we performed two analyses: a single drug mode analysis to quantify the drug response of a single drug, and a pairwise recommendation mode to rank-order potential drug treatments for any given patient. The single drug mode was done on 4 breast cancer drugs and 7 additional Pan-Cancer drugs. We limited our drug response information to the same information used in [66] to compare the performance of the two methods as closely as possible. We then used the results of the classifier to determine whether our drug prescription is associated with any clinical outcomes such as survival. The PRM analyses ranked a list of 11 breast cancer drugs and 22 Pan-Cancer drugs for each patient, and we visualized the data in terms of its pairwise classifier performance, overall rank precision performance, and its ability to recommend novel and/or infrequently prescribed drugs.

4.2.2 Scattershot accurately predicts drug response for single drugs

We first used Scattershot in SDM to build a drug response classifier for each drug to evaluate the overall performance of the model. We limited our results to drugs with physician-coded response in at least 5% of prescribed patients, leaving 4 breast cancer drugs and 11 Pan-Cancer drugs (including the 4 breast cancer drugs). We chose 5% because it is the smallest cutoff in all analyses where the number of cross-validation samples n will be larger than the average number of features p, which avoids the small n large p problem [153]. At a 5% cutoff in breast cancer, for example, a drug has to have at least 30 physician coded responses . The average number of features in a breast cancer single drug mode is 24.25. We evaluated our method using standard binary metrics in sensitivity, specificity, AUC, and accuracy. We also used at the Cohen's kappa coefficient which evaluates the model performance by comparing to a chance agreement. The kappa statistic is defined as: $(p_o - p_e)/(1 - p_e)$, where p_e is the expected probability that the classifier will output the result by chance and p_o is the observed probability that the classifier will output the result. In other words, given a 2×2 confusion comparing the classifier predictions with the true results.

	(+)	**(-)**
(+)	*a*	*b*
(-)	*c*	*d*

p_o is the observed accuracy of the confusion matrix $(a + d)/(a + b + c + d)$ while p_e is the expected probability of random agreement, calculated by the sum of the marginal probabilities $(marginal_a + marginal_b)/(a + b + c + d)$ where $marginal_a = ((a + b)(a + c))/(a + b + c + d)$ and $marginal_b = ((d + c)(d + b))/(a + b + c + d)$.

A kappa statistic is similar to a correlation measure and the output ranges from -1 to 1 where 1 indicates a perfect classification in which all predictions made by the model are not by chance or guessing, -1 indicates that the classifier built on random guessing always performs better than the model, and 0 indicates no distinguishable difference between the model and the result by chance or guessing. We interpret the kappa statistic using Landis and Koch’s approach where 0-0.2 is “weak”, 0.2-0.4 is “fair”, 0.4-0.6 is “moderate”, 0.6-0.8 is “substantial”, and 0.8 to 1 is “almost perfect” [154].

Figure 14A shows the mean AUC performance of three of the drug response classifiers in BRCA from 10-fold cross validation, and reported the median value from the classifer. We excluded Doxorubicin because it had highly skewed class imbalances where <10% of the data was a disease state, which did not yield enough data points to accurately assess the performance of the Doxorubicin classifier. Hence, we observed excellent performances of Doxorubicin accuracy of 91%, AUC of 90%, but a poor Doxorubicin kappa of 0.02. The remaining single drug classifiers in breast cancer had more reliable results, with an 88.3% AUC for Anastrozole and a “moderate” kappa of 0.463, a 94.1% AUC for Tamoxifen and a “fair” kappa of 0.384, and an 81.8% AUC for Paclitaxel and a “moderate” kappa of 0.435.

In Figure 14C, we extended our Scattershot single drug mode classifiers of the Pan-Cancer dataset. With many more samples, the Pan-Cancer dataset allows us to evaluate more drugs than the BRCA dataset alone. The median of AUC of the classifiers is 86.1% with a median kappa of 0.277. The Pan-Cancer results provide analysis of drugs that we were not able to categorize in the breast cancer analysis due to lack of data. This includes Cisplatin (AUC: 90.9%), Carboplatin (AUC: 82.0%), Cyclophosphalamide (AUC: 85.5%), Doxorubicin (AUC: 86.7%), Gemcitabine (AUC: 76.5%), and Temzolomide (AUC: 74.3%). We next determined whether or not the patients that we predict to respond to a certain drug would actually have a clinically significant response. We used survival analysis in breast cancer patients to determine whether patients with a Scattershot predicted positive response to Anastrozole, Paclitaxel, or Tamoxifen treatment would have a different clinical outcome in terms of survival (see Figure 14B). We looked at TCGA clinical 5-year survival data. The significance of survival was calculated using a chi-squared test, and we found that the patients predicted to respond to Anastrozole and Tamoxifen exhibited a statistically higher survival rate than patients expected to respond poorly to these drugs (p-value 0.012 and p-value < 2e-16, respectively). No significant difference in survival was found in Paclitaxel (p-value 0.516).

Scattershot identifies patients in which Anastrozole and Tamoxifen administration significantly improve survival. Both Anastrozole and Tamoxifen inhibit aromatase, an enzyme that synthesizes estrogen. Unsurprisingly, Aromatase Inhibitors are often prescribed for ER+ breast cancer patients [155]. We examined features of the Anastrozole and Tamoxifen classifier to identify which features in the model are the most important survival indicators in ER+ breast cancer. Tamoxifen has the *ESR2* drug target as its fourth most common feature. The ESR2 gene codes for Estrogen Receptor beta, a key pathway in ER+ breast cancer [156]. Clinically, *ESR2* is widely targeted in BRCA, and the *ESR2* molecular marker is highly correlated with survival [157]. Anastrozole is another drug that serves as a survival predictor . In Scattershot's Anastrozole classifier, the second most important feature is *CTNNB1*. *CTNNB1* coordinates cell-to-cell adhesion and gene transcription, and it promotes the *Wnt* signaling pathway, a prominent signaling pathway which controls cell fate specification, cell migration, and G1/S cell proliferation. *CTNNB1* and the *Wnt* signaling pathway is commonly perturbed in ER+ breast cancer [158]. Like with *ESR2*, *CTNNB1* is also widely targeted and it is strongly associated with BRCA survival [159]. Survival markers such as *ESR2* and *CTNNB1* in ER+ breast cancer explain how Scattershot identifies patients where Tamoxifen and Anastrozole can be administered to improve survival.

4.2.3 Scattershot achieves better performance in single drug prediction than previous methods

We compared Scattershot to the method described in [66], which was the method most similar to Scattershot where it attempts to predict the same physician-coded drug response. Using the same response variables, the method in [66] built 4 models corresponding to expression, miRNA, copy number, and methylation for Paclitaxel in BRCA (no other BRCA-specific drugs were reported in that work), and reported that their best model was the miRNA model, which had a mean AUC performance of 67.3%. In contrast, Scattershot's average Paclitaxel AUC performance is 81.8%. [66] also modeled the drug response for Carboplatin and Cisplatin with respect to the Pan-Cancer analysis. Scattershot's performance respect to Cisplatin AUC is 90.9%, and Scattershot's performance with respect to Carboplatin AUC is 81.2%. In contrast, the strongest Cisplatin model in [66] was miRNA with an AUC 68.4% and the mRNA expression model of Cisplatin had an AUC of 62.6%. The best Carboplatin model in [66] was expression with an AUC of 58.0%. These performance results suggest that Scattershot provides a framework that can better predict the drug response of single drugs.

Three potential reasons Scattershot shows stronger performance than previous methods. 1) Scattershot identifies a data integration step that allows for the integration of data from multiple sources, whereas the model presented in [66] only allows for one type of genomic data. 2) Scattershot's feature selection step includes an initial cutoff that selects for significant cancer genes. 3) Scattershot's inclusion of pathway features, and drug target features present add crucial clinically significant predictive features.

Scattershot is able to incorporate multiple data types (binary variables, continuous variables, whole numbers, and integers), using the non-parametric Random Forest classifier which can incorporate multiple types input data to be present in the model without requiring normalization steps which result in information loss [160]. To explore the impact of data-integration, we ran Scattershot using one type of data only (expression-only model, copy number-only model, and mutation-only model) to predict drug response. For Paclitaxel in BRCA, the best model performance was gene expression with an AUC of 70.0%, lower than the fully-integrated model of AUC 81.8%. For Pan-Cancer Cisplatin, the best model performance was the mutation model, with an AUC of 75.2%, lower than the fully-integrated Scattershot model of 90.9%. For Pan Cancer Carboplatin, the best model performance was the expression model, with an AUC of 61.9%, lower than the fully-integrated Scattershot model of 82.0%.

We then examined the impact of Scattershot's initial feature selection step. Unlike the previous method, Scattershot's feature selection includes an initial cutoff of functional cancer genes from three sources: 1) known drivers found by the Cancer Gene Census [101], 2) drivers that have pathway impact from DawnRank [67], and 3) drivers that are associated with functional events in cancer [58]. This criteria leads to a feature space of 65 expression features, 76 copy number features, and 76 mutation features. The smaller feature space reduces the chances of overfitting and prevents the model from selecting genomic features that provide little or no functional impact. To quantify the impact of this initial feature selection, we ran Scattershot without the initial feature elimination step and compared our results to the full Scattershot model. For Paclitaxel in BRCA, no feature selection yielded an AUC of 70.7%, higher than the model in [66] which had an AUC 67.3% but lower than the fully-integrated model of AUC 81.8%. For Cisplatin in the Pan-Cancer analysis, no feature selection yielded an AUC of 81.4%, lower than the fully-integrated Scattershot model of 90.9%. For Pan-Cancer Carboplatin, no feature selection yielded an AUC of 52.1%, lower than the fully-integrated Scattershot model of 82.0%.

Lastly, we examined the impact of pathway and drug target features in the model. The Paclitaxel model for BRCA, for example, consists of several features from drug targets and pathway features. This model consists of 10 features, and three of those features are pathway features: Cell Cycle Control (the most important feature), Receptor Tyrosine Kinase (RTK) signaling and Folate Transport. Additional two features are the *PTEN* and *KRAS* drug target. Mechanistically, these new features are quite important in Paclitaxel response. Paclitaxel mainly serves to enhance the polymerization of tubulin to stable microtubules, which are required to pass the G2/M phase of mitosis [161]. This explains why the Cell Cycle Control pathway variable is the single-most important predictor of Paclitaxel response. Receptor tyrosine kinases are cell surface receptors polypeptide growth factors, cytokines, and hormones that are key regulators in many cell processes. Paclitaxel and Trastuzumab (Herceptin) target RTK and are associated with stronger drug response [162]. Folic acid targets cell membranes and enhances endocytosis of nanoparticles, which facilitates the uptake of Paclitaxel to cancer cells, increasing its bioavailability [163] [164]. *PTEN* is a phosphatase and tensin homolog that plays a major role in cell cycle progression and proliferation [165], and the *PTEN* signaling pathway has been linked to reversing chemoresistance to paclitaxel in p53 mutated cancer cells [166]. *KRAS* is a GTPase and is an early player in many signal transduction pathways [167], and Paclitaxel has been involved as a chemotherapeutic agent in *KRAS* mutated cell lines to improve drug response [168].

4.2.4 Scattershot can accurately predict drug in pairwise recommendation mode

We next ran Scattershot in pairwise recommendation mode to provide a ranked list of drugs using the results of the pairwise preference classifiers. As with the single drug classifier in SDM, we studied any drug that was prescribed in at least 5% of the patients to keep the *n* samples larger than the *p* features. This resulted in 11 eligible breast cancer drugs to rank and 22 eligible Pan-Cancer drugs to rank. Unlike the single drug mode, we were only concerned with whether a drug was prescribed, not whether the drug had a disease state or response outcome (see Methods). This was done in part to increase the number of drugs to rank to provide meaningful ranking results, in part because the vast majority (68%) of the prescribed patients exhibit a positive response, and in part to simplify the problem to keep the pairwise classifier a binary classifier for the rank aggregation method. The goal of each pairwise drug classifier is to determine whether or not a test patient would prefer one of the two drugs based on the patient's genomic profiles.

The results of the pairwise preference classifiers for breast cancer are shown in Figure 15. Figure 3A shows the AUC evaluation and Figure 15B displays the kappa statistic for each pairwise preference classifier. In breast cancer, we evaluated all but 3 of the potential 55 pairwise drug classifiers. We excluded Cyclophosphamide-Trastuzumab, Cyclophosphamide-Epirubicin, and Cyclophosphamide-Fluorouracil from the study due to severe class imbalance, where almost all (>99%) or none of the patients that were treated with Trastuzumab, Epirubicin, or Fluorouracil were also treated with Cyclophosphamide.

For breast cancer, the average pairwise AUC is 82.8% with an average accuracy of 84.3%. Only one pairwise response, Epirubicin vs. Doxorubicin, out of 52 did not have a significant AUC when comparing with the model with a 0.5 AUC baseline. The average kappa statistic is 0.381. Looking at the kappa statistic, 10 of the 52 classifiers had a “slight” kappa score of 0-0.2, 20 of the 52 classifiers had a kappa score of 0.21-0.4, 11 of the 52 classifiers had a kappa score of 0.41 to 0.6, and 11 of the 52 classifiers had a kappa score of 0.61 or higher. The distribution of kappa scores indicates that while the performance of the classifiers in general are not due to chance. However, the association is not very strong in many cases. One explanation for low kappa scores in some classifiers and high kappa scores in others may be related to drug mechanisms. Anastrozole, Exemestane, and Letrozole are all aromatase inhibitors with very similar mechanisms in estrogen receptor positive BRCA [169]. Due to drug response similarity, Scattershot has difficulty in comparing these drugs which explains why the kappa value of Anastrozole performs worst when paired with Exemestane and Letrozole, exhibiting a kappa of 0.06 and 0.22.

With regards to the Pan-Cancer analysis, the AUC results are higher than the breast cancer predictors with an AUC of 93.6%, and the kappa statistic is 0.765, and only 19 out of 209 classifiers returned a kappa of less than 0.2. This indicates that the overall classifiers for Pan-Cancer analysis is very strong and that the vast majority of classifiers' performance is not by chance. The kappa score for the pairwise classifier is substantially higher than they are for the breast cancer data meaning that the performance of Pan-Cancer classifiers is much less likely to be due to chance than in breast cancer. There are several differences in the data that are potential sources of this discrepancy. One reason is the larger *n*. The average preference classifier in Pan-Cancer contains 2.3 times as many samples as the average preference classifier in breast cancer. Patients across multiple types of cancer have more distinguishable genomic features than patients within only one type of cancer. Cancer mechanisms vary from cancer to cancer, and some chemotherapeutic drugs are cancer-specific. An example of this is Bleomycin in Testicular Cancer (TGCT). Bleomycin is heavily prescribed in TGCT. 53 out of 162 TGCT patients, but it is not prescribed in patients in any other cancer.

We next looked at the feature selection process for each of the subtypes in breast cancer. Figure 15C highlights the most commonly selected features for each source of data with respect to the pairwise preference classifier. Selected features in Figure 15C represent the most commonly selected features when building a pairwise preference classifiers regarding the drug. With regards to expression data, the most selected feature is *GATA3,* which was heavily selected in the models of every drug, and *EGFR* and *ERBB2*, which are important features in 10/11 drug models. These three genes are known to be highly predictive of breast cancer subtypes, which are often used to prescribe drugs *GATA3* along with *BRCA1* is involved in pathogenesis of basal and triple negative breast cancer [170]. *EGFR* and *ERBB2* are well-known for their driving potential in Her2 breast cancer [171]. For copy number analysis, *PIK3R1* is the most selected feature in breast cancer. *PIK3R1* is selected in 8/11 drug models. *PIK3R1* activated in response to activations in tyrosine kinases such as *EGFR, VEGFR2,* and *ERBB2,* and its involvement in multiple cancer pathways make it an ideal marker for predicting drug response [172]. *BRCA2* is the most important mutation feature in the pairwise preference classifiers. *BRCA2* is selected in 10/11 drug models. The *BRCA2* mutation, is a well-known hereditary mutation is involved in DNA repair mechanisms [173]. *BRCA2*-induced cancers are more likely to be ER+ and less likely to be Her2, and therefore, it is a strong treatment marker for ER+ prescribing drugs [174]. Pairwise classifiers involving Epirubicin and Docetaxel rely heavily on drug target features. Epirubicin-based classifiers use 80% of the available drug target features and Docetaxel-based features use 67% of the available drug target features. Both Epirubicin and Docetaxel have very similar drug target with common features such as *ABCC6, NAT2, XRCC3, PRDX2, PLD2, SLCR10A2, TUBB,* and *NR112*. Docetaxel and Epirubicin are often co-prescribed with targeted chemotherapy to improve drug response [175]. Some features such as multi-drug resistance proteins such as *ABCC6* are associated with resistance to Docetaxel and Epirubicin treatment [176]. Pathway features may be quite drug specific, but many of the pathway features selected by our model agree with the current literature in breast cancer treatments. For example, 5-Fluorouracil and Epirubicin are associated with telomerase length, and these two drugs have the telomere maintenance pathway as an important select breast cancer pathway [177]. In addition to genomic data and drug target-driver gene interaction, there are also important clinical features in breast cancer predictors, including ER status and Triple Negative status. This agrees with prior knowledge as the ER status or the lack thereof is often the most important current manual decision-making steps in breast cancer drug prescription as drugs [155].

4.2.5 Scattershot in pairwise recommendation mode rankings are well associated with the prescribed drugs

After building the pairwise response matrix, we used the FAS-Pivot algorithm to rank aggregate the pairwise comparisons and output a ranked list of drugs for each patient by relevance (see Methods). To evaluate the rankings, we compared the ranked list of drugs for each patient to the drugs that were actually prescribed to the patient by the health care provider. We used the precision @ *k* score to test the performance. The precision @ *k* score measures the precision, the percentage of drugs that were prescribed vs. all Scattershot proposed drugs at rank *k* [178]. Figure 16A illustrates the precision @ *k* score for each breast cancer drug. The precision @ 1 for Scattershot is 73.1% and the precision @ 2 is 56.6%. This means that for all patients, Scattershot's top recommended drug was actually prescribed to the patient over 73% of the time, which indicates that Scattershot is able to reproduce a substantial number of physician recommendations in cancer. Scattershot's second choice was prescribed over 56% of the time. The precision curve decreases with *k*, meaning that predictions ranked low are unlikely to be actually prescribed.

To confirm that Scattershot recommends prescribed drugs higher than it does drugs that were not prescribed, we compared the distribution of the rankings for each drug when it was prescribed with the distribution of the rankings of the drugs when they were not prescribed (Figure 16B). A Mann Whitney Rank-Sum test showed that the rank difference between the Scattershot rank when the drug was prescribed compared to when it was not was statistically significant. The p-value for each breast drug test was less than 0.05 for all 11 Scattershot breast cancer drugs, meaning that Scattershot rankings for each drug were higher when the drug was actually prescribed. Similar results were found with the Pan-Cancer analysis, with a Mann Whitney Rank-Sum test showing that the rank difference was statistically significant for all 22 drugs ($p<0.05$).

In breast cancer, the largest difference between Scattershot rankings is in Trastuzumab, or Herceptin. Scattershot's rankings for Trastuzumab in patients where Trastuzumab was prescribed was 2.0 (See Figure 16B) while the median ranking for Trastuzumab where Trastuzumab was not prescribed was 11.0. Trastuzumab is often prescribed in aggressive Her2 BRCA patients, targeting growth factors such as *ERBB2* and *ERBB3*, activating the *PIK3* apoptosis pathways, and contributing to inhibiting cancer angiogenesis [179]. We looked at the features of the Trastuzumab pairwise classifiers and found that the most commonly selected features include *ERBB2* and *ERBB3* expression, copy number, and mutation features, *PIK3R1* copy number, the Her2 clinical subtype, and the Angiogenesis pathways. Each of these features is a hallmark property in Trastuzumab response which may largely explain the reason the Trastuzumab ranking is so reliable.

4.2.6 A cluster of Scattershot rankings reveals subtypes that are consistent with known breast cancer subtypes

We further visualized the Scattershot personalized drug rankings by clustering the breast cancer patients based on the predicted drug rankings using hierarchical clustering with Ward's linkage over the Spearman's footrule distance. This resulted in five clusters from Scattershot recommendations. We compared the breast cancer clusters based on Scattershot rankings to that of established cancer subtypes such as the ER/PR/Her2 [39] subtype classification as well as the commonly accepted PAM50 gene expression subtypes of Basal, Her2, Luminal A, Luminal B, and Normal-like [42] (see Figure 16C). We found that the Scattershot subtypes correspond very closely to each clinical subtype as well as each PAM50 subtype. A chi-square test for association was performed to determine the significance of the clusters, and a significant association was found between Scattershot drug recommendations subtypes with both the ER/PR/Her2 subtype (p-value < 2e-16 for each marker) and PAM50 subtypes (p-value < 2e-16). We found 5 subtypes in breast cancer based on drug prescription that were strongly associated both the clinical and gene expression subtypes. Three of the clusters (named Clusters 1, 3 and 5 in Figure 16C) are related to ER+ and PR+ breast cancer and the Luminal A and Luminal B PAM50 subtype (Chi-square association test p <2e-16). Another cluster (Cluster 2) is related to the Her2+ subtype and the PAM50 Her2 subtype (Chi-square association test p <2e-16). The last cluster (cluster 4) is related to triple negative breast cancer and the PAM50 Basal subtype (Chi-square association test p <2e-16).

We then compared the drugs that are associated with each cluster, using a Mann-Whitney Rank-Sum test to compare the rankings of each drug within the cluster to those outside the cluster to determine if there is a significant association between the drug and the cluster. We used a Bonferroni correction for multiple testing. The ranking of Trastuzumab, which is associated strongly with the Her2 subtype, is strongly associated with high ranks in cluster 2 which corresponds to many Her2 patients (p <2e-16) and is also associated with low ranks in clusters 1, 3, and 4. Aromatase inhibitors in Anastrozole and Letrozole as well as anthracycline drugs in

Epirubicin and Doxorubicin are significantly associated with Cluster 1, a Luminal A and B and ER/PR+ subtype. Clusters 2 and 5 are other Luminal A and B and ER/PR+, however, and while most of the Cluster 1 drugs are still significant in clusters 3 and 5, the defining feature separating Clusters 3 and 5 from Cluster 1 is Tamoxifen, which is not significantly associated in Cluster 1, but is associated in Cluster 3 (p=1.68e-6) and cluster 5 (p=8.33e-5). Although Tamoxifen acts targeting the ER receptor, its mechanism is different from other ER drugs in that it causes a change in the folding of the steroid binding domain that prevents gene activation [180]. The drugs that define clusters 3 and 5 is Docetaxel which is not significant in Cluster 3, but significant in Cluster 5 (p = 1.63e-13). Docetaxel has been shown to effectively treat ER+ breast cancer patients, but the efficacy varies due to the level of ER expression [181].

The most interesting conclusions from Figure 16C come from Scattershot's recommendations for Triple Negative Breast Cancer. Triple Negative Breast Cancer are defined by the lack of ER, PR, and Her2 receptors, and they are known for their low survival rates due to the lack of targeted therapies available [182][40]. In Triple Negative Breast Cancer, Scattershot tends to rank the drugs Cyclophosphamide, Fluorouracil, Epirubicin, and Doxorubicin high with statistical significance. Evidently, this corresponds to literature studies which show that both CEF (Cyclophosphamide, Epirubicin, Fluorouracil) and CDF (Cyclophosphamide, Doxorubicin, Fluorouracil) chemotherapies outperform traditional chemotherapy regimens for TNBC patients [183]. Triple Negative Breast Cancer patients are shown to be sensitive to anthracyclines such as Epirubicin and other DNA destabilization agents to a degraded DNA repair cascade in TNBC [184]. This result further demonstrates that the Scattershot clusters can be used to stratify breast cancer patients to well defined drug response subtypes. Therefore, Scattershot may be a useful tool which could help physicians select a treatment by taking into account integrated genomic, drug target, and clinical outcome, complementing current strategies of identifying clinically relevant subtypes.

4.3 Methods

Scattershot operates under one of two modes: 1) single drug mode (SDM) and 2) pairwise recommendation mode (PRM). Single drug mode uses a classifier to predict the clinical response to a single drug in a group of test patients. Pairwise recommendation mode is a multilabel classification method (a method that predicts multiple responses simultaneously) that provides a ranked list of drugs in a test patient ordered from most to least likely to respond. We use multilabel problem transformation techniques [185] to transform a comparison of many drugs to a comparison of any two drugs in a binary classifier in which test patients are classified a "preference" between the two drug. The preferences binary classifiers are then rank-aggregated using the pairwise rank aggregation FAS-Pivot algorithm to determine a final rank (see the multilabel Rank Prioritization Section). The classifiers in both modes are based on Random Forest incorporating a wide array of integrated features from a multitude of data sources including molecular signature data in expression, mutation, and copy number information combined with drug target, pathway interaction, and clinical data (see Figure 13).

4.3.1 Data Collection

All genomic features and drug response data were downloaded originally from TCGA [152]. For common genomic features, we obtained preprocessed and curated data from cBioPortal [127]. We used the genomic information including: mRNA expression (median z-score), copy number information (based on GISTIC [128]), and mutation information. Although methylation information was shown by [58] to improve drug prediction models on GDSC cell lines, we excluded methylation data in this work due to a large number of missing values in breast cancer samples (29%) and an absence of reliable methylation HumanMethylation27 BeadChip information in ovarian cancer. We obtained genomic information from 15 cancers with a substantial number of patients with both genomic feature information and drug information: BLCA (bladder cancer), BRCA (breast cancer), CESC (cervical cancer), GBM (glioblastoma), HNSC (head and neck squamous cell), KIRC (kidney cancer), LGG (low grade glioma), LUAD (lung adenocarcinoma), LUSC (lung squamous cell carcinoma), OV (ovarian cancer), PAAD (pancreatic cancer), PRAD (prostate cancer), TGCT (testicular cancer), UCS (uterine carcinoma).

Drug response information was also obtained from TCGA through the Broad institute [186] and the TCGABiolinks R package [187]. Drug response was recorded in the TCGA as one of five outcomes. The five outcomes, ordered from best to worse, are as stated: “Complete Response”, “Partial Response”, “Stable Disease”, “Radiographically Progressive Disease”, and “Clinically Progressive Disease”. The names of the recorded drugs and treatments in TCGA, however, are not standardized and require curation due to the use of formatting differences and the use of differing names for the same drug (e.g., Generic and Trade Name) [65]. The paper [66] provides a dictionary which translates all TCGA prescriptions to a standardized DrugBank ID [188] for all treatments that were prescribed in a patient that yielded a drug response. We added curations to original dictionary to include prescriptions that yielded a missing or unknown responses using the same methodology from [66], standardizing prescription names to the corresponding DrugBank ID [188].

We included missing values in the pairwise recommendation mode analysis for several reasons. First, the vast majority of drug prescriptions in TCGA (68%) have no corresponding drug response. This can contribute to the small *n* large *p* problem where the number of features is much greater than the number of samples to train [189], or it can severely limit the number of drugs that we can apply to Scattershot, reducing the scope of the problem. Second, the mere prescription of the drug in an actual clinical setting implies that the physician believes that the drug will elicit a positive response in the patient. A majority of patients treated with any drug will respond favorably. 63.9% of patients prescribed with a certain drug had a "Complete Response", 70.4% patients had a "Complete" or "Partial" response, and 83.0% of the patients had a "Stable Disease" response or higher. Therefore, we included all drug prescription information with missing and unknown values in order to increase the power of our classifiers.

For drug target specific features, we gathered gene network and pathway information with respect to drug target and mutations. We used the gene network of 8726 genes from [67], which is a network combined with curated KEGG data [190] as well as non-curated interactions from [191]. We used 13 cancer pathways defined as "General" cancer pathways from cBioPortal [127]. Drug targets were obtained using the Drug Gene Interaction Database (DGIdb) [147]. Although, its compendium is not complete, the DGIdb compendium is one of the most extensive databases to characterize drug interactions with the genome, integrating data from multiple well known databases. For mutation information, we focused only on mutations in known driver genes defined by the Cancer Gene Census [101]. All data was accessed on 7/9/2016. All in all, 1508 patients across 15 cancers were analyzed. Of that, the cancer type with the drug information was breast cancer, with 647 patients. Because breast cancer is the cancer with the most prescription information by far, we did two analyses on drug response: one with only breast cancer patients and the other Pan-Cancer analysis with all 15 cancers. With the dataset, we made a training and test dataset with the training dataset consisting of 90% of the data and the test dataset consisting of the remaining 10%. Ten training datasets were created this way for a 10-fold cross-validation.

4.3.2 Binary Classification

Both single drug and pairwise recommendation modes use binary classification at the heart of their method. In single drug mode, the binary response variable is 0 if the treatment elicits no drug response and 1 if the treatment does elicit a drug response. We define a drug response as a recorded outcome corresponding to a response state: either “Complete Response” or “Partial Response”. We define no response as a recorded outcome corresponding to a disease state: either “Stable Disease”, “Clinically Progressively Disease”, or “Radiographic Progressive Disease”. In pairwise recommendation mode, we created a binary classifier for each pairwise drug comparison using a training dataset of 90% of the data and a testing dataset of 10% of the dataset with 10-fold cross validation. Each pairwise binary classifier represents a preference of one of two drugs. If both drugs are prescribed in the same patient, the preference goes to the drug that elicits the stronger response, and the patient is removed from the classifier if both drugs elicit the same response.

Algorithm 4.1 listed below explains the building of the binary classifier. The first set of Algorithm 4.1 is to calculate process the features used in the model. We initially process the genomic features **F**, the distance features **T**, and the pathway features **S**, separately and merge them together to make the combined feature space for the model. We then implemented a Random Forest classifier with recursive feature elimination to predict the final outcome. The feature elimination step involves building the model with all the features, eliminating the features that provide the least amount of information according the Random Forest GINI index, and rerunning the model again until the model performance no longer performs better than the previous model. In single drug mode, it is the physician-coded drug response, and in pairwise recommendation mode, it is an indicator of which drug is most likely to be prescribed.

We selected Random Forest as the binary classification model for three main reasons. 1) Random Forests have few parameters to train. 2) Random Forests do not require normalization 3) Random Forests are more robust (though not immune) to small *n* large *p* problems. The only parameters that Random Forests are required to train are the number of features for each decision tree and the number of trees that make up the forest [192]. This is less than similarly performing methods such as SVM which require more parameters for any given kernel. Random Forests also do not require normalized data. Random Forests are better able to handle small *n* large *p* because it is an ensemble method aggregating results of multiple models (Random Trees) with a small number of features [193]. To evaluate our method, we ran Random Forest with 10-fold cross validation training with 10, 50, 100, 500, 1000, and 5000 and we found that 5000 trees yielded the highest performance. We trained 1 to 10 features per tree and we selected the parameter on a per-model basis based on the 10-fold CV result, and we reported the median result from our classifier. We used the R package RandomForest to analyze our model.

The genomic features **F** used in each binary classifier fall under one of three groups: (i) genomic features, (ii) drug-specific pathway target features, and (iii) clinical features. The genomic features include information from mutation, expression, and copy number information. Because the number of potential features from this data is large (~33,000) and can contribute to the small *n* large *p* problem, we limited genomic features to features that satisfy each of the following criteria: (i) genomic features with known tumorigenic properties, or driver genes; (ii) genomic features with network impact; and (iii) genomic features that have been previously identified as clinically relevant in cell line studies. We used the 580 driver genes in the Cancer Gene Census (CGC) [101]. Highly impactful genes were calculated by DawnRank [67], selecting a corresponding number of highly-ranked, significant, impactful genes in the cancer pathway. We used the cell line study [58] to detect Cancer Functional Events (CFEs), which are features from cell line data associated with drug response. Using the intersect of all three of the following criteria, we used 65 expression features, 76 copy number features, and 76 mutation features. For drugs that had DGIdb drug targets outside these 76 genes, the expression, copy number, and mutation features of those target genes were also included in the model. For breast cancer, we also used basic clinical information as well. We used patient information such as age as well as specific tumor staging (T, M, N information) [194] and subtype and tissue type information

[195]. The subtypes used for the breast cancer analysis are the clinical subtype (ER, PR, Her2) and not the PAM50 subtype.

In addition to basic genomic and clinical features, we also sought to include information that quantifies the interaction between the drug's targets with the patient's driver genes. This is matrix **T** in algorithm 4.1. The drug target / driver gene interaction has been hypothesized to be predictive of drug response. Studies such as [63] operated under the paradigm that drugs that directly target a driver gene or target a gene that interacts with the driver should be candidates for targeted therapy. The authors used the drug target / driver gene interaction to assign targeted therapy to patients based on how close the drug target was to the patient's driver genes. We also calculated the drug target / driver gene interaction in our model. For each drug target, we calculated the Dijkstra's shortest path distance [196] for the drug target corresponding to the patient's nearest predicted driver mutation (shortest Dijkstra's path distance). This outputs a distance feature for each drug target. In pairwise recommendation mode, we consider two drugs at a time, so we used the drug targets of both drugs as features. For each drug, we also calculated the absolute minimum distance between all drug targets and all driver genes. The absolute minimum distance represents the smallest possible interaction distance between any drug target and any driver gene, which indicates the overall most likely mechanism in which a cancer drug would act on the patient.

```
function CLASSIFY($d1, d2, \mathbf{F}$)          ▷ Where $d1$ - Drug1, $d2$- Drug2 (Optional), $\mathbf{F}$ - Features
                                                 ▷ Output 1 if patient should be prescribed with Drug1, 0 otherwise

    Let $\mathbf{F}$ be common genomic features (Mutations, Copy Number, Expression.. etc)
    Let $\mathbf{T}$ be distance matrix for each drug target and it's closest driver alteration
    Let $\mathbf{S}$ be the shortest drug target driver gene interaction passing through each pathway

    $\mathbf{T_{d1}}$ = calculateTargetDriverInteraction($d1$)
    $\mathbf{S_{d1}}$ = calculatePathwayInteraction($d1$)
    $\mathbf{F}$ = merge($\mathbf{F}, \mathbf{T_{d1}}, \mathbf{S_{d1}}$)          ▷ Add specific features concerning drug 1

    if $d2$ exists then
        $\mathbf{T_{d2}}$ = calculateTargetDriverInteraction($d2$)
        $\mathbf{S_{d2}}$ = calculatePathwayInteraction($d2$)
        $\mathbf{F}$ = merge($\mathbf{F}, \mathbf{T_{d2}}, \mathbf{S_{d2}}$)      ▷ Add specific features concerning drug 2
    end if

    $\mathbf{F}$ = recursiveFeatureElimination($\mathbf{F}$)                 ▷ Remove Redundant Features
    return randomForest($\mathbf{F}$)                                         ▷ Random Forest Classifier

end function
```

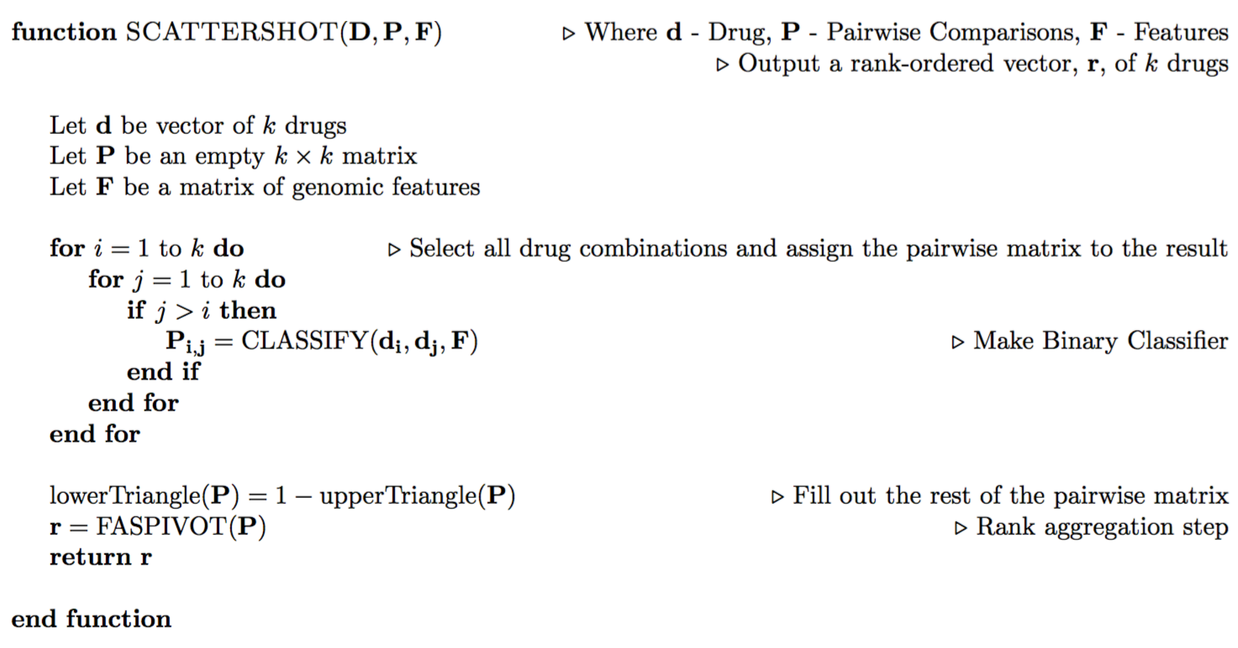
(Algorithm 4.1)

To complement the drug target-driver gene interaction, we further complemented our model with pathway information. This is matrix **S** in Algorithm 4.2. Driver genes affect tumorigenesis by acting on cancer pathways which act in tandem to produce a phenotypic effect. Important pathway features such as PIKC3A/AKT's effect on apoptosis have been shown to be clinically significant features when perturbed in cell line models, which showed that drug target and driver gene interacting within the same pathway have an impact on predicting drug response [58]. We captured this type of interaction by mapping the distance of a driver to a specific cancer pathway by calculating the shortest path distance between the any driver gene with any gene in the pathway. Values closer to 0 indicate that the pathway interacts more closely with a patient's drivers while values equal to 0 indicate that the pathway is directly perturbed in a patient. We used the R package igraph to calculate the pathway interaction values [197].

The features in the model **F**, **T**, and **S** were then merged to form the feature space, which range from 217 to 265 total features. Although this feature space is much smaller than the potential 30,000 features possible in our model, the limited availability of treatment information in many drugs still leaves us prone to overfitting from the small *n* large *p* problem. To rectify this, we used Recursive Feature Elimination (RFE). RFE eliminates redundant or irrelevant features to yield the most precise set of genes with the greatest predictive accuracy, and it has also been shown to have high predictive power in predicting cell line response [198]. RFE works by building a full model and calculating its performance, then rank ordering each variable by its importance, and then eliminating the least important features from the model and reevaluating the method to determine if there is an improvement in performance. The importance for our RFE was the Gini coefficient, the entropy calculation of Random Forest classifiers. The RFE process is repeated until there is no improvement in the model from eliminating features. The RFE step in our model was built using the R package, Caret [199]. One instance of Scattershot takes 1 hour and 15 minutes on an 8 GB ram personal computer.

4.3.3 Pairwise Rank Prioritization

For Scattershot to run in single drug mode, only the binary classification step is needed. However, the drug prioritization step of pairwise recommendation mode requires an additional step to create the ranked list for drug prioritization. Algorithm 4.2 below describes the Scattershot approach. Comparing the results of many related labels (in this case, drugs in vector **d**), is a challenging machine learning problem because the labels do not act independently, and a multilabel framework is designed to better handle inter-dependencies and interactions of the labels (drugs in this case) [69].



```
function SCATTERSHOT($\mathbf{D}, \mathbf{P}, \mathbf{F}$)    ▷ Where $\mathbf{d}$ - Drug, $\mathbf{P}$ - Pairwise Comparisons, $\mathbf{F}$ - Features
                                                  ▷ Output a rank-ordered vector, $\mathbf{r}$, of $k$ drugs

    Let $\mathbf{d}$ be vector of $k$ drugs
    Let $\mathbf{P}$ be an empty $k \times k$ matrix
    Let $\mathbf{F}$ be a matrix of genomic features

    for $i = 1$ to $k$ do    ▷ Select all drug combinations and assign the pairwise matrix to the result
        for $j = 1$ to $k$ do
            if $j > i$ then
                $\mathbf{P_{i,j}} = \text{CLASSIFY}(\mathbf{d_i}, \mathbf{d_j}, \mathbf{F})$    ▷ Make Binary Classifier
            end if
        end for
    end for

    $\text{lowerTriangle}(\mathbf{P}) = 1 - \text{upperTriangle}(\mathbf{P})$    ▷ Fill out the rest of the pairwise matrix
    $\mathbf{r} = \text{FASPIVOT}(\mathbf{P})$    ▷ Rank aggregation step
    return $\mathbf{r}$

end function
```

(Algorithm 4.2)

We use binary problem transformation to convert the multi-label problem into single label problems. We compare two labels at a time for every combination of labels in one-vs-one comparison over the feature space **F** described in the previous section. This differs from the traditional one-vs-all problem formulation in which a classifier is built for each drugs in that in that it maps drug interactions and dependencies where one-vs-all methods cannot. One-vs-all problem formulation is the default model used most current methods in predicting drug response in cell line data, including the most recent [58].

The result of each classifier indicates whether or not a test patient is more similar to patients prescribed with one of two drugs, resulting in a “preference” of one of these two drugs. The advantage of using problem transformation is that we account for any potential interactions and interdependencies which may confound our data while at the same time simplifying the problem to binary, single-label, preference. The pairwise preferences, calculated as the median result from the 10-fold CV, are used to build a pairwise comparison matrix, **P**. **P** is unique for every patient in the test set. When comparing *k* drugs, **P**, is a *k* x *k* matrix holding the result of the drug preferences (results of the binary classifiers) for the test patient in which $\mathbf{P}_{i,j}$ is 1 if the patient is more likely to prefer drug *i* and 0 if the patient is more likely to prefer drug *j*. When *i =j,* no value is given in the matrix. A pairwise rank aggregation step is then done on **P** to obtain the final result. Scattershot's model follows that of pairwise classification described in [200], which showed that pairwise, one-vs-one classification can be utilized to output promising experimental results compared to traditional one-vs-all methods.

Algorithm 4.3 listed below describes the FAS-Pivot pairwise rank aggregation algorithm to determine the final drug rankings, which is a special case of the FAS-Tournament sports algorithm designed to rank sports teams in the wake of a large amount of inconsistent information [201]. FAS-Pivot provides a globally consistent rank solution when there is potential for large number of disagreements and inconsistent information in the pairwise ranking matrix **P**. FAS-Pivot is especially important for drug prediction in patient samples because the dataset itself is subject to many confounding factors beyond the scope of genomic data that may lead to inconsistent information. Confounding factors include patient demographics, patient medical histories, and environmental information which are not well recorded and difficult to adjust [202]. FAS-Pivot works by first selecting a random drug pivot *q* among all drugs in **P**. Afterwards, it splits each remaining drug into one of two vectors $\mathbf{v_L}$ and $\mathbf{v_R}$. $\mathbf{v_L}$ contains all drugs preferred over the pivot and $\mathbf{v_R}$ contains all drugs not preferred over the pivot. A pairwise matrix **P[$\mathbf{v_L}$,$\mathbf{v_L}$]** consisting of only drugs prefered over the pivot and then FAS-Pivot is run recursively with the input **P[$\mathbf{v_L}$,$\mathbf{v_L}$]**. The results are appended to the left of the pivot. A pairwise matrix **P[$\mathbf{v_R}$,$\mathbf{v_R}$]** is also made for drugs not preferred over the pivot, and then FAS-Pivot is run recursively with the input **P[$\mathbf{v_R}$,$\mathbf{v_R}$]**, and the results are appended to the right of the pivot. The algorithm runs the input of FAS-Pivot is a 1 x 1 matrix in which only the pivot is returned or a 0 x 0 matrix in which nothing is returned.

```
function FASPIVOT($\mathbf{P}$)                     ▷ Where $\mathbf{P}$ - Pairwise Comparison Matrix
                                                    ▷ Output a rank-ordered vector, $\mathbf{r}$, of $k$ drugs

    Let $\mathbf{P}$ be an $k \times k$ matrix of all pairwise comparisons
    Let $q$ be a random pivot vertex from all drugs in $\mathbf{P}$
    Let $\mathbf{v_R}$ be vector representing drugs ranked higher than the pivot
    Let $\mathbf{v_L}$ be vector representing drugs ranked lower than the pivot

    $q = \text{selectPivot}(\mathbf{P})$                               ▷ Select Random Pivot

    for $i = 1$ to $k$ do
        if $\mathbf{P_{q,i}} < 1$ then
            $\mathbf{v_L} = \text{append}(\mathbf{v_L}, i)$                  ▷ Add vertex to $\mathbf{v_L}$
        else
            $\mathbf{v_R} = \text{append}(\mathbf{v_R}, i)$                  ▷ Add vertex to $\mathbf{v_R}$
        end if
    end for

    $\mathbf{r} = \text{append}(\text{FASPIVOT}(\mathbf{P}[\mathbf{V_L}, \mathbf{V_L}]), q, \text{FASPIVOT}(\mathbf{P}[\mathbf{V_R}, \mathbf{V_R}]))$  ▷ Recursively split matrix and re-rank
    return $\mathbf{r}$

end function
```

(Algorithm 4.3)

FAS-Pivot has a distinct advantage over traditional rank-prioritization methods such as Condorcet Voting in that it will output a ranked list in all circumstances while Condorcet Voting may be trapped in cyclical ranks [203]. Cyclical ranks are avoided by FAS-Pivot because FAS-Pivot forces a rank by comparing all drugs to a single pivot. Missing values in the pairwise rank aggregation matrix were imputed by allowing downstream calls to rank the drug and applying rank-balancing as seen in [201] by using Spearman’s footrule distance to obtain a consensus rank over 100 FAS-Pivot calls. We implemented FAS-Pivot in R and used the R RankAggreg package for rank balancing via Spearman’s footrule distance [110].

4.4 Discussion

The results of Scattershot highlight a method that can both predict the response of a single drug as well as rank-prioritize a list of drugs for any given test patient. Running Scattershot in Single Drug Mode found that Scattershot greatly outperforms previous methods in terms of predicting the response of the drug. Scattershot models several drugs such as Anastrozole and Tamoxifen using genomic markers that are indicative for survival. In Pairwise Recommendation Mode, we found that the pairwise classifiers predict the assigned drug with a high accuracy. The rank list of Scattershot recommendations indicate that the most recommended drugs are drugs that were actually prescribed with a high precision. BRCA subtypes based the Scattershot rankings are highly predictive of previously defined BRCA subtypes such as clinical subtypes and PAM50 gene expression subtypes.

The Scattershot method does have its limitations, and further work needs to be done to confirm and improve the results of Scattershot. One limitation of Scattershot lies within the quality of data in Scattershot. This extends to both the qualitative nature of the drug response information in the TCGA patient histories which are subject to subjectivity by physicians. Additionally, some of the input data may need to be examined further. One example of this lies in the drug target information found in DGIdb. The database, while extensive is not complete, and it draws from drug target information from multiple sources. To ensure the highest quality in results with respect to drug target information, manual curation is a potential future step to fill in potential gaps of the non-curated drug target database.

Another main difficulty in the classification step of Scattershot was the small number of samples for any given classifier. The small *n* may be one of the leading explanations to some of the low Cohen’s Kappa score in some classifiers. As more data is recorded, Scattershot may improve over time by more reliably predicting the response and prescription of more drugs. As more drug information becomes available in the future, we also plan to improve the Scattershot pairwise classifiers so that they simultaneously take into account prescription and the magnitude of response rather than a simple binary to determine drug prescription.

Scattershot is a new computational method that can help clinicians prioritize potential drug treatments and predict the response of a certain drug to a patent. Researchers can utilize the Scattershot pipeline to select for important features that define the drug response for specific drugs. The pairwise classifier may also provide insight of drug-interactions as it directly compares the response of two drugs. Taken together, Scattershot shows strong promise in its application to predict drug response and to recommend drugs on a personalized basis.

FIGURES

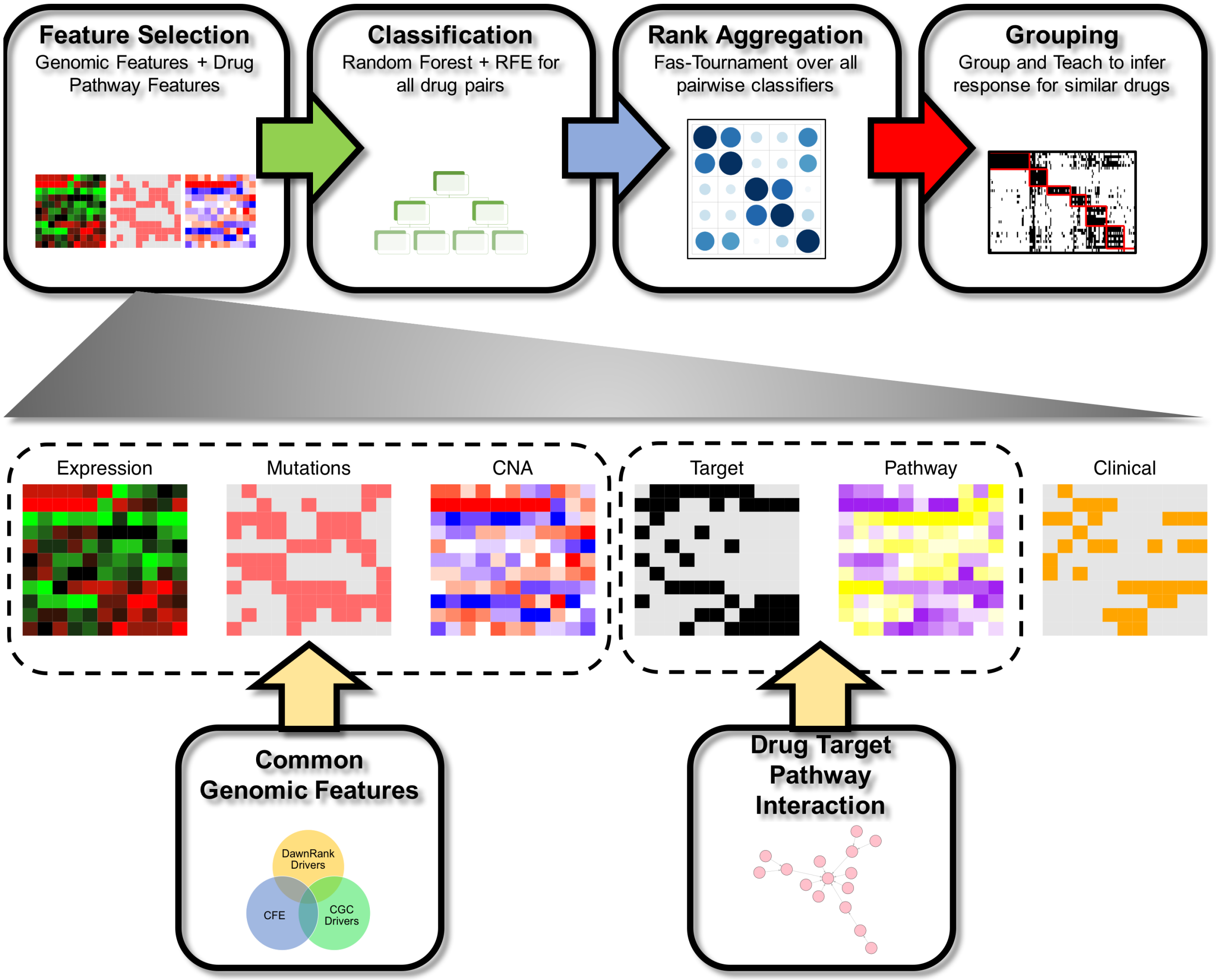

Feature Selection
Genomic Features + Drug Pathway Features
Classification
Random Forest + RFE for all drug pairs
Rank Aggregation
Fas-Tournament over all pairwise classifiers
Grouping
Group and Teach to infer response for similar drugs
Expression
Mutations
CNA
Target
Pathway
Clinical
Common Genomic Features
DawnRank Drivers
CFE
CGC Drivers
Drug Target Pathway Interaction

Figure 13: The workflow of Scattershot. It describes the steps necessary to run Scattershot in pairwise recommendation mode. In single drug mode, the drug response in the output so the method stops at the classification step. The lower portion of the figure illustrates the features and the feature types selected by Scattershot.

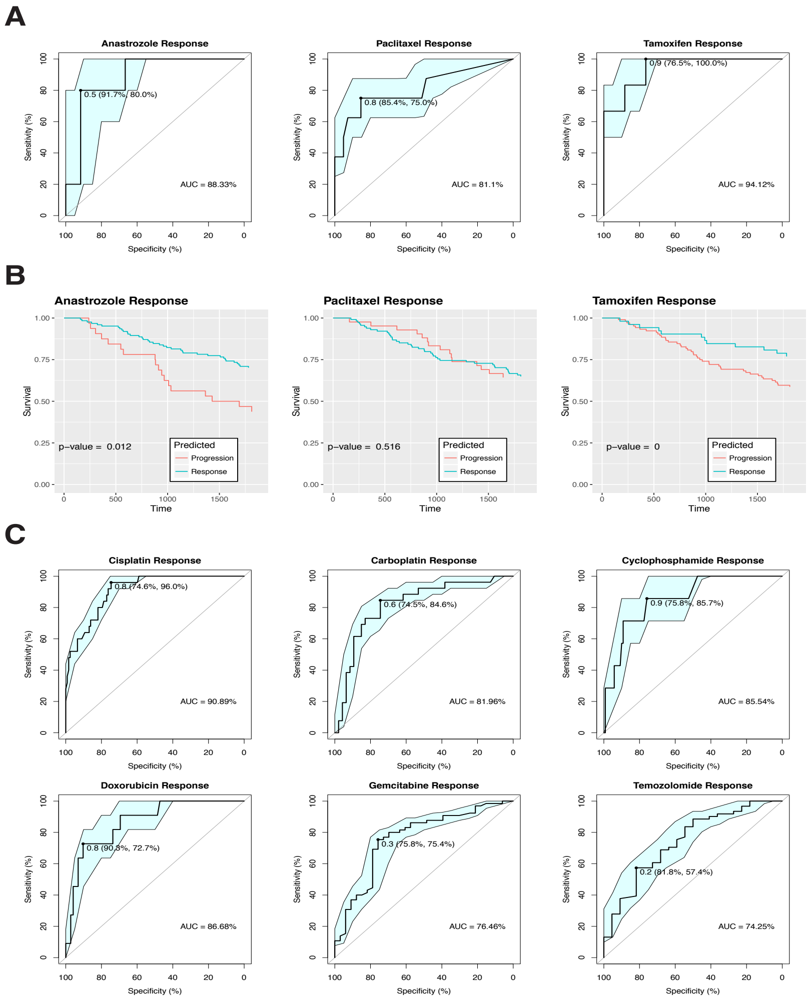


Figure 14: The performance evaluation of Scattershot run in single drug mode to measure drug response. (A) The ROC plots show the results of single drug mode Scattershot with breast cancer patients only in terms of Specificity (X axis, in reverse) and Sensitivity (Y-axis) for …. (B) The survival differences are shown for patients with predicted positive responses to the drug in question vs. predicted negative responses. The X axis represents survival time in days up to 5 years and the Y axis represents the percentage of patients surviving. (C) The ROC plots show the results of single drug mode Scattershot with Pan-Cancer patients in terms of Specificity, the true negative rate compared to all negatives (X-axis, in reverse) and Sensitivity, the true positive rate compared to all positives (Y-axis).

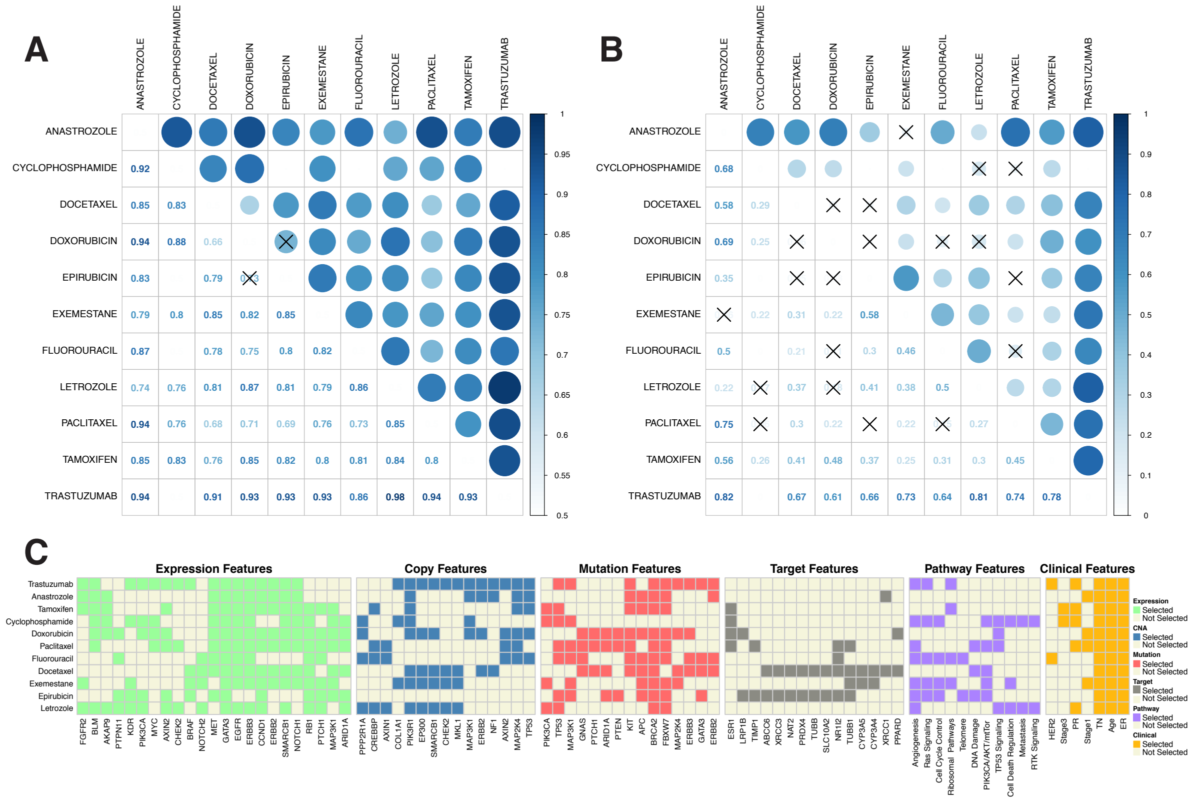


Figure 15: A summary plot of the breast cancer performances of the pairwise binary relevance classifiers when Scattershot is run in pairwise recommendation mode (PRM). (**A)** The median AUC values from the 10-fold CV of each pairwise classifier are shown in the lower triangle and the corresponding color and size intensity are in the upper triangle. (**B)** The C value of each pairwise classifier are shown in the lower triangle and the corresponding color and size intensity are in the upper triangle. **(C)** We show the feature selection variables for each type of data. The X-axis represents the feature type and they Y-axis represents the drug. A value is colored if the feature was selected in at least 25% of the models involving the drug.

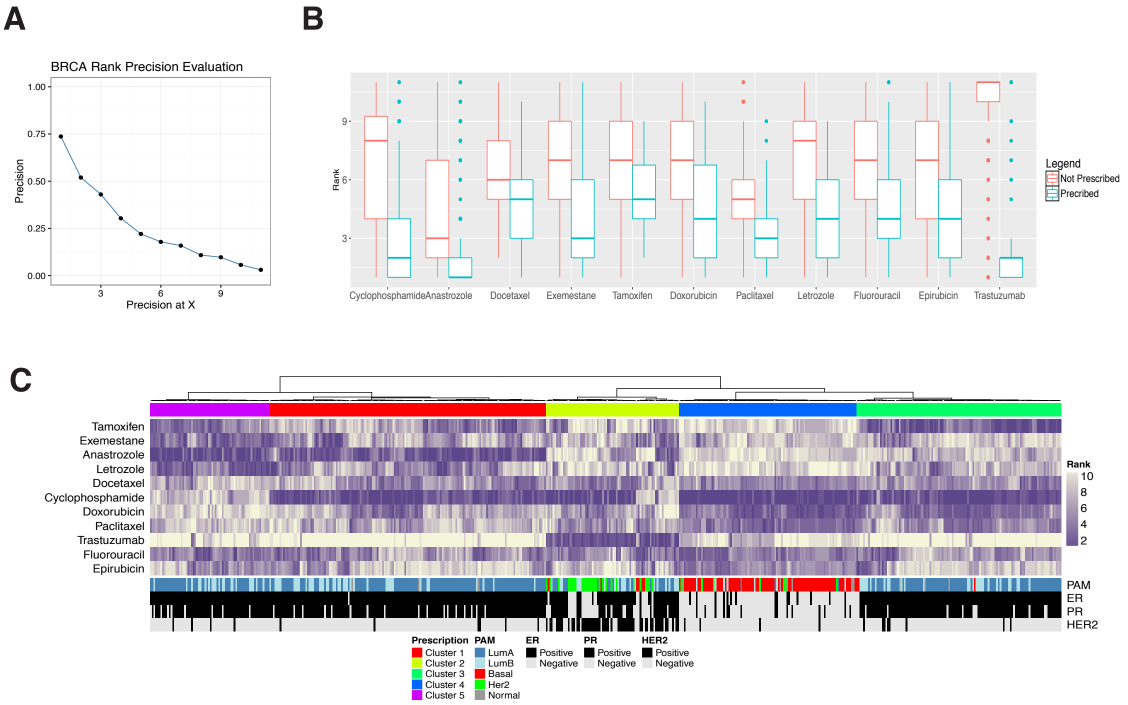


Figure 16: A summary plot of the recommendations in breast cancer after Scattershot's Rank Aggregation step was performed on the pairwise binary relevance classifiers in pairwise recommendation mode. (**A)** We show the Precision @ *k* Score with the X-axis indicating the ranking and the Y axis indicating the Precision at that ranking. (**B)** We visualize the rank distribution of all 11 breast cancer drugs (X axis) ordered by prescription frequency, between all Scattershot ranks where the drug was actually prescribed (Blue-Green) and the ranking when not prescribed (red). The Y axis represents the final rank. Note that lower rankings indicate the top Scattershot recommendations while upper rankings indicate the worst Scattershot recommendations. (**C)** A clustering landscape of breast cancer patients (X axis) and drugs (Y-axis) is shown. The intensity of purple signifies higher rank. The Spearman's footrule Hierarchical Clustering is seen at the top followed by the *k*=5 split for Scattershot clusters. The bottom tracks indicate the clusters of other breast cancer subtypes in PAM50 and ER, PR, and Her2 markers.

CHAPTER 5: Conclusions

5.1 Summary

The process of bringing cancer treatment models from the lab bench to the patient bedside remains one of the daunting challenges in making personalized medicine a reality. This dissertation identifies several key aspects of this ordeal and proposes several new computational methods to overcome these challenges. First, our thesis identifies the drivers of cancer. We built the method DawnRank which integrates mutation data, gene expression, and network information to discover drivers in a personalized manner that is geared towards finding especially rare and novel drivers which may have been masked by previous methods. We further demonstrated the power of DawnRank by using it to identify driver subtypes in BRCA. DawnRank, coupled with Consensus Clustering found 5 novel subtypes in BRCA while defining driving chromosomal hotspots of copy number alterations in breast cancer, including 1q amplification, 8q amplification, 11q loss, and 16q loss. Three subtypes correlate highly with the Luminal A subtypes, one with Basal/Her2, and the final with LumB/Her2. Additionally, the subgroups correlate with known clinical markers such as the estrogen and progesterone receptors with the Luminal subtypes, TP53 mutation in the Basal/Her2 subtypes, and worsened overall survival in the Basal/Her2 subtype. DawnRank's BRCA subtype analysis provides a proof-of-concept which can be used to stratify patients into subgroups that can later be defined by potential personalized treatment.

We also addressed the concept of multiple drivers and the pathway-level impact involved in cancer progression. We described a novel method, termed C^3, which has the potential to precisely and efficiently identify clusters of gene modules with mutually exclusive mutation patterns. The C^3 algorithm uses large-scale cancer genomics datasets which are pre-processed to yield parameters governing novel constrained correlation clustering techniques. The optimization criteria used in clustering include patterns of mutual exclusivity of mutations, patient sample coverage, and network driver concentration. C^3 improves over previous methods that use randomized with fixed cluster sizes approaches without the guarantee that multiple runs of the methods on the same data will produce compatible results for any cluster size. C^3 was able to identify several potential driver pathways when applied to BRCA and GBM data that could guide new drug targets and new drug mechanisms.

Finally, we presented a novel method that ties in the insights we obtained from molecular signature and pathway information to prescribe treatments to cancer patients. Scattershot's comprehensive genomic, pathway, and clinical data to predict the drug response of a patient and make a ranked list of drug recommendations for any given patient *in silico*. We applied Scattershot to 647 breast cancer patients and a Pan-Cancer study of 1508 patients from the publicly available TCGA database. Scattershot's integrated approach has outperformed previous methods in predicting drug response in actual patient samples, and its novel recommender has consistently ranked actual prescribed drugs highly in a large majority of patients. We believe that Scattershot provides a framework which can be used to personalized treatment approaches in cancer.

5.2 Future Directions

The conclusions from the chapters of the thesis pave the way for several potential future directions of our projects. In **Chapter 2**, one of the areas of interest lies in the construction of the gene network. We are also limited by the biased by the curated pathway used to evaluate the networks. The gene network is not complete, with many interactions incomplete. Additionally, the interactions between the genes themselves may change in a cancer genome. One future direction to model the interaction of the gene network is to utilize a dynamic network where the nodes and edges are specific to an individual cancer patient. Additionally, we would like to access the effect of the driver genes over time comparing pre-treatment and post-treatment samples) to determine if there are any changes in the driver function after treatment. We are also interested in looking at drivers that participate in the metastatic process. Metastases is the leading cause of cancer related deaths, and oftentimes a small percentage clone in the primary causes seeding of distant metastases. Thus, drivers identified in the primary may not be the main causes of metastasis or the genes that need to be targeted to halt metastatic progression. Future studies on large cohorts of matched primaries and metastases will soon answer these questions. *In vitro* and *in vivo* studies can also be used to confirm our findings.

Several directions of future work are also present in **Chapter 3**. From a technical standpoint, several improvements can be made on determining the weights of the algorithm. Weights in C^3 were determined heuristically, using a brute force method to test C^3 on multiple weighting parameters. This manner of selecting weights is less efficient and time consuming if optimal weight parameters change in different cancer studies if the optimal weights for C^3 vary from cancer to cancer. In the future, we plan on applying C^3 to other data sets and other cancers to determine to determine whether the optimal weights for expression, coverage and mutual exclusivity vary within different biological context. Also, *In vitro* and *in vivo* studies can also be used to confirm our findings.

Future work can also address limitations in Scattershot in **Chapter 4**. To a large amount of missing information, Scattershot's drug response may be incomplete. This is especially true for many of the cancers outside of the 14 chosen in the Scattershot parameter selection. When more information from TCGA is made available, we will be able to update the results of Scattershot. Additionally, more scrutiny and curation can be made in the DGIdb derived drug target information. A manual curation of drug targets with like-mechanism drugs serving as a baseline may be useful in supplementing some potential incomplete information a generalized, non-curated method like DGIdb provides from drug targets. Scattershot should be run in the future to create a more complete analysis of Pan-Cancer response once the TCGA data is updated. For this part of the analysis, *in vitro* cell line analysis may be used to confirm the findings of Scattershot.

References

[1] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2016.," *CA. Cancer J. Clin.*, vol. 66, no. 1, pp. 7–30.

[2] M. S. Lawrence, P. Stojanov, P. Polak, G. V Kryukov, K. Cibulskis, A. Sivachenko, S. L. Carter, C. Stewart, C. H. Mermel, S. A. Roberts, A. Kiezun, P. S. Hammerman, A. McKenna, Y. Drier, L. Zou, A. H. Ramos, T. J. Pugh, N. Stransky, E. Helman, J. Kim, C. Sougnez, L. Ambrogio, E. Nickerson, E. Shefler, M. L. Cortés, D. Auclair, G. Saksena, D. Voet, M. Noble, D. DiCara, P. Lin, L. Lichtenstein, D. I. Heiman, T. Fennell, M. Imielinski, B. Hernandez, E. Hodis, S. Baca, A. M. Dulak, J. Lohr, D.-A. Landau, C. J. Wu, J. Melendez-Zajgla, A. Hidalgo-Miranda, A. Koren, S. A. McCarroll, J. Mora, R. S. Lee, B. Crompton, R. Onofrio, M. Parkin, W. Winckler, K. Ardlie, S. B. Gabriel, C. W. M. Roberts, J. A. Biegel, K. Stegmaier, A. J. Bass, L. A. Garraway, M. Meyerson, T. R. Golub, D. A. Gordenin, S. Sunyaev, E. S. Lander, and G. Getz, "Mutational heterogeneity in cancer and the search for new cancer-associated genes.," *Nature*, vol. 499, no. 7457, pp. 214–8, Jul. 2013.

[3] A. Marusyk and K. Polyak, "Tumor heterogeneity: causes and consequences.," *Biochim. Biophys. Acta*, vol. 1805, no. 1, pp. 105–17, Jan. 2010.

[4] R. A. Burrell, N. McGranahan, J. Bartek, and C. Swanton, "The causes and consequences of genetic heterogeneity in cancer evolution.," *Nature*, vol. 501, no. 7467, pp. 338–45, Sep. 2013.

[5] S. Aparicio and C. Caldas, "The Implications of Clonal Genome Evolution for Cancer Medicine," *N. Engl. J. Med.*, vol. 368, no. 9, pp. 842–851, Feb. 2013.

[6] S. Calza, P. Hall, G. Auer, J. Bjöhle, S. Klaar, U. Kronenwett, E. T. Liu, L. Miller, A. Ploner, J. Smeds, J. Bergh, and Y. Pawitan, "Intrinsic molecular signature of breast cancer in a population-based cohort of 412 patients.," *Breast Cancer Res.*, vol. 8, no. 4, p. R34, 2006.

[7] I. S. Chan and G. S. Ginsburg, "Personalized Medicine: Progress and Promise," *Annu. Rev. Genomics Hum. Genet.*, vol. 12, no. 1, pp. 217–244, Sep. 2011.

[8] L. A. Garraway, J. Verweij, and K. V. Ballman, "Precision Oncology: An Overview," *J. Clin. Oncol.*, vol. 31, no. 15, pp. 1803–1805, May 2013.

[9] E. M. Goldblatt and W.-H. Lee, "From bench to bedside: the growing use of translational research in cancer medicine.," *Am. J. Transl. Res.*, vol. 2, no. 1, pp. 1–18, 2010.

[10] D. Hanahan and R. A. Weinberg, "Hallmarks of cancer: The next generation," *Cell*, vol. 144, no. 5. pp. 646–674, 2011.

[11] B. Vogelstein, N. Papadopoulos, V. E. Velculescu, S. Zhou, L. A. Diaz, K. W. Kinzler, and K. W. Kinzler, "Cancer genome landscapes.," *Science*, vol. 339, no. 6127, pp. 1546–58, Mar. 2013.

[12] P. Eroles, A. Bosch, J. Alejandro Pérez-Fidalgo, and A. Lluch, "Molecular biology in breast cancer: Intrinsic subtypes and signaling pathways," *Cancer Treat. Rev.*, vol. 38, no. 6, pp. 698–707, 2012.

[13] L. Wang, H. L. McLeod, and R. M. Weinshilboum, "Genomics and Drug Response," *http://dx.doi.org/10.1056/NEJMra1010600*, 2011.

[14] G. H. Fernald, E. Capriotti, R. Daneshjou, K. J. Karczewski, and R. B. Altman, "Bioinformatics challenges for personalized medicine.," *Bioinformatics*, vol. 27, no. 13, pp. 1741–8, Jul. 2011.

[15] C. Greenman, P. Stephens, R. Smith, G. L. Dalgliesh, C. Hunter, G. Bignell, H. Davies, J.

Teague, A. Butler, C. Stevens, S. Edkins, S. O'Meara, I. Vastrik, E. E. Schmidt, T. Avis, S. Barthorpe, G. Bhamra, G. Buck, B. Choudhury, J. Clements, J. Cole, E. Dicks, S. Forbes, K. Gray, K. Halliday, R. Harrison, K. Hills, J. Hinton, A. Jenkinson, D. Jones, A. Menzies, T. Mironenko, J. Perry, K. Raine, D. Richardson, R. Shepherd, A. Small, C. Tofts, J. Varian, T. Webb, S. West, S. Widaa, A. Yates, D. P. Cahill, D. N. Louis, P. Goldstraw, A. G. Nicholson, F. Brasseur, L. Looijenga, B. L. Weber, Y.-E. Chiew, A. DeFazio, M. F. Greaves, A. R. Green, P. Campbell, E. Birney, D. F. Easton, G. Chenevix-Trench, M.-H. Tan, S. K. Khoo, B. T. Teh, S. T. Yuen, S. Y. Leung, R. Wooster, P. A. Futreal, and M. R. Stratton, "Patterns of somatic mutation in human cancer genomes.," *Nature*, vol. 446, no. 7132, pp. 153–8, Mar. 2007.

[16] L. Ding, G. Getz, D. A. Wheeler, E. R. Mardis, M. D. McLellan, K. Cibulskis, C. Sougnez, H. Greulich, D. M. Muzny, M. B. Morgan, L. Fulton, R. S. Fulton, Q. Zhang, M. C. Wendl, M. S. Lawrence, D. E. Larson, K. Chen, D. J. Dooling, A. Sabo, A. C. Hawes, H. Shen, S. N. Jhangiani, L. R. Lewis, O. Hall, Y. Zhu, T. Mathew, Y. Ren, J. Yao, S. E. Scherer, K. Clerc, G. A. Metcalf, B. Ng, A. Milosavljevic, M. L. Gonzalez-Garay, J. R. Osborne, R. Meyer, X. Shi, Y. Tang, D. C. Koboldt, L. Lin, R. Abbott, T. L. Miner, C. Pohl, G. Fewell, C. Haipek, H. Schmidt, B. H. Dunford-Shore, A. Kraja, S. D. Crosby, C. S. Sawyer, T. Vickery, S. Sander, J. Robinson, W. Winckler, J. Baldwin, L. R. Chirieac, A. Dutt, T. Fennell, M. Hanna, B. E. Johnson, R. C. Onofrio, R. K. Thomas, G. Tonon, B. A. Weir, X. Zhao, L. Ziaugra, M. C. Zody, T. Giordano, M. B. Orringer, J. A. Roth, M. R. Spitz, I. I. Wistuba, B. Ozenberger, P. J. Good, A. C. Chang, D. G. Beer, M. A. Watson, M. Ladanyi, S. Broderick, A. Yoshizawa, W. D. Travis, W. Pao, M. A. Province, G. M. Weinstock, H. E. Varmus, S. B. Gabriel, E. S. Lander, R. A. Gibbs, M. Meyerson, and R. K. Wilson, "Somatic mutations affect key pathways in lung adenocarcinoma.," *Nature*, vol. 455, no. 7216, pp. 1069–75, Oct. 2008.

[17] S. Jones, X. Zhang, D. W. Parsons, J. C.-H. Lin, R. J. Leary, P. Angenendt, P. Mankoo, H. Carter, H. Kamiyama, A. Jimeno, S.-M. Hong, B. Fu, M.-T. Lin, E. S. Calhoun, M. Kamiyama, K. Walter, T. Nikolskaya, Y. Nikolsky, J. Hartigan, D. R. Smith, M. Hidalgo, S. D. Leach, A. P. Klein, E. M. Jaffee, M. Goggins, A. Maitra, C. Iacobuzio-Donahue, J. R. Eshleman, S. E. Kern, R. H. Hruban, R. Karchin, N. Papadopoulos, G. Parmigiani, B. Vogelstein, V. E. Velculescu, and K. W. Kinzler, "Core signaling pathways in human pancreatic cancers revealed by global genomic analyses.," *Science*, vol. 321, no. 5897, pp. 1801–6, Sep. 2008.

[18] S. Banerji, K. Cibulskis, C. Rangel-Escareno, K. K. Brown, S. L. Carter, A. M. Frederick, M. S. Lawrence, A. Y. Sivachenko, C. Sougnez, L. Zou, M. L. Cortes, J. C. Fernandez-Lopez, S. Peng, K. G. Ardlie, D. Auclair, V. Bautista-Piña, F. Duke, J. Francis, J. Jung, A. Maffuz-Aziz, R. C. Onofrio, M. Parkin, N. H. Pho, V. Quintanar-Jurado, A. H. Ramos, R. Rebollar-Vega, S. Rodriguez-Cuevas, S. L. Romero-Cordoba, S. E. Schumacher, N. Stransky, K. M. Thompson, L. Uribe-Figueroa, J. Baselga, R. Beroukhim, K. Polyak, D. C. Sgroi, A. L. Richardson, G. Jimenez-Sanchez, E. S. Lander, S. B. Gabriel, L. A. Garraway, T. R. Golub, J. Melendez-Zajgla, A. Toker, G. Getz, A. Hidalgo-Miranda, and M. Meyerson, "Sequence analysis of mutations and translocations across breast cancer subtypes.," *Nature*, vol. 486, no. 7403, pp. 405–9, Jun. 2012.

[19] N. D. Dees, Q. Zhang, C. Kandoth, M. C. Wendl, W. Schierding, D. C. Koboldt, T. B. Mooney, M. B. Callaway, D. Dooling, E. R. Mardis, R. K. Wilson, and L. Ding, "MuSiC: identifying mutational significance in cancer genomes.," *Genome Res.*, vol. 22, no. 8, pp.

1589–98, Aug. 2012.

[20] H. Carter, J. Samayoa, R. H. Hruban, and R. Karchin, “Prioritization of driver mutations in pancreatic cancer using cancer-specific high-throughput annotation of somatic mutations (CHASM).,” *Cancer Biol. Ther.*, vol. 10, no. 6, pp. 582–7, Sep. 2010.

[21] U. D. Akavia, O. Litvin, J. Kim, F. Sanchez-Garcia, D. Kotliar, H. C. Causton, P. Pochanard, E. Mozes, L. A. Garraway, D. Pe’er“An integrated approach to uncover drivers of cancer.,” *Cell*, vol. 143, no. 6, pp. 1005–17, Dec. 2010.

[22] S. Ng, E. A. Collisson, A. Sokolov, T. Goldstein, A. Gonzalez-Perez, N. Lopez-Bigas, C. Benz, D. Haussler, and J. M. Stuart, “PARADIGM-SHIFT predicts the function of mutations in multiple cancers using pathway impact analysis.,” *Bioinformatics*, vol. 28, no. 18, pp. i640–i646, Sep. 2012.

[23] A. Bashashati, G. Haffari, J. Ding, G. Ha, K. Lui, J. Rosner, D. G. Huntsman, C. Caldas, S. A. Aparicio, and S. P. Shah, “DriverNet: uncovering the impact of somatic driver mutations on transcriptional networks in cancer.,” *Genome Biol.*, vol. 13, no. 12, p. R124, Jan. 2012.

[24] Y. Chen, J. Hao, W. Jiang, T. He, X. Zhang, T. Jiang, R. Jiang, “Identifying potential cancer driver genes by genomic data integration,” *Sci. Rep.*, vol. 3, pp. 652–4, Dec. 2013.

[25] Y. Liu, F. Tian, Z. Hu, C. DeLisi “Evaluation and integration of cancer gene classifiers: identification and ranking of plausible drivers,” *Sci. Rep.*, vol. 5, p. 10204, May 2015.

[26] K. D. Korthauer and C. Kendziorski, “MADGiC: a model-based approach for identifying driver genes in cancer.” Bioinformatics 2015 May 15;31(10):1526-35.

[27] I.-F. Chung, C.-Y. Chen, S.-C. Su, C.-Y. Li, K.-J. Wu, H.-W. Wang, and W.-C. Cheng, “DriverDBv2: a database for human cancer driver gene research.,” *Nucleic Acids Res.*, vol. 44, no. D1, pp. D975-9, Jan. 2016.

[28] B. Vogelstein and K. W. Kinzler, “Cancer genes and the pathways they control,” *Nat. Med.*, vol. 10, no. 8, pp. 789–799, Aug. 2004.

[29] G. Ciriello, E. Cerami, C. Sander, and N. Schultz, “Mutual exclusivity analysis identifies oncogenic network modules.,” *Genome Res.*, vol. 22, no. 2, pp. 398–406, Feb. 2012.

[30] J. Zhang, S. Zhang, Y. Wang, and X.-S. Zhang, “Identification of mutated core cancer modules by integrating somatic mutation, copy number variation, and gene expression data.,” *BMC Syst. Biol.*, vol. 7 Suppl 2, p. S4, 2013.

[31] F. Vandin, E. Upfal, and B. J. Raphael, “De novo discovery of mutated driver pathways in cancer.,” *Genome Res.*, vol. 22, no. 2, pp. 375–85, Feb. 2012.

[32] J. Zhao, S. Zhang, L.-Y. Wu, and X.-S. Zhang, “Efficient methods for identifying mutated driver pathways in cancer.” *Bioinformatics (2012) 28 (22): 2940-2947.*

[33] M. D. M. Leiserson, D. Blokh, R. Sharan, B. J. Raphael “Simultaneous Identification of Multiple Driver Pathways in Cancer,” *PLoS Comput. Biol.*, vol. 9, no. 5, p. e1003054, May 2013.

[34] J. Zhang, L.-Y. Wu, X.-S. Zhang, S. Zhang “Discovery of co-occurring driver pathways in cancer,” *BMC Bioinformatics*, vol. 15, no. 1, p. 271, 2014.

[35] M. D. M. Leiserson, F. Vandin, H.-T. Wu, J. R. Dobson, J. V Eldridge, J. L. Thomas, A. Papoutsaki, Y. Kim, B. Niu, M. McLellan, M. S. Lawrence, A. Gonzalez-Perez, D. Tamborero, Y. Cheng, G. A. Ryslik, N. Lopez-Bigas, G. Getz, L. Ding, and B. J. Raphael, “Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes.,” *Nat. Genet.*, vol. 47, no. 2, pp. 106–14, Feb. 2015.

[36] J. S. Parker, M. Mullins, M. C. U. Cheang, S. Leung, D. Voduc, T. Vickery, S. Davies, C.

Fauron, X. He, Z. Hu, J. F. Quackenbush, I. J. Stijleman, J. Palazzo, J. S. Marron, A. B. Nobel, E. Mardis, T. O. Nielsen, M. J. Ellis, C. M. Perou, and P. S. Bernard, "Supervised risk predictor of breast cancer based on intrinsic subtypes.," *J. Clin. Oncol.*, vol. 27, no. 8, pp. 1160–7, Mar. 2009.

[37] K. D. Voduc, M. C. U. Cheang, S. Tyldesley, K. Gelmon, T. O. Nielsen, and H. Kennecke, "Breast cancer subtypes and the risk of local and regional relapse.," *J. Clin. Oncol.*, vol. 28, no. 10, pp. 1684–91, Apr. 2010.

[38] A. Prat and C. M. Perou, "Deconstructing the molecular portraits of breast cancer.," *Mol. Oncol.*, vol. 5, no. 1, pp. 5–23, Feb. 2011.

[39] A. A. Onitilo, J. M. Engel, R. T. Greenlee, and B. N. Mukesh, "Breast cancer subtypes based on ER/PR and Her2 expression: comparison of clinicopathologic features and survival.," *Clin. Med. Res.*, vol. 7, no. 1–2, pp. 4–13, Jun. 2009.

[40] C. M. Perou, "Molecular stratification of triple-negative breast cancers.," *Oncologist*, vol. 16 Suppl 1, pp. 61–70, 2011.

[41] C. A. Hudis and L. Gianni, "Triple-Negative Breast Cancer: An Unmet Medical Need," *Oncologist*, vol. 16, no. Supplement 1, pp. 1–11, Jan. 2011.

[42] R. R. L. Bastien, Á. Rodríguez-Lescure, M. T. W. Ebbert, A. Prat, B. Munárriz, L. Rowe, P. Miller, M. Ruiz-Borrego, D. Anderson, B. Lyons, I. Álvarez, T. Dowell, D. Wall, M. Á. Seguí, L. Barley, K. M. Boucher, E. Alba, L. Pappas, C. A. Davis, I. Aranda, C. Fauron, I. J. Stijleman, J. Palacios, A. Antón, E. Carrasco, R. Caballero, M. J. Ellis, T. O. Nielsen, C. M. Perou, M. Astill, P. S. Bernard, and M. Martín, "PAM50 breast cancer subtyping by RT-qPCR and concordance with standard clinical molecular markers.," *BMC Med. Genomics*, vol. 5, no. 1, p. 44, Jan. 2012.

[43] C. Curtis, S. P. Shah, S.-F. Chin, G. Turashvili, O. M. Rueda, M. J. Dunning, D. Speed, A. G. Lynch, S. Samarajiwa, Y. Yuan, S. Gräf, G. Ha, G. Haffari, A. Bashashati, R. Russell, S. McKinney, A. Langerød, A. Green, E. Provenzano, G. Wishart, S. Pinder, P. Watson, F. Markowetz, L. Murphy, I. Ellis, A. Purushotham, A.-L. Børresen-Dale, J. D. Brenton, S. Tavaré, C. Caldas, and S. Aparicio, "The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups.," *Nature*, vol. 486, no. 7403, pp. 346–52, Jun. 2012.

[44] "Comprehensive molecular portraits of human breast tumours.," *Nature*, vol. 490, no. 7418, pp. 61–70, Oct. 2012.

[45] M. L. Gatza, G. O. Silva, J. S. Parker, C. Fan, and C. M. Perou, "An integrated genomics approach identifies drivers of proliferation in luminal-subtype human breast cancer," *Nat. Genet.*, vol. 46, no. 10, pp. 1051–1059, Aug. 2014.

[46] J. U. Adams, "Genetics: Big hopes for big data," *Nature*, vol. 527, no. 7578, pp. S108–S109, Nov. 2015.

[47] M. A. Rubin, "Health: Make precision medicine work for cancer care.," *Nature*, vol. 520, no. 7547, pp. 290–1, Apr. 2015.

[48] M. Ladanyi and W. Pao, "Lung adenocarcinoma: guiding EGFR-targeted therapy and beyond," *Mod. Pathol.*, vol. 21, pp. S16–S22, May 2008.

[49] J. A. Incorvati, S. Shah, Y. Mu, and J. Lu, "Targeted therapy for HER2 positive breast cancer.," *J. Hematol. Oncol.*, vol. 6, p. 38, Jun. 2013.

[50] M. M. Gottesman, "Mechanisms of cancer drug resistance.," *Annu. Rev. Med.*, vol. 53, pp. 615–27, 2002.

[51] M. Iskar, G. Zeller, X.-M. Zhao, V. Van Noort, and P. Bork, "Drug discovery in the age of

systems biology: the rise of computational approaches for data integration," *Curr. Opin. Biotechnol.*, vol. 23, pp. 1–8, 2011.

[52] H. Sun, X.-B. Li, Y. Meng, L. Fan, M. Li, and J. Fang, "TRAF6 upregulates expression of HIF-1α and promotes tumor angiogenesis.," *Cancer Res.*, vol. 73, no. 15, pp. 4950–9, Aug. 2013.

[53] W. Yang, J. Soares, P. Greninger, E. J. Edelman, H. Lightfoot, S. Forbes, N. Bindal, D. Beare, J. A. Smith, I. R. Thompson, S. Ramaswamy, P. A. Futreal, D. A. Haber, M. R. Stratton, C. Benes, U. McDermott, and M. J. Garnett, "Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells.," *Nucleic Acids Res.*, vol. 41, no. Database issue, pp. D955-61, Jan. 2013.

[54] J. Barretina, G. Caponigro, N. Stransky, K. Venkatesan, A. A. Margolin, S. Kim, C. J. Wilson, J. Lehár, G. V Kryukov, D. Sonkin, A. Reddy, M. Liu, L. Murray, M. F. Berger, J. E. Monahan, P. Morais, J. Meltzer, A. Korejwa, J. Jané-Valbuena, F. A. Mapa, J. Thibault, E. Bric-Furlong, P. Raman, A. Shipway, I. H. Engels, J. Cheng, G. K. Yu, J. Yu, P. Aspesi, M. de Silva, K. Jagtap, M. D. Jones, L. Wang, C. Hatton, E. Palescandolo, S. Gupta, S. Mahan, C. Sougnez, R. C. Onofrio, T. Liefeld, L. MacConaill, W. Winckler, M. Reich, N. Li, J. P. Mesirov, S. B. Gabriel, G. Getz, K. Ardlie, V. Chan, V. E. Myer, B. L. Weber, J. Porter, M. Warmuth, P. Finan, J. L. Harris, M. Meyerson, T. R. Golub, M. P. Morrissey, W. R. Sellers, R. Schlegel, and L. A. Garraway, "The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity.," *Nature*, vol. 483, no. 7391, pp. 603–7, Mar. 2012.

[55] Z. Dong, N. Zhang, C. Li, H. Wang, Y. Fang, J. Wang, X. Zheng, "Anticancer drug sensitivity prediction in cell lines from baseline gene expression through recursive feature selection," *BMC Cancer*, vol. 15, no. 1, p. 489, Dec. 2015.

[56] S. Gupta, K. Chaudhary, R. Kumar, A. Gautam, J. S. Nanda, S. K. Dhanda, S. K. Brahmachari, G. P. S. Raghava, "Prioritization of anticancer drugs against a cancer using genomic features of cancer cells: A step towards personalized medicine," *Sci. Rep.*, vol. 6, p. 23857, Mar. 2016.

[57] J. Jack, D. Rotroff, and A. Motsinger-Reif, "Lymphoblastoid cell lines models of drug response: successes and lessons from this pharmacogenomic model.," *Curr. Mol. Med.*, vol. 14, no. 7, pp. 833–40, 2014.

[58] F. Iorio, T. A. Knijnenburg, D. J. Vis, G. R. Bignell, M. P. Menden, M. Schubert, N. Aben, E. Gonçalves, S. Barthorpe, H. Lightfoot, T. Cokelaer, P. Greninger, E. van Dyk, H. Chang, H. de Silva, H. Heyn, X. Deng, R. K. Egan, Q. Liu, T. Mironenko, X. Mitropoulos, L. Richardson, J. Wang, T. Zhang, S. Moran, S. Sayols, M. Soleimani, D. Tamborero, N. Lopez-Bigas, P. Ross-Macdonald, M. Esteller, N. S. Gray, D. A. Haber, M. R. Stratton, C. H. Benes, L. F. A. Wessels, J. Saez-Rodriguez, U. McDermott, M. J. Garnett, "A Landscape of Pharmacogenomic Interactions in Cancer," *Cell*, vol. 166, no. 3, pp. 740–754, Jul. 2016.

[59] P. Geeleher, E. R. Gamazon, C. Seoighe, N. J. Cox, and R. S. Huang, "Consistency in large pharmacogenomic studies," *Nature*, vol. 540, no. 7631, pp. E1–E2, Nov. 2016.

[60] T. Minamoto, M. Mai, and Z. Ronai, "Environmental factors as regulators and effectors of multistep carcinogenesis.," *Carcinogenesis*, vol. 20, no. 4, pp. 519–27, Apr. 1999.

[61] P. Geeleher, N. J. Cox, and R. Huang, "Clinical drug response can be predicted using baseline gene expression levels and in vitro drug sensitivity in cell lines," *Genome Biol.*, vol. 15, no. 3, p. R47, 2014.

[62] S. B. Amin, W.-K. Yip, S. Minvielle, A. Broyl, Y. Li, B. Hanlon, D. Swanson, P. K. Shah, P. Moreau, B. van der Holt, M. van Duin, F. Magrangeas, P. Pieter Sonneveld, K. C. Anderson, C. Li, H. Avet-Loiseau, and N. C. Munshi, “Gene expression profile alone is inadequate in predicting complete response in multiple myeloma,” *Leukemia*, vol. 28, no. 11, pp. 2229–2234, Nov. 2014.

[63] C. Rubio-Perez, D. Tamborero, M. P. Schroeder, A. A. Antolín, J. Deu-Pons, C. Perez-Llamas, J. Mestres, A. Gonzalez-Perez, N. Lopez-Bigas “In silico prescription of anticancer drugs to cohorts of 28 tumor types reveals targeting opportunities.,” *Cancer Cell*, vol. 27, no. 3, pp. 382–96, Mar. 2015.

[64] S. Nabavi, “Identifying candidate drivers of drug response in heterogeneous cancer by mining high throughput genomics data,” *BMC Genomics*, vol. 17, no. 1, p. 638, Dec. 2016.

[65] R. Louhimo, M. Laakso, D. Belitskin, J. Klefström, R. Lehtonen, and S. Hautaniemi, “Data integration to prioritize drugs using genomics and curated data.,” *BioData Min.*, vol. 9, p. 21, 2016.

[66] Z. Ding, S. Zu, and J. Gu, “Evaluating the molecule-based prediction of clinical drug responses in cancer,” *Bioinformatics*, p. btw344, Jun. 2016.

[67] J. P. Hou and J. Ma, “DawnRank: discovering personalized driver genes in cancer,” *Genome Med.*, vol. 6, no. 7, p. 56, 2014.

[68] J. P. Hou, A. Emad, G. J. Puleo, J. Ma, and O. Milenkovic, “A new correlation clustering method for cancer mutation analysis.,” *Bioinformatics*, p. btw546, Aug. 2016.

[69] G. Tsoumakas and I. Katakis, “Multi-Label Classification: An Overview.” *International Journal of Data Warehousing and Mining*, David Taniar (Ed.), Idea Group Publishing, 3(3), pp. 1-13, 2007.

70] M. R. Stratton, P. J. Campbell, and P. A. Futreal, “The cancer genome.,” *Nature*, vol. 458, no. 7239, pp. 719–24, Apr. 2009.

[71] D. Pe’er, N. Hacohen “Principles and strategies for developing network models in cancer.,” *Cell*, vol. 144, no. 6, pp. 864–73, Mar. 2011.

[72] L. Ding, M. C. Wendl, D. C. Koboldt, and E. R. Mardis, “Analysis of next-generation genomic data in cancer: accomplishments and challenges.,” *Hum. Mol. Genet.*, vol. 19, no. R2, pp. R188-96, Oct. 2010.

[73] J. Reimand and G. D. Bader, “Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers.,” *Mol. Syst. Biol.*, vol. 9, p. 637, 2013.

[74] C. M. Perou and A.-L. Børresen-Dale, “Systems biology and genomics of breast cancer.,” *Cold Spring Harb. Perspect. Biol.*, vol. 3, no. 2, p. a003293-, Feb. 2011.

[75] O. Yersal and S. Barutca, “Biological subtypes of breast cancer: Prognostic and therapeutic implications.,” *World J. Clin. Oncol.*, vol. 5, no. 3, pp. 412–24, Aug. 2014.

[76] B. D. Lehmann, J. A. Bauer, X. Chen, M. E. Sanders, A. B. Chakravarthy, Y. Shyr, J. A. Pietenpol, “Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies.,” *J. Clin. Invest.*, vol. 121, no. 7, pp. 2750–67, Jul. 2011.

[77] Cancer Genome Atlas Research Network, “Comprehensive genomic characterization defines human glioblastoma genes and core pathways.,” *Nature*, vol. 455, no. 7216, pp. 1061–8, Oct. 2008.

[78] D. Bell, A. Berchuck, M. Birrer, J. Chien, D. W. Cramer, F. Dao, R. Dhir, P. DiSaia, H. Gabra, P. Glenn, A. K. Godwin, J. Gross, L. Hartmann, M. Huang, D. G. Huntsman, M. Iacocca, M. Imielinski, S. Kalloger, B. Y. Karlan, D. A. Levine, G. B. Mills, C. Morrison, D. Mutch, N. Olvera, S. Orsulic, K. Park, N. Petrelli, B. Rabeno, J. S. Rader, B. I. Sikic, K. Smith-McCune, A. K. Sood, D. Bowtell, R. Penny, J. R. Testa, K. Chang, H. H. Dinh, J. A. Drummond, G. Fowler, P. Gunaratne, A. C. Hawes, C. L. Kovar, L. R. Lewis, M. B. Morgan, I. F. Newsham, J. Santibanez, J. G. Reid, L. R. Trevino, Y.-Q. Wu, M. Wang, D. M. Muzny, D. A. Wheeler, R. A. Gibbs, G. Getz, M. S. Lawrence, K. Cibulskis, A. Y. Sivachenko, C. Sougnez, D. Voet, J. Wilkinson, T. Bloom, K. Ardlie, T. Fennell, J. Baldwin, S. Gabriel, E. S. Lander, L. Ding, R. S. Fulton, D. C. Koboldt, M. D. McLellan, T. Wylie, J. Walker, M. O'Laughlin, D. J. Dooling, L. Fulton, R. Abbott, N. D. Dees, Q. Zhang, C. Kandoth, M. Wendl, W. Schierding, D. Shen, C. C. Harris, H. Schmidt, J. Kalicki, K. D. Delehaunty, C. C. Fronick, R. Demeter, L. Cook, J. W. Wallis, L. Lin, V. J. Magrini, J. S. Hodges, J. M. Eldred, S. M. Smith, C. S. Pohl, F. Vandin, B. J. Raphael, G. M. Weinstock, E. R. Mardis, R. K. Wilson, M. Meyerson, W. Winckler, G. Getz, R. G. W. Verhaak, S. L. Carter, C. H. Mermel, G. Saksena, H. Nguyen, R. C. Onofrio, M. S. Lawrence, D. Hubbard, S. Gupta, A. Crenshaw, A. H. Ramos, K. Ardlie, L. Chin, A. Protopopov, J. Zhang, T. M. Kim, I. Perna, Y. Xiao, H. Zhang, G. Ren, N. Sathiamoorthy, R. W. Park, E. Lee, P. J. Park, R. Kucherlapati, D. M. Absher, L. Waite, G. Sherlock, J. D. Brooks, J. Z. Li, J. Xu, R. M. Myers, P. W. Laird, L. Cope, J. G. Herman, H. Shen, D. J. Weisenberger, H. Noushmehr, F. Pan, T. Triche Jr, B. P. Berman, D. J. Van Den Berg, J. Buckley, S. B. Baylin, P. T. Spellman, E. Purdom, P. Neuvial, H. Bengtsson, L. R. Jakkula, S. Durinck, J. Han, S. Dorton, H. Marr, Y. G. Choi, V. Wang, N. J. Wang, J. Ngai, J. G. Conboy, B. Parvin, H. S. Feiler, T. P. Speed, J. W. Gray, D. A. Levine, N. D. Socci, Y. Liang, B. S. Taylor, N. Schultz, L. Borsu, A. E. Lash, C. Brennan, A. Viale, C. Sander, M. Ladanyi, K. A. Hoadley, S. Meng, Y. Du, Y. Shi, L. Li, Y. J. Turman, D. Zang, E. B. Helms, S. Balu, X. Zhou, J. Wu, M. D. Topal, D. N. Hayes, C. M. Perou, G. Getz, D. Voet, G. Saksena, J. Zhang, H. Zhang, C. J. Wu, S. Shukla, K. Cibulskis, M. S. Lawrence, A. Sivachenko, R. Jing, R. W. Park, Y. Liu, P. J. Park, M. Noble, L. Chin, H. Carter, D. Kim, R. Karchin, P. T. Spellman, E. Purdom, P. Neuvial, H. Bengtsson, S. Durinck, J. Han, J. E. Korkola, L. M. Heiser, R. J. Cho, Z. Hu, B. Parvin, T. P. Speed, J. W. Gray, N. Schultz, E. Cerami, B. S. Taylor, A. Olshen, B. Reva, Y. Antipin, R. Shen, P. Mankoo, R. Sheridan, G. Ciriello, W. K. Chang, J. A. Bernanke, L. Borsu, D. A. Levine, M. Ladanyi, C. Sander, D. Haussler, C. C. Benz, J. M. Stuart, S. C. Benz, J. Z. Sanborn, C. J. Vaske, J. Zhu, C. Szeto, G. K. Scott, C. Yau, K. A. Hoadley, Y. Du, S. Balu, D. N. Hayes, C. M. Perou, M. D. Wilkerson, N. Zhang, R. Akbani, K. A. Baggerly, W. K. Yung, G. B. Mills, J. N. Weinstein, R. Penny, T. Shelton, D. Grimm, M. Hatfield, S. Morris, P. Yena, P. Rhodes, M. Sherman, J. Paulauskis, S. Millis, A. Kahn, J. M. Greene, R. Sfeir, M. A. Jensen, J. Chen, J. Whitmore, S. Alonso, J. Jordan, A. Chu, J. Zhang, A. Barker, C. Compton, G. Eley, M. Ferguson, P. Fielding, D. S. Gerhard, R. Myles, C. Schaefer, K. R. Mills Shaw, J. Vaught, J. B. Vockley, P. J. Good, M. S. Guyer, B. Ozenberger, J. Peterson, and E. Thomson, "Integrated genomic analyses of ovarian carcinoma," *Nature*, vol. 474, no. 7353, pp. 609–615, Jun. 2011.

[79] M. D. Wilkerson and D. N. Hayes, "ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking.," *Bioinformatics*, vol. 26, no. 12, pp. 1572–3, Jun. 2010.

[80] A. R. Dabney, "ClaNC: point-and-click software for classifying microarrays to nearest centroids.," *Bioinformatics*, vol. 22, no. 1, pp. 122–3, Jan. 2006.

[81] V. G. Tusher, R. Tibshirani, and G. Chu, "Significance analysis of microarrays applied to the ionizing radiation response.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 98, no. 9, pp. 5116–21, Apr. 2001.

[82] R. Bhargava, S. Beriwal, D. J. Dabbs, U. Ozbek, A. Soran, R. R. Johnson, A. M. Brufsky, B. C. Lembersky, and G. M. Ahrendt, "Immunohistochemical surrogate markers of breast cancer molecular classes predicts response to neoadjuvant chemotherapy: a single institutional experience with 359 cases.," *Cancer*, vol. 116, no. 6, pp. 1431–9, Mar. 2010.

[83] H. Kouros-Mehr, E. M. Slorach, M. D. Sternlicht, and Z. Werb, "GATA-3 maintains the differentiation of the luminal cell fate in the mammary gland.," *Cell*, vol. 127, no. 5, pp. 1041–55, Dec. 2006.

[84] Y. Shrestha, E. J. Schafer, J. S. Boehm, S. R. Thomas, F. He, J. Du, S. Wang, J. Barretina, B. A. Weir, J. J. Zhao, K. Polyak, T. R. Golub, R. Beroukhim, and W. C. Hahn, "PAK1 is a breast cancer oncogene that coordinately activates MAPK and MET signaling.," *Oncogene*, vol. 31, no. 29, pp. 3397–408, Jul. 2012.

[85] L. C. Chen, C. Dollbaum, and H. S. Smith, "Loss of heterozygosity on chromosome 1q in human breast cancer.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 86, no. 18, pp. 7204–7, 1989.

[86] W. Yu, Y. Kanaan, Y. K. Bae, Y.-K. Baed, and E. Gabrielson, "Chromosomal changes in aggressive breast cancers with basal-like features.," *Cancer Genet. Cytogenet.*, vol. 193, no. 1, pp. 29–37, Aug. 2009.

[87] J. Faridi, L. Wang, G. Endemann, and R. A. Roth, "Expression of Constitutively Active Akt-3 in MCF-7 Breast Cancer Cells Reverses the Estrogen and Tamoxifen Responsivity of these Cells in Vivo," *Clin. Cancer Res.*, vol. 9, no. 8, pp. 2933–2939, Aug. 2003.

[88] G. O. Silva, X. He, J. S. Parker, M. L. Gatza, L. A. Carey, J. P. Hou, S. L. Moulder, P. K. Marcom, J. Ma, J. M. Rosen, and C. M. Perou, "Cross-species DNA copy number analyses identifies multiple 1q21-q23 subtype-specific driver genes for breast cancer.," *Breast Cancer Res. Treat.*, vol. 152, no. 2, pp. 347–56, Jun. 2015.

[89] V. Gelsi-Boyer, B. Orsetti, N. Cervera, P. Finetti, F. Sircoulomb, C. Rougé, L. Lasorsa, A. Letessier, C. Ginestier, F. Monville, S. Esteyriès, J. Adélaïde, B. Esterni, C. Henry, S. P. Ethier, F. Bibeau, M.-J. Mozziconacci, E. Charafe-Jauffret, J. Jacquemier, F. Bertucci, D. Birnbaum, C. Theillet, and M. Chaffanet, "Comprehensive profiling of 8p11-12 amplification in breast cancer.," *Mol. Cancer Res.*, vol. 3, no. 12, pp. 655–67, 2005.

[90] J. Xu, Y. Chen, and O. I. Olopade, "MYC and Breast Cancer.," *Genes Cancer*, vol. 1, no. 6, pp. 629–40, Jun. 2010.

[91] M. Wagner, M. Koslowski, C. Paret, M. Schmidt, O. Türeci, and U. Sahin, "NCOA3 is a selective co-activator of estrogen receptor α-mediated transactivation of PLAC1 in MCF-7 breast cancer cells.," *BMC Cancer*, vol. 13, p. 570, Jan. 2013.

[92] G. S. Pryhuber, H. L. Huyck, R. J. Staversky, J. N. Finkelstein, and M. A. O'Reilly, "Tumor necrosis factor-alpha-induced lung cell expression of antiapoptotic genes TRAF1 and cIAP2.," *Am. J. Respir. Cell Mol. Biol.*, vol. 22, no. 2, pp. 150–6, Feb. 2000.

[93] R. Natrajan, M. B. K. Lambros, F. C. Geyer, C. Marchio, D. S. P. Tan, R. Vatcheva, K.-K. Shiu, D. Hungermann, S. M. Rodriguez-Pinilla, J. Palacios, A. Ashworth, H. Buerger, and J. S. Reis-Filho, "Loss of 16q in high grade breast cancer is associated with estrogen receptor status: Evidence for progression in tumors with a luminal phenotype?," *Genes. Chromosomes Cancer*, vol. 48, no. 4, pp. 351–65, Apr. 2009.

[94] G. Berx and F. Van Roy, “The E-cadherin/catenin complex: an important gatekeeper in breast cancer tumorigenesis and malignant progression.,” *Breast Cancer Res.*, vol. 3, no. 5, pp. 289–93, Jan. 2001.

[95] K. Shukla, A. K. Sharma, A. Ward, R. Will, T. Hielscher, A. Balwierz, C. Breunig, E. Münstermann, R. König, I. Keklikoglou, and S. Wiemann, “MicroRNA-30c-2-3p negatively regulates NF-κB signaling and cell cycle progression through downregulation of TRADD and CCNE1 in breast cancer.,” *Mol. Oncol.*, vol. 9, no. 6, pp. 1106–19, Jun. 2015.

[96] F. André and C. C. Zielinski, “Optimal strategies for the treatment of metastatic triple-negative breast cancer with currently approved agents.,” *Ann. Oncol.*, vol. 23 Suppl 6, no. suppl 6, p. vi46-51, Aug. 2012.

[97] L. Page, S. Brin, R. Motwani, and T. Winograd, “The PageRank Citation Ranking: Bringing Order to the Web.” Stanford InfoLab, 11-Nov-1999.

[98] S. Brin and L. Page, “The Anatomy of a Large-Scale Hypertextual Web Search Engine,” Stanford InfoLab, 1998.

[99] E. Khurana, Y. Fu, V. Colonna, X. J. Mu, H. M. Kang, T. Lappalainen, A. Sboner, L. Lochovsky, J. Chen, A. Harmanci, J. Das, A. Abyzov, S. Balasubramanian, K. Beal, D. Chakravarty, D. Challis, Y. Chen, D. Clarke, L. Clarke, F. Cunningham, U. S. Evani, P. Flicek, R. Fragoza, E. Garrison, R. Gibbs, Z. H. Gümüs, J. Herrero, N. Kitabayashi, Y. Kong, K. Lage, V. Liluashvili, S. M. Lipkin, D. G. MacArthur, G. Marth, D. Muzny, T. H. Pers, G. R. S. Ritchie, J. A. Rosenfeld, C. Sisu, X. Wei, M. Wilson, Y. Xue, F. Yu, 1000 Genomes Project Consortium, E. T. Dermitzakis, H. Yu, M. A. Rubin, C. Tyler-Smith, and M. Gerstein, “Integrative annotation of variants from 1092 humans: application to cancer genomics.,” *Science*, vol. 342, no. 6154, p. 1235587, Oct. 2013.

[100] M. D’Antonio, F. D. Ciccarelli, "Integrated analysis of recurrent properties of cancer genes to identify novel drivers",*Genome Biol.*, vol. 14, no. 5, p. R52, 2013.

[101] P. A. Futreal, L. Coin, M. Marshall, T. Down, T. Hubbard, R. Wooster, N. Rahman, and M. R. Stratton, “A census of human cancer genes.,” *Nat. Rev. Cancer*, vol. 4, no. 3, pp. 177–83, Mar. 2004.

[102] J. L. Morrison, R. Breitling, D. J. Higham, D. R. Gilbert, A. Langville, C. Meyer, “GeneRank: Using search engine technology for the analysis of microarray experiments,” *BMC Bioinformatics*, vol. 6, no. 1, p. 233, 2005.

[103] A. L. Tarca, S. Draghici, P. Khatri, S. S. Hassan, P. Mittal, J.-S. Kim, C. J. Kim, J. P. Kusanovic, and R. Romero, “A novel signaling pathway impact analysis.,” *Bioinformatics*, vol. 25, no. 1, pp. 75–82, Jan. 2009.

[104] C. Winter, G. Kristiansen, S. Kersting, J. Roy, D. Aust, T. Knösel, P. Rümmele, B. Jahnke, V. Hentrich, F. Rückert, M. Niedergethmann, W. Weichert, M. Bahra, H. J. Schlitt, U. Settmacher, H. Friess, M. Büchler, H.-D. Saeger, M. Schroeder, C. Pilarsky, R. Grützmann, “Google Goes Cancer: Improving Outcome Prediction for Cancer Patients by Network-Based Ranking of Marker Genes,” *PLoS Comput. Biol.*, vol. 8, no. 5, p. e1002511, May 2012.

[105] Y. Liu, Q. Gu, J. P. Hou, J. Han, and J. Ma, “A network-assisted co-clustering algorithm to discover cancer subtypes based on gene expression.,” *BMC Bioinformatics*, vol. 15, no. 1, p. 37, Jan. 2014.

[106] F. Vandin, E. Upfal, and B. J. Raphael, “Algorithms for detecting significantly mutated pathways in cancer.,” *J. Comput. Biol.*, vol. 18, no. 3, pp. 507–22, Mar. 2011.

[107] E. O. Paull, D. E. Carlin, M. Niepel, P. K. Sorger, D. Haussler, and J. M. Stuart, "Discovering causal pathways linking genomic events to transcriptional states using Tied Diffusion Through Interacting Events (TieDIE).," *Bioinformatics*, vol. 29, no. 21, pp. 2757–64, Nov. 2013.

[108] X. Wang, T. Tao, X. Wang, A. Shakery, and C. Zhai, "DirichletRank: Solving the Zero-One Gap Problem of PageRank DirichletRank: Solving the Zero-One Gap Problem of," *ACM J. Name*, vol. 20, pp. 0–27.

[109] N. X. Vinh, M. Chetty, R. Coppel, and P. P. Wangikar, "Issues impacting genetic network reverse engineering algorithm validation using small networks," *Biochim. Biophys. Acta - Proteins Proteomics*, vol. 1824, no. 12, pp. 1434–1441, 2012.

[110] V. Pihur, S. Datta, "RankAggreg, an R package for weighted rank aggregation," *BMC Bioinformatics*, vol. 10, no. 1, p. 62, 2009.

[111] M. P. Schroeder, C. Rubio-Perez, D. Tamborero, A. Gonzalez-Perez, and N. Lopez-Bigas, "OncodriveROLE classifies cancer driver genes in loss of function and activating mode of action.," *Bioinformatics*, vol. 30, no. 17, pp. i549-55, Sep. 2014.

[112] H. Carter, S. Chen, L. Isik, S. Tyekucheva, V. E. Velculescu, K. W. Kinzler, B. Vogelstein, and R. Karchin, "Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations.," *Cancer Res.*, vol. 69, no. 16, pp. 6660–7, Aug. 2009.

[113] M. S. Lawrence, P. Stojanov, P. Polak, G. V Kryukov, K. Cibulskis, A. Sivachenko, S. L. Carter, C. Stewart, C. H. Mermel, S. A. Roberts, A. Kiezun, P. S. Hammerman, A. McKenna, Y. Drier, L. Zou, A. H. Ramos, T. J. Pugh, N. Stransky, E. Helman, J. Kim, C. Sougnez, L. Ambrogio, E. Nickerson, E. Shefler, M. L. Cortés, D. Auclair, G. Saksena, D. Voet, M. Noble, D. DiCara, P. Lin, L. Lichtenstein, D. I. Heiman, T. Fennell, M. Imielinski, B. Hernandez, E. Hodis, S. Baca, A. M. Dulak, J. Lohr, D.-A. Landau, C. J. Wu, J. Melendez-Zajgla, A. Hidalgo-Miranda, A. Koren, S. A. McCarroll, J. Mora, R. S. Lee, B. Crompton, R. Onofrio, M. Parkin, W. Winckler, K. Ardlie, S. B. Gabriel, C. W. M. Roberts, J. A. Biegel, K. Stegmaier, A. J. Bass, L. A. Garraway, M. Meyerson, T. R. Golub, D. A. Gordenin, S. Sunyaev, E. S. Lander, and G. Getz, "Mutational heterogeneity in cancer and the search for new cancer-associated genes.," *Nature*, vol. 499, no. 7457, pp. 214–8, Jul. 2013.

[114] A. Manolakos, I. Ochoa, K. Venkat, A. J. Goldsmith, O. Gevaert "CaMoDi: a new method for cancer module discovery," *BMC Genomics*, vol. 15, no. Suppl 10, p. S8, 2014.

[115] A. Gonzalez-Perez and N. Lopez-Bigas, "Functional impact bias reveals cancer drivers.," *Nucleic Acids Res.*, vol. 40, no. 21, p. e169, Nov. 2012.

[116] S. Ng, E. A. Collisson, A. Sokolov, T. Goldstein, A. Gonzalez-Perez, N. Lopez-Bigas, C. Benz, D. Haussler, and J. M. Stuart, "PARADIGM-SHIFT predicts the function of mutations in multiple cancers using pathway impact analysis.," *Bioinformatics*, vol. 28, no. 18, pp. i640–i646, Sep. 2012.

[117] N. Bansal, A. Blum, and S. Chawla, "Correlation Clustering." In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing* (STOC '15).

[118] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A K-Means Clustering Algorithm," *Appl. Stat.*, vol. 28, no. 1, p. 100, 1979.

[119] M. Charikar, V. Guruswami, and A. Wirth, "Clustering with Qualitative Information." *J. Comput. Syst. Sci.* 71, 3, 360-383, Oct. 2005

[121] A. Torkamani and N. J. Schork, "Identification of rare cancer driver mutations by network

reconstruction.," *Genome Res.*, vol. 19, no. 9, pp. 1570–8, Sep. 2009.

[122] B. Vogelstein, N. Papadopoulos, V. E. Velculescu, S. Zhou, L. A. Diaz, K. W. Kinzler, L. "Cancer genome landscapes.," *Science*, vol. 339, no. 6127, pp. 1546–58, Mar. 2013.

[123] C. Tomasetti, L. Marchionni, M. A. Nowak, G. Parmigiani, and B. Vogelstein, "Only three driver gene mutations are required for the development of lung and colorectal cancers.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 112, no. 1, pp. 118–23, Jan. 2015.

[124] M. D. Leiserson, H.-T. Wu, F. Vandin, B. J. Raphael, "CoMEt: a statistical approach to identify combinations of mutually exclusive alterations in cancer," *Genome Biol.*, vol. 16, no. 1, p. 160, Dec. 2015.

[125] G. J. Puleo and O. Milenkovic, "Correlation Clustering with Constrained Cluster Sizes and Extended Weights Bounds," *SIAM J. Optim.*, vol. 25, no. 3, pp. 1857–1872, Jan. 2015.

[126] S. Sridhar, V. Bittorf, J. Liu, C. Zhang, C. Ré, and S. J. Wright, "An Approximate, Efficient Solver for LP Rounding."

[127] J. Gao, B. A. Aksoy, U. Dogrusoz, G. Dresdner, B. Gross, S. O. Sumer, Y. Sun, A. Jacobsen, R. Sinha, E. Larsson, E. Cerami, C. Sander, and N. Schultz, "Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal.," *Sci. Signal.*, vol. 6, no. 269, p. pl1, Apr. 2013.

[128] C. H. Mermel, S. E. Schumacher, B. Hill, M. L. Meyerson, R. Beroukhim, and G. Getz, "GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers.," *Genome Biol.*, vol. 12, no. 4, p. R41, Jan. 2011.

[129] R. A. Fisher, "On the Mathematical Foundations of Theoretical Statistics," *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.*, vol. 222, no. 594–604, pp. 309–368, Jan. 1922.

[130] Ö. Babur, M. Gönen, "Systematic identification of cancer driving signaling pathways based on mutual exclusivity of genomic alterations," *Genome Biol.*, vol. 16, no. 1, p. 45, Dec. 2015.

[131] B. Rosner and D. Grove, "Use of the Mann-Whitney U-test for clustered data.," *Stat. Med.*, vol. 18, no. 11, pp. 1387–400, Jun. 1999.

[132] D. W. Zimmerman, "Comparative Power of Student *T* Test and Mann-Whitney *U* Test for Unequal Sample Sizes and Variances," *J. Exp. Educ.*, vol. 55, no. 3, pp. 171–174, Apr. 1987.

[133] V. Stambolic, A. Suzuki, J. L. de la Pompa, G. M. Brothers, C. Mirtsos, T. Sasaki, J. Ruland, J. M. Penninger, D. P. Siderovski, T. W. Mak, N. "Negative regulation of PKB/Akt-dependent cell survival by the tumor suppressor PTEN.," *Cell*, vol. 95, no. 1, pp. 29–39, Oct. 1998.

[134] S. Inoue, Z. Hao, A. J. Elia, D. Cescon, L. Zhou, J. Silvester, B. Snow, I. S. Harris, M. Sasaki, W. Y. Li, M. Itsumi, K. Yamamoto, T. Ueda, C. Dominguez-Brauer, C. Gorrini, I. I. C. Chio, J. Haight, A. You-Ten, S. McCracken, A. Wakeham, D. Ghazarian, L. J. Z. Penn, G. Melino, and T. W. Mak, "Mule/Huwe1/Arf-BP1 suppresses Ras-driven tumorigenesis by preventing c-Myc/Miz1-mediated down-regulation of p21 and p15.," *Genes Dev.*, vol. 27, no. 10, pp. 1101–14, May 2013.

[135] Y.-J. Shann, C. Cheng, C.-H. Chiao, D.-T. Chen, P.-H. Li, and M.-T. Hsu, "Genome-wide mapping and characterization of hypomethylated sites in human tissues and breast cancer cell lines.," *Genome Res.*, vol. 18, no. 5, pp. 791–801, May 2008.

[136] D. Roy, S.-H. Sin, B. Damania, and D. P. Dittmer, "Tumor suppressor genes FHIT and

WWOX are deleted in primary effusion lymphoma (PEL) cell lines.," *Blood*, vol. 118, no. 7, pp. e32-9, Aug. 2011.

[137] I. Azimi, S. J. Roberts-Thomson, and G. R. Monteith, "Calcium influx pathways in breast cancer: opportunities for pharmacological intervention," *Br. J. Pharmacol.*, vol. 171, no. 4, pp. 945–960, Feb. 2014.

[138] C. Magnusson, J. Liu, R. Ehrnström, J. Manjer, K. Jirström, T. Andersson, and A. Sjölander, "Cysteinyl leukotriene receptor expression pattern affects migration of breast cancer cells and survival of breast cancer patients.," *Int. J. cancer*, vol. 129, no. 1, pp. 9–22, Jul. 2011.

[139] A. Vazquez, E. E. Bond, A. J. Levine, and G. L. Bond, "The genetics of the p53 pathway, apoptosis and cancer therapy.," *Nat. Rev. Drug Discov.*, vol. 7, no. 12, pp. 979–87, Dec. 2008.

[140] K. Stemke-Hale, A. M. Gonzalez-Angulo, A. Lluch, R. M. Neve, W.-L. Kuo, M. Davies, M. Carey, Z. Hu, Y. Guan, A. Sahin, W. F. Symmans, L. Pusztai, L. K. Nolden, H. Horlings, K. Berns, M.-C. Hung, M. J. van de Vijver, V. Valero, J. W. Gray, R. Bernards, G. B. Mills, and B. T. Hennessy, "An integrative genomic and proteomic analysis of PIK3CA, PTEN, and AKT mutations in breast cancer.," *Cancer Res.*, vol. 68, no. 15, pp. 6084–91, Aug. 2008.

[141] V. Clement, P. Sanchez, N. de Tribolet, I. Radovanovic, A. Ruiz i Altaba, "HEDGEHOG-GLI1 signaling regulates human glioma growth, cancer stem cell self-renewal, and tumorigenicity.," *Curr. Biol.*, vol. 17, no. 2, pp. 165–72, Jan. 2007.

[142] N. V. Serão, K. R. Delfino, B. R. Southey, J. E. Beever, S. L. Rodriguez-Zas, "Cell cycle and aging, morphogenesis, and response to stimuli genes are individualized biomarkers of glioblastoma progression and survival," *BMC Med. Genomics*, vol. 4, no. 1, p. 49, Dec. 2011.

[143] C. Gratas, Y. Tohma, E. G. Van Meir, M. Klein, M. Tenan, N. Ishii, O. Tachibana, P. Kleihues, and H. Ohgaki, "Fas ligand expression in glioblastoma cell lines and primary astrocytic brain tumors.," *Brain Pathol.*, vol. 7, no. 3, pp. 863–9, Jul. 1997.

[144] J. A. DiMasi, J. M. Reichert, L. Feldman, and A. Malins, "Clinical approval success rates for investigational cancer drugs.," *Clin. Pharmacol. Ther.*, vol. 94, no. 3, pp. 329–35, Sep. 2013.

[145] K. S. Garman, J. R. Nevins, and A. Potti, "Genomic strategies for personalized cancer therapy.," *Hum. Mol. Genet.*, no. R2, pp. R226-32, Oct. 2007.

[146] R. V. J. Chari, "Targeted cancer therapy: conferring specificity to cytotoxic drugs.," *Acc. Chem. Res.*, vol. 41, no. 1, pp. 98–107, Jan. 2008.

[147] M. Griffith, O. L. Griffith, A. C. Coffman, J. V Weible, J. F. McMichael, N. C. Spies, J. Koval, I. Das, M. B. Callaway, J. M. Eldred, C. A. Miller, J. Subramanian, R. Govindan, R. D. Kumar, R. Bose, L. Ding, J. R. Walker, D. E. Larson, D. J. Dooling, S. M. Smith, T. J. Ley, E. R. Mardis, and R. K. Wilson, "DGIdb: mining the druggable genome.," *Nat. Methods*, vol. 10, no. 12, pp. 1209–10, Dec. 2013.

[148] J. Mestres, E. Gregori-Puigjané, S. Valverde, and R. V Solé, "Data completeness—the Achilles heel of drug-target networks," *Nat. Biotechnol.*, vol. 26, no. 9, pp. 983–984, Sep. 2008.

[149] J. C. Costello, L. M. Heiser, E. Georgii, M. Gönen, M. P. Menden, N. J. Wang, M. Bansal, M. Ammad-Ud-Din, P. Hintsanen, S. A. Khan, J.-P. Mpindi, O. Kallioniemi, A. Honkela, T. Aittokallio, K. Wennerberg, J. J. Collins, D. Gallahan, D. Singer, J. Saez-Rodriguez, S.

Kaski, J. W. Gray, and G. Stolovitzky, “A community effort to assess and improve drug sensitivity prediction algorithms.,” *Nat. Biotechnol.*, vol. 32, no. 12, pp. 1202–1212, Jun. 2014.

[150] I. Cortés-Ciriano, G. J. P. van Westen, G. Bouvier, M. Nilges, J. P. Overington, A. Bender, and T. E. Malliavin, “Improved large-scale prediction of growth inhibition patterns using the NCI60 cancer cell line panel.,” *Bioinformatics*, vol. 32, no. 1, pp. 85–95, Jan. 2016.

[151] L. C. Wienkers and T. G. Heath, “Predicting in vivo drug interactions from in vitro drug discovery data,” *Nat. Rev. Drug Discov.*, vol. 4, no. 10, pp. 825–833, Oct. 2005.

[152] Cancer Genome Atlas Research Network, J. N. Weinstein, E. A. Collisson, G. B. Mills, K. R. M. Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, and J. M. Stuart, “The Cancer Genome Atlas Pan-Cancer analysis project.,” *Nat. Genet.*, vol. 45, no. 10, pp. 1113–20, Oct. 2013.

[153] J. Hua, Z. Xiong, J. Lowey, E. Suh, and E. R. Dougherty, “Optimal number of features as a function of sample size for various classification rules.,” *Bioinformatics*, vol. 21, no. 8, pp. 1509–15, Apr. 2005.

[154] J. R. Landis and G. G. Koch, “The Measurement of Observer Agreement for Categorical Data,” *Biometrics*, vol. 33, no. 1, p. 159, Mar. 1977.

[155] I. E. Smith and M. Dowsett, “Aromatase Inhibitors in Breast Cancer,” *http://dx.doi.org/10.1056/NEJMra023246*, 2009.

[156] C. Palmieri, G. J. Cheng, S. Saji, M. Zelada-Hedman, A. Wärri, Z. Weihua, S. Van Noorden, T. Wahlstrom, R. C. Coombes, M. Warner, and J.-A. Gustafsson, “Estrogen receptor beta in breast cancer.,” *Endocr. Relat. Cancer*, vol. 9, no. 1, pp. 1–13, Mar. 2002.

[157] L.-A. Haldosén, C. Zhao, and K. Dahlman-Wright, “Estrogen receptor beta in breast cancer,” *Mol. Cell. Endocrinol.*, vol. 382, no. 1, pp. 665–672, 2014.

[158] T. Schlange, Y. Matsuda, S. Lienhard, A. Huber, and N. E. Hynes, “Autocrine WNT signaling contributes to breast cancer cell proliferation via the canonical WNT pathway and EGFR transactivation.,” *Breast Cancer Res.*, vol. 9, no. 5, p. R63, 2007.

[159] F. C. Geyer, M. Lacroix-Triki, K. Savage, M. Arnedos, M. B. Lambros, A. MacKay, R. Natrajan, and J. S. Reis-Filho, “β-Catenin pathway activation in breast cancer is associated with triple-negative phenotype but not with CTNNB1 mutation,” *Mod. Pathol.*, vol. 24, no. 2, pp. 209–231, Feb. 2011.

[160] P. Gustafson and D. Le Nhu, “Comparing the effects of continuous and discrete covariate mismeasurement, with emphasis on the dichotomization of mismeasured predictors.,” *Biometrics*, vol. 58, no. 4, pp. 878–87, Dec. 2002.

[161] S. B. Horwitz, “Taxol (paclitaxel): mechanisms of action.,” *Ann. Oncol.*, vol. 5 Suppl 6, pp. S3-6, 1994.

[162] E. Zwick, J. Bange, and A. Ullrich, “Receptor tyrosine kinase signalling as a target for cancer intervention strategies,” *Endocr. Relat. Cancer*, vol. 8, no. 8, pp. 161–173, 2001.

[163] F. Wang, Y. Chen, D. Zhang, Q. Zhang, D. Zheng, L. Hao, Y. Liu, C. Duan, L. Jia, and G. Liu, “Folate-mediated targeted and intracellular delivery of paclitaxel using a novel deoxycholic acid-O-carboxymethylated chitosan-folic acid micelles.,” *Int. J. Nanomedicine*, vol. 7, pp. 325–37, 2012.

[164] E. Roger, S. Kalscheuer, A. Kirtane, B. R. Guru, A. E. Grill, J. Whittum-Hudson, and J. Panyam, “Folic acid functionalized nanoparticles for enhanced oral drug delivery.,” *Mol. Pharm.*, vol. 9, no. 7, pp. 2103–10, Jul. 2012.

[165] M. Keniry and R. Parsons, “The role of PTEN signaling perturbations in cancer and in targeted therapy,” *Oncogene*, vol. 27, no. 41, pp. 5477–5485, Sep. 2008.

[166] J. Li, Y. Zhang, J. Zhao, F. Kong, and Y. Chen, “Overexpression of miR-22 reverses paclitaxel-induced chemoresistance through activation of PTEN signaling in p53-mutated colon cancer cells.,” *Mol. Cell. Biochem.*, vol. 357, no. 1–2, pp. 31–8, Nov. 2011.

[167] O. Kranenburg, “The KRAS oncogene: Past, present, and future,” *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer*, vol. 1756, no. 2. pp. 81–82, 2005.

[168] X.-H. Zhang, J.-Y. Shin, J.-O. Kim, J.-E. Oh, S.-A. Yoon, C.-K. Jung, and J.-H. Kang, “Synergistic antitumor efficacy of sequentially combined paclitaxel with sorafenib in vitro and in vivo NSCLC models harboring KRAS or BRAF mutations,” *Cancer Lett.*, vol. 322, no. 2, pp. 213–222, 2012.

[169] C. J. Fabian, “The what, why and how of aromatase inhibitors: hormonal agents for treatment and prevention of breast cancer.,” *Int. J. Clin. Pract.*, vol. 61, no. 12, pp. 2051–63, Dec. 2007.

[170] J. Chou, S. Provot, and Z. Werb, “GATA3 in development and cancer differentiation: cells GATA have it!,” *J. Cell. Physiol.*, vol. 222, no. 1, pp. 42–9, Jan. 2010.

[171] F. Milanezi, S. Carvalho, and F. C. Schmitt, “EGFR/HER2 in breast cancer: a biological approach for molecular diagnosis and therapy.,” *Expert Rev. Mol. Diagn.*, vol. 8, no. 4, pp. 417–34, Jul. 2008.

[172] P. Liu, H. Cheng, T. M. Roberts, and J. J. Zhao, “Targeting the phosphoinositide 3-kinase pathway in cancer.,” *Nat. Rev. Drug Discov.*, vol. 8, no. 8, pp. 627–44, Aug. 2009.

[173] R. Wooster, S. L. Neuhausen, J. Mangion, Y. Quirk, D. Ford, N. Collins, K. Nguyen, S. Seal, T. Tran, and D. Averill, “Localization of a breast cancer susceptibility gene, BRCA2, to chromosome 13q12-13.,” *Science*, vol. 265, no. 5181, pp. 2088–90, Sep. 1994.

[174] W. D. Foulkes, K. Metcalfe, P. Sun, W. M. Hanna, H. T. Lynch, P. Ghadirian, N. Tung, O. I. Olopade, B. L. Weber, J. McLennan, I. A. Olivotto, L. R. Bégin, and S. A. Narod, “Estrogen receptor status in BRCA1- and BRCA2-related breast cancer: the influence of age, grade, and histological type.,” *Clin. Cancer Res.*, vol. 10, no. 6, pp. 2029–34, Mar. 2004.

[175] C. Sessa and O. Pagani, “Docetaxel and epirubicin in advanced breast cancer.,” *Oncologist*, vol. 6 Suppl 3, pp. 13–6, 2001.

[176] Z.-S. Chen and A. K. Tiwari, “Multidrug resistance proteins (MRPs/ABCCs) in cancer chemotherapy and genetic diseases.,” *FEBS J.*, vol. 278, no. 18, pp. 3226–45, Sep. 2011.

[177] C. P. Schröder, G. B. Wisman, S. de Jong, W. T. van der Graaf, M. H. Ruiters, N. H. Mulder, L. F. de Leij, A. G. van der Zee, and E. G. de Vries, “Telomere length in breast cancer patients before and after chemotherapy with or without stem cell transplantation.,” *Br. J. Cancer*, vol. 84, no. 10, pp. 1348–53, May 2001.

[178] T. Ge, S. Zdonik, and S. Madden, “Top-k Queries on Uncertain Data: On Score Distribution and Typical Answers.” In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data* (SIGMOD '09), Carsten Binnig and Benoit Dageville (Eds.). ACM, New York, NY, USA, 375-388.

[179] D. Gajria and S. Chandarlapaty, “HER2-amplified breast cancer: mechanisms of trastuzumab resistance and novel targeted therapies.,” *Expert Rev. Anticancer Ther.*, vol. 11, no. 2, pp. 263–75, Feb. 2011.

[180] V. C. Jordan, “Molecular mechanisms of antiestrogen action in breast cancer.,” *Breast Cancer Res. Treat.*, vol. 31, no. 1, pp. 41–52, 1994.

[181] F. Andre, K. Broglio, H. Roche, M. Martin, J. R. Mackey, F. Penault-Llorca, G. N. Hortobagyi, and L. Pusztai, “Estrogen receptor expression and efficacy of docetaxel-containing adjuvant chemotherapy in patients with node-positive breast cancer: results from a pooled analysis.,” *J. Clin. Oncol.*, vol. 26, no. 16, pp. 2636–43, Jun. 2008.

[182] C. M. Perou, T. Sørlie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, C. A. Rees, J. R. Pollack, D. T. Ross, H. Johnsen, L. A. Akslen, O. Fluge, A. Pergamenschikov, C. Williams, S. X. Zhu, P. E. Lønning, A. L. Børresen-Dale, P. O. Brown, and D. Botstein, “Molecular portraits of human breast tumours.,” *Nature*, vol. 406, no. 6797, pp. 747–52, Aug. 2000.

[183] M. Martin, A. Villar, A. Sole-Calvo, R. Gonzalez, B. Massuti, J. Lizon, C. Camps, A. Carrato, A. Casado, M. T. Candel, J. Albanell, J. Aranda, B. Munarriz, J. Campbell, and E. Diaz-Rubio, “Doxorubicin in combination with fluorouracil and cyclophosphamide (i.v. FAC regimen, day 1, 21) versus methotrexate in combination with fluorouracil and cyclophosphamide (i.v. CMF regimen, day 1, 21) as adjuvant chemotherapy for operable breast cancer,” *Ann. Oncol.*, vol. 14, no. 6, pp. 833–842, Jun. 2003.

[184] S. Cleator, W. Heller, and R. C. Coombes, “Triple-negative breast cancer: therapeutic options.,” *Lancet. Oncol.*, vol. 8, no. 3, pp. 235–44, Mar. 2007.

[185] E. A. Cherman, M. C. Monard, and J. Metz, “Multi-label Problem Transformation Methods: a Case Study,” *CLEI Electron. J.*, vol. 14, no. 4, 2011.

[186] V. Marx, “Drilling into big cancer-genome data,” *Nat. Methods*, vol. 10, no. 4, pp. 293–297, Mar. 2013.

[187] A. Colaprico, T. C. Silva, C. Olsen, L. Garofano, C. Cava, D. Garolini, T. S. Sabedot, T. M. Malta, S. M. Pagnotta, I. Castiglioni, M. Ceccarelli, G. Bontempi, and H. Noushmehr, “*TCGAbiolinks* : an R/Bioconductor package for integrative analysis of TCGA data,” *Nucleic Acids Res.*, vol. 44, no. 8, pp. e71–e71, May 2016.

[188] D. S. Wishart, C. Knox, A. C. Guo, D. Cheng, S. Shrivastava, D. Tzur, B. Gautam, and M. Hassanali, “DrugBank: a knowledgebase for drugs, drug actions and drug targets.,” *Nucleic Acids Res.*, vol. 36, no. Database issue, pp. D901-6, Jan. 2008.

[189] I. M. Johnstone, D. M. Titterington, “Statistical challenges of high-dimensional data.,” *Philos. Trans. A. Math. Phys. Eng. Sci.*, vol. 367, no. 1906, pp. 4237–53, Nov. 2009.

[190] M. Kanehisa and S. Goto, “KEGG: kyoto encyclopedia of genes and genomes.,” *Nucleic Acids Res.*, vol. 28, no. 1, pp. 27–30, Jan. 2000.

[191] G. Wu, X. Feng, and L. Stein, “A human functional protein interaction network and its application to cancer data analysis.,” *Genome Biol.*, vol. 11, no. 5, p. R53, 2010.

[192] L. Breiman, “Random Forests,” *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[193] M. B. Kursa, “Robustness of Random Forest-based gene selection methods,” *BMC Bioinformatics*, vol. 15, no. 1, p. 8, 2014.

[194] S. Dawood, S. D. Merajver, P. Viens, P. B. Vermeulen, S. M. Swain, T. A. Buchholz, L. Y. Dirix, P. H. Levine, A. Lucci, S. Krishnamurthy, F. M. Robertson, W. A. Woodward, W. T. Yang, N. T. Ueno, and M. Cristofanilli, “International expert panel on inflammatory breast cancer: consensus statement for standardized diagnosis and treatment.,” *Ann. Oncol.*, vol. 22, no. 3, pp. 515–23, Mar. 2011.

[195] D. P. Atchley, C. T. Albarracin, A. Lopez, V. Valero, C. I. Amos, A. M. Gonzalez-Angulo, G. N. Hortobagyi, and B. K. Arun, “Clinical and Pathologic Characteristics of Patients With BRCA-Positive and BRCA-Negative Breast Cancer,” *J. Clin. Oncol.*, vol. 26, no. 26, pp. 4282–4288, Sep. 2008.

[196] E. W. Drrksrra, “A Note on Two Problems in Connexion with Graphs,” *Numer. Math.*, pp. 269–27.

[197] G. Csárdi and T. Nepusz, “The igraph software package for complex network research.”

[198] Z. Dong, N. Zhang, C. Li, H. Wang, Y. Fang, J. Wang, and X. Zheng, “Anticancer drug sensitivity prediction in cell lines from baseline gene expression through recursive feature selection.,” *BMC Cancer*, vol. 15, no. 1, p. 489, Jan. 2015.

[199] A. Max Kuhn Contributions form Jed Wing, S. Weston, A. Williams, and M. Max Kuhn, “The caret Package Title Classification and Regression Training Description Misc functions for training and plotting classification and regression models,” 2008.

[200] E. Hüllermeier, J. Fürnkranz, W. Cheng, and K. Brinker, “Label Ranking by Learning Pairwise Preferences.” *Artif. Intell.* 172, 16-17. Nov. 2008.

[201] N. Ailon, M. Charikar, A. Newman, A. P. Sloan, and H. B. Wentz Jr, “Aggregating Inconsistent Information: Ranking and Clustering.” *J. ACM* 55, 5, Article 23. Nov 2008

[202] N. P. Tatonetti, P. P. Ye, R. Daneshjou, R. B. Altman, “Data-driven prediction of drug effects and interactions.,” *Sci. Transl. Med.*, vol. 4, p. 125, Mar. 2012.

[203] C. Dwork, R. Kumar, M. Naor, and D. Sivakumar, “Rank Aggregation Revisited.” In *Proceedings of the 10th international conference on World Wide Web* (WWW '01). ACM, New York, NY, USA, 613-622.