THE CANCER TARGETOME AND ITS APPLICATION TO A PATHWAY PERSPECTIVE ON TARGETED THERAPY RESPONSE IN ACUTE MYELOID LEUKEMIA

By

Aurora Skye Blucher

A DISSERTATION

Presented to the Department of Medical Informatics and Clinical Epidemiology and the
Oregon Health & Science University
School of Medicine
in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

September 2018

School of Medicine
Oregon Health & Science University

Certificate of Approval

This is to certify that the PhD Dissertation of

Aurora S. Blucher

"The Cancer Targetome and its Application to a Pathway Perspective on Targeted Therapy Response in Acute Myeloid Leukemia"

Has been approved

_____
Shannon K. McWeeney, PhD

_____
Aaron Cohen, MD

_____
Guanming Wu, PhD

_____
Arie Baratt, PhD

_____
Jeffrey Tyner, PhD

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

This work was made possible by the support of many over the last six years. My deepest gratitude to family, friends, colleagues, and mentors who have been a part of this journey.

Thank you to my dissertation advisory committee for their support, guidance, and enthusiasm: Drs. Aaron Cohen, Arie Baratt, Jeffrey Tyner, Guanming Wu, and Shannon McWeeney. Guanming and Shannon, it is difficult to find words to express my appreciation for your mentorship. Thank you from the bottom of my heart. I hope to put back into the world some of the light that you have inspired within me.

Thank you to DMICE faculty, staff, and especially the National Library of Medicine Fellows for an endless supply of support. Special thanks to Drs. Dian Chase, Ted Laderas, Dave Gibbs, Nathan Lazar, and Dana Womack for thoughtful words along the way.

Thank you to the Head and Neck Squamous Cell Carcinoma Group and the BeatAML Consortium for the incredible opportunities to collaborate in cutting-edge and inspiring work. Thank you to the patients who have donated their precious samples to these research efforts.

Throughout my predoctoral training, I have been fortunate to be supported by the National Library of Medicine Informatics Training Grant (T15LM007088). This fellowship has given me the opportunity to nurture both my voice and my vocation.

Finally, thank you to Saejin for being my rock.

ABSTRACT

Despite incredible advances in cancer treatment in recent years, precision oncology faces many ongoing challenges in selecting the best treatment for each patient. First, we have yet to untangle the mechanisms by which many of our drugs act, especially with regards to the targets (receptors) to which they bind in the human body. To this aim we present the Cancer Targetome, which collects drug-target interactions for FDA-approved cancer drugs and proposes an evidence framework for the supporting information accompanying each drug-target interaction. This framework enables us to prioritize drug-target interactions according to the strength of supporting evidence and then select the interactions for which we have the most confidence. Second, we have much to learn about how individual gene-level mutations affect a patient's overall response to drug therapies. Here, we take a pathway perspective on patient mutational burden in acute myeloid leukemia (AML) and show that this allows us to capture meaningful associations with drug sensitivity (or drug resistance). Importantly, these relationships are not necessarily captured when considering only the gene-level perspective. We further investigate the relationship between pathway mutational burden and drug response through factor graph modeling, which allows us to infer the impact of gene mutations on pathway signaling. In parallel, we model drug binding on these same pathways using quantitative binding assay information from the Cancer Targetome. This work is foundational for future efforts for combining both patient mutation and drug information into a unified model to help guide selection of the best therapy for each cancer patient.

CHAPTER 1. INTRODUCTION

Modern cancer treatment is centered on the paradigm of targeted therapies, or cancer drugs that interact with biological targets (receptors) specific to cancer cells. Beginning with the drug imatinib, the first targeted therapy approved in 2001 and a break-though treatment for chronic myeloid leukemia (CML), targeted therapies have ushered in the promise of treating cancers based on alterations unique to cancer cells and not present in normal or healthy cells [1,2]. Imatinib, or Gleevec, was the first in a class of drugs known as kinase inhibitors and works by binding to the target kinase ABL1 portion of the BCR-ABL fusion protein, caused by a chromosomal fusion known as the Philadelphia chromosome. This chromosomal event is the driving cause behind CML and results in the BCR-ABL protein being turned continually “on” in cancer cells [2,3]. The advent of targeted therapies presented a game-changer for cancer therapies, which traditionally are cytotoxic (cell-killing) and thus kill both healthy and normal cells, resulting in severe side effects for the patient [1]. At present, over a hundred targeted therapies are currently FDA-approved for cancer treatment, with many more under development.

However, despite many successful targeted therapies, cancer treatment today faces a new set of pressing challenges. Two critical challenges include lack of patient response to therapy and cancer relapse, or re-occurrence, following treatment [1]. In the first scenario, many patients do not respond to standard-of-care therapies for their cancer types, and may not necessarily possess any biomarkers, or genetic alterations, that we currently can match with

an appropriate targeted therapy. In the second scenario, patients may exhibit an initial (often strong) response to therapy but then experience cancer relapse and often exhibit resistance to cancer drugs. In recent years it has come to light that this resistance may be due to the tumor harboring multiple sub-clones of cancer cells (tumor heterogeneity), each with distinct aberrations that can contribute to resistance [1,4]. Combination of therapies have been hailed as a way to overcome these challenges, by targeting multiple aberrations or pathways at once to overcome drug resistance or address tumor subclonal populations.

In order to meet these challenges in precision cancer treatment, we require an improved mechanistic understanding of both how cancer drugs interact with targets and how individual patient genetic aberrations affect this process. While it has become well-accepted in recent years that cancer drugs can and often do interact with multiple targets [5,6], characterization of such drug-to-target relationships is often lacking. In particular, finding target information across all known targets for any particular drug presents as a daunting endeavor that requires substantial manual effort and reconciliation across sources [7]. Current efforts in cancer biology, genetics, and bioinformatics require that we have a working knowledge of these drug-target interactions that is accessible and accompanied by transparent supporting information. In particular, we must have an understanding of both primary and secondary targets for our drugs if we are to pursue a more complete understanding of drug mechanistic action. Building upon drug-to-target relationships, we must translate this information to the biological context of cellular pathways, or groups of

functionally interacting molecular entities. By moving to a pathway context, we can begin to think about the ‘scope’ or reach of our drugs and begin to ask questions about the potential druggable space for particular cancers. This is especially key if we are to meet the challenge of drug resistance, which can be mediated at the pathway level through compensatory pathway signaling [1].

In addition to improving our mechanistic understanding of drug action, we require improved knowledge of how individual patient aberrations, such as gene mutations, effect this process and the overall resulting effect on patient response to a drug. Traditionally, pharmacogenomics approaches have studied relationships between gene-level aberrations and drug response. But given the challenges we face, we must begin to also translate this knowledge into a biological pathway context. Patient genetic mutations represent a profile of dysregulation, and through mapping to pathways, we may be able to create a bigger, more connected picture of dysregulation. This framework allows us to consider the impact of multiple mutations, in the same pathway or across pathways, and the interplay with drug action, all in the context of biological pathways. In this manner, we are working towards frameworks that incorporate both individual patient mutations and targeted therapy mechanisms to create the most effective match between patient and therapy.

Statement of Aims

Specific Aim 1. Create an Evidence-Based Framework for Assessing Drug-Target Interaction Data. We will create an evidence-based framework for evaluating existing drug-target interactions for FDA-approved cancer drugs. This framework will allow for a critical appraisal of the evidence supporting each drug-target interaction, such as literature references and experimental binding assays. This framework will enable us to assess drug-target interactions in a manner that reflects both confidence in supporting evidence and applicability of the interaction to drug research pipelines.

Specific Aim 2. Use Pathway-Based Analysis and Modeling to Characterize Somatic Mutation Intrinsic Perturbation and Drug-Targeted Extrinsic Perturbation, in Acute Myeloid Leukemia Patient Samples. We will map somatic mutation information for acute myeloid leukemia patient samples from the BeatAML Consortium to pathways and investigate the relationship between pathway-level dysregulation and drug response. Using probabilistic graphical modeling, we will model the impact of somatic mutations on pathways and also the impact of drug-target binding on pathways. We will use these models to better understand the relationship between patient-level pathway impact and drug response and make recommendations for developments for a unified model of mutation and drug pathway impact.

Background

Aim 1 Background

In recent years, it has become apparent that the one-drug-to-one-target model of paradigm of drug binding is much too simplified to accurately capture how many of our current pharmaceutical drugs work. In reality, drugs are chemical compounds capable of binding to many different protein targets or receptors in the human body, and each of these targets has the potential to elicit different changes at the cellular and overall physiological level [8,9]. This phenomenon is known as drug polypharmacology or drug promiscuity and it has gradually come to light as the rule rather than the exception for many of our current pharmaceutical drugs [10,11].

Drug polypharmacology has implications for almost every area of drug research and development [11,12]. Instead of viewing drug-target binding as a selective process whereby a drug binds to a particular target to elicit a particular cellular response (and ultimately a therapeutic effect), we are now beginning to think of drug-binding as a process that occurs across a range or profile of targets, each capable of causing different physiological effects once bound. One or more of these targets may be those that we have traditionally thought of as “primary” targets, or those that are responsible for eliciting a therapeutic response. However, non-primary targets (often referred to as “secondary” or “off-targets”) can also be responsible for additional physiological effects, such as those that we often think of as side effects.

For instance, the well-known hERG channel is an off-target that when bound can result in severe adverse cardiac events. The U.S. Food and Drug Administration (FDA) now requires that drug candidates be screened against hERG and show no binding activity as part of the FDA approval process [12,13]. We note here that off-targets do not necessarily have to result in adverse physiological effects – they can certainly also result in positive side effects or even contribute to the overall therapeutic efficacy of a drug [11,14]. Drug binding across multiple targets also has implications for drug resistance, drug-drug interactions, and combination therapies. Drug polypharmacology is also particularly relevant to research on novel or repurposed therapeutics, where information regarding additional drug targets may lead to an alternative or "off-label" use of a drug.

However, while we have begun to recognize that many of our small molecule drugs exhibit polypharmacology, the true extent of this phenomenon is poorly characterized. In particular, there is disagreement in the literature as to the number of targets we might expect (on average) for any given drug. Some estimates indicate that every drug binds to two or three targets, while some estimates indicate that the number is much closer to six or seven [6,15,16]. Such estimates appear to vary drastically depending on the data source used for drug-target interactions. For instance, Mestres et al. (2008) demonstrated that the degree of drug polypharmacology can increase drastically with the amount of drug-target interaction data that is included in a network [6]. Using only drug-target interaction data from the database DrugBank, the authors showed that

drugs have an average of 1.7 protein targets. However, including binding data from the literature into the network increased the average number of targets per drug to 2.7, and further including of predicted target interactions increased the average number of targets per drug to 6.3 [6]. More recent work on drug-disease networks using drug-target interaction data from the latest version of DrugBank has estimated an average of 3.5 targets per drug [17]. Despite the lack of agreement on the extent and range of polypharmacology exhibited by drugs, there is strong evidence that many of our drugs bind to more than one target, i.e. exhibit polypharmacology.

There has been limited work regarding the extent of drug promiscuity that occurs across target gene families. While some drugs target multiple proteins that are very similar (with regards to biological sequence or structure), drug promiscuity can also involve targets that are structurally dissimilar [18]. In a widely cited work, Paolini et al. (2006) found that while the majority of pharmaceutical drugs are active against protein targets originating from the same gene family, a quarter of all promiscuous compounds were found to be active at protein targets spanning different gene families. However, elsewhere it has been proposed that estimates of drug promiscuity across target families likely depend on the type of binding assay measured [16]. Hu et al. (2013) found that using IC50 binding assay measurements resulted in much higher estimates of drug promiscuity than those obtained using Ki binding assay measurements. However, the authors postulate that this may be a bias inherent in the screening pipeline; IC50 assays are often used much earlier in the pipeline while Ki assays

are often used to confirm more well-supported hits. Thus it may be expected that the larger number of compound-target pairs tested in IC50 screens would result in higher estimates for promiscuity compared to Ki measurements [16]. In a more focused analysis on Ki assay results only, the authors found that out of a total of 12,159 compounds with Ki measurements, approximately 36% of compounds were active against multiple targets from the same family while only 2.4% of compounds were active against targets spanning families [16].

Thus, while drug polypharmacology has been identified a crucial issue for drug research and development, it nonetheless remains a relatively underexplored area of research. Knowledge of additional targets for a drug (beyond a primary target) is potentially of high value for many types of drug research projects, especially those that concern drug side effects, drug resistance, or drug combinations. Given the potential impact this knowledge could have on decisions in such research projects it is critical that we begin to accommodate complete targeting profiles for drugs into our research pipelines. That is, in addition to binding information for so-called “primary” targets, we also need information on drugs for so-called “secondary” or “off” targets. The question now arises as to whether our current resources for drug-target interaction information accommodate this emerging polypharmacology paradigm of drug action.

In order to incorporate complete targeting profiles for drugs into research pipelines, there are several critical pieces of information that we must have. We use the term “complete” here to refer to profiles that reflect the current knowledge

base of drug-target information, dispersed across many resources. In reality, knowledge of a drug's complete targeting profile would require the infeasible task of testing drugs against all currently known human proteins. First, we require drug-target interaction information spanning multiple targets. Second, we must have some way of quantifying the strength of the interaction between a drug and each of its targets. Traditionally such an interaction is assessed using binding data, such as IC50 (half maximal inhibitory concentration) values or K_i (inhibition constant) values. IC50 values measure the concentration of a drug required to induce approximately a 50% inhibition of some measured response (such as enzymatic activity, activity of a competitive inhibitor, cellular growth, or cellular proliferation). Ki values also represent the concentration required for 50% inhibition. However Ki values are theoretically assay independent and more accurate than IC50 values [16]. These pieces of information constitute minimum requirements for assessing drug-target interactions across multiple targets.

In **Chapter 2**, we will explore two public repositories for high throughput drug screening data, PubChem Bioassay and ChemBank [19–21]. These resources contain the results from screens testing tens of thousands to hundreds of thousands of compounds for activity, usually against one biological target or receptor of interest [22,23]. Thus they are potentially a huge resource for mining for information about compound-target interactions to facilitate future drug development and repositioning projects. We will highlight case examples (one from PubChem and one from ChemBank) and discuss data quality and normalization issues. In particular, we address data accessibility issues (lack of

raw activity results and strong batch effects) and discuss the implications for re-use of the data.

By comparison, public resources for drug-target interaction data are generally one of two types: resources that are at least partially manually curated, or secondary resources that aggregate binding data from other sources. In general, they usually do not directly contain assay results from experiments. Instead, they contain citations or citations and reported binding assay values as mined from publications or the resources that directly house assay results. These resources vary in their coverage of compounds (approved drug, investigational drug candidates, or chemical-tool compounds) and targets, the supporting evidence supplied for interactions, and the level of curation provided. These resources will be reviewed extensively in **Chapter 3** and key issues with this data will be addressed.

In **Chapter 3**, we will also present our work for Aim 1 in aggregating cancer drug-target interaction information across public resources and introduce an evidence framework for the supporting evidence of each drug-target interaction. Collectively known as the Cancer Targetome, these aggregated drug-target relationships and evidence framework allow for prioritization and triage of drug-target interaction information in an empirical and evidence-driven manner. We explore how this allows for characterization of drug-target interactions and assessment of targets with stronger evidence, both in type of assay and strength of binding value. In particular, we will discuss the contribution of different types of evidence and highlight the complementarity of different assay binding types. For

the first time, this framework allows for characterization of uncertainty in the drug-target interaction space, a key aspect for rigorously assessing drug-target interaction data in research applications.

To translate the drug-target interaction and supporting evidence framework of the Cancer Targetome into a biological pathway context, we will map target interactions to biological pathways in Reactome. Reactome is a comprehensive, open-access database of biological pathways. Pathways are hierarchical collections of biological entities, with the bottom-most (or most detailed) level being biochemical reactions, and scaling up to broader parent groupings of pathways by function [24,25]. As part of this mapping, we will introduce and explore the concepts of 'light' and 'dark' pathways. Briefly, light pathways are those harboring a target of an approved cancer drug, while dark pathways are currently not targetable by any approved cancer drugs. This framework allows us to assess the scope or reach of current FDA-approved cancer drugs from a pathway perspective and will be critical in the pursuit of combinations of therapies to combat drug resistance. We will return to the concept of 'druggable' pathways in **Chapter 4** when we apply our probabilistic graphical modeling approach to model drug-impacted pathways for a select set of cancer drugs.

Aim 2 Background

Acute Myeloid Leukemia

Acute myeloid leukemia (AML) has seen little improvement in treatment and overall survival in the last forty years [26]. Notably, AML exhibits considerable heterogeneity across patients, where some cases are characterized by large chromosomal rearrangements but many cases possess mutations at the individual gene level [26,27]. The Cancer Genome Atlas (TCGA) found that AML is characterized by a smaller number of gene mutations compared to other cancers in adults, with an average of 13 gene mutations per patient and only 5 of these being found recurrent in the cohort [28]. Canonical AML mutations occur in FLT3 (internal tandem duplications (ITD) or tyrosine kinase domain (TKD) mutations), NPM1, CEPBA, DNMT3A, TET2, IDH1, IDH2, KRAS, NRAS, KRAS, TP53, and c-KIT [26,29]. The majority of AML cases are de novo malignancies, occurring in patients who were previously healthy. Alternative types of AML can occur in patients with other hematological diseases or as a result of prior treatment (therapy-related AML) [26]. Standard-of-care treatment for AML is the '7+3' regimen of cytarabine and anthracycline. While some success has been found in treating AML with targeted therapies, such as FLT3 inhibitors, (e.g. sorafenib, midostaurin, crenolanib, and quizartinib), STAT inhibitors, and IDH1/IDH2 inhibitors, for many patients, an initial response to a single-agent therapy will be followed soon after by cancer relapse [26].

In **Chapter 4**, our Aim 2 work on mutation impacted pathways and patient drug response on will leverage AML patient somatic mutation and drug screening

data from the BeatAML Consortium [30]. This consortium was a multi-year and multi-institution effort to collect AML patient samples and conduct extensive genetic, clinical, and drug screening data [30]. In Aim 2, we will use paired sample mutation data from 187 patient samples with de novo acute myeloid leukemia. A subset of these samples are accompanied by drug screening data, in which samples have been tested against a panel of drugs [31]. We note that the combination of both mutational aberrational information and functional drug screening data for this cohort of patients is critical to enable a more in-depth assessment of the relationship between genetic alterations and overall patient drug response [30–32].

Previously, in **Chapter 3**, we cast drug-target interactions into a pathway perspective to provide biological context. Supported by the evidence framework of the Cancer Targetome, this enabled us to assess whether biological pathways are potentially 'druggable'. Translating drug-target mechanistic information to the pathway perspective is critical for moving cancer targeted therapies into combination treatment; many drug resistance mechanisms in cancer are mediated at the pathway level [1,33]. Complementary to our work in **Chapter 3** on the Cancer Targetome characterizing drug-target interactions, targeted pathways, and supporting levels of evidence, we also pursue a pathway perspective for individual patient mutation aberrations. In **Chapter 4**, we map mutations from de novo AML patient samples to biological pathways and investigate the relationship between pathway-level dysregulation and drug response. Given the heterogeneity, especially many mutations that are low

frequency, and the need for improved therapies, there is a need for improved biomarkers to guide treatment, and understanding to inform combination strategies.

Pathway-Based Analysis and Modeling

For mapping to pathways, we will use the Reactome Pathway database [24,25]. By mapping mutations to these pathways, we can begin to understand the effect of mutations in terms of greater, overarching biological functions. In particular, we can aggregate multiple mutations to the same pathway and get an idea of pathway mutational load, and group samples together based on shared dysregulated pathway signaling, even if the mutated genes are not shared amongst samples. Additionally, this opens the door for assessment at the level of patterns of share dysregulated pathways, rather than patterns (i.e. clustering) of shared mutated genes.

Traditionally, investigation of mutation and drug response has been restricted to the relationship between mutational or chromosomal aberrations as biomarkers and drug response [34]. In work by Iorio et al. (2016), the relationship between pathway-level aberrations and drug response was investigated for cancer cell lines and a wide range of functional events (CFEs) [34]. In **Chapter 4**, we use a similar ANOVA assessment to investigate pathway-level dysregulation and drug sensitivity or resistance. Briefly, when samples with pathway dysregulation have a significantly different mean AUC than samples without dysregulation, there is a relationship between the pathway response.

However, to interrogate the relationship between pathway dysregulation and drug response, we require a modeling framework that can capture the impact of a mutation on a pathway and thus model differing effects of mutations on a pathway based on the gene which is affected. In this work, we use functional impact score to represent the impact of a gene mutation, which include PolyPhen, SIFT, and Variant Effect Predictor [35,36]. These scores predict the impact of particular SNP on a gene. They each provide a continuous measure between 0 and 1 that reflects how deleterious the mutation is predicted to be. For PolyPhen, higher numbers are more deleterious (0 being no effect and 1 being deleterious) while for SIFT, lower numbers are more deleterious (0 being deleterious and 1 being no effect).

In **Chapter 4**, we will use an adaptation of factor graph approach implemented originally in PARADIGM [37], and assess the impact of patient-level mutations on pathway signaling. To provide background, probabilistic graphical models (PGMs) are a broad class of models that utilize core concepts from graph theory to represent complex probability distributions [38]. One type of PGM is the factor graph, which is a bipartite graph containing two types of nodes: variable nodes and factor nodes (Figure 1) [39]. Factor graphs are a generic way of representing probabilistic graphical models, as both directed and undirected models can be formulated as a factor graph. One advantage to using a factor graph representation is that it makes the factors or relationships between variables very apparent [40]. Factor graphs lend themselves well to describing biological signaling networks, as biological entities (such as genes or proteins)

can be described using variable nodes while relationships between such entities can be captured using factor nodes. Additionally, factor graphs can also accommodate cycles, which is useful for applications to biological networks, as will be discussed below [37–39,41]. This is in contrast to Bayesian networks, which cannot accommodate cycles in the graph structure.

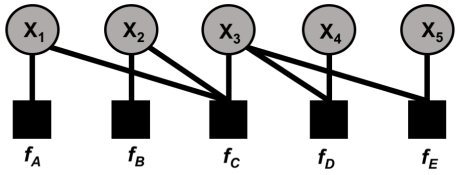


Figure 1. Example Factor Graph. A factor graph is a bipartite network consisting of variable nodes (circles) and factor nodes (squares). Factor nodes describe the relationships between variable nodes. Figure adapted from Kschischang, et al. (2001) [39].

Factor graphs have successfully been used to integrate different types of genomic data for prediction of pathway activity in individual patients. In their factor graph approach known as PARADIGM, Vaske et al. (2010) used variable nodes to represent biological entities such as proteins, RNA, and DNA . Factor nodes were used to describe the relationships (or influence) between these entities. In their work, the authors made use of additional “observed” nodes to attach genomic data such as mRNA expression data and copy number variation. Using message passing algorithms, the influence of observed data was propagated to additional hidden nodes throughout the graph representing unseen or unmeasured biological entities in each pathway. The authors applied their approach to the analysis of glioblastoma and breast cancer genomic data and showed that significantly up or down-regulated entities could be used to cluster

similarly impacted samples and thereby distinguish between multiple subtypes of glioblastoma as well as the four canonical subtypes of breast cancer [37].

We will use a modified version of the PARADIGM method to model mutational impact on pathway signaling. Similar to PARADIGM's pathway impact score, we use the average log odds of entity belief states to measure the impact of a mutation on a pathway. We will investigate the impact of different AML de novo mutations on overall pathway impact and the relationship with drug sensitivity or resistance using the accompanying drug panel results.

In parallel, we will also investigate the impact of drug-target binding across targets within a pathway. For those drugs that are in both the Cancer Targetome and the BeatAML drug screening panel, we will retrieve high confidence (Level III binding assay supported) target interactions. The Cancer Targetome data will be used to assign the impact of drug-target binding at relevant nodes in the pathway, which will then be propagated to neighboring nodes via message passing. This will allow us to investigate the impact of drugs on pathways. Using these two independent modeling approaches (mutation impact and drug impact), we will explore how to unite these pieces of information to better understand patient level drug response. We will also make recommendations for future modeling iterations towards an eventual unified model of mutation and drug response.

The probabilistic graphical modeling framework offers several advantages to other modeling approaches for tackling such a problem. Using a factor graph approach for such a task will allow us to probabilistically model both mutation

impact and drug-target binding impact across multiple nodes in the graph. Using factor nodes, we can encode the relationships between mutation and gene or drug and target in a probabilistic manner to (respectively) reflects the functional impact of mutations or higher confidence drug-target interactions. In this way, we can capture a more accurate picture of drug binding across multiple targets and thereby represent drug polypharmacology in our computational model. Secondly, by representing biological entities (such as proteins that neighbor drug targets) using a factor graph approach, we can incorporate topological information into our model that provides value over solely using a binary indication of mapped versus unmapped genes. Lastly, the factor graph approach allows us to create a patient-specific model of impacted pathways for each patient sample. Such a model can then be used for a personalized approach to understanding somatic mutational profiles, drug binding, and overall drug response in patients.

CHAPTER 2. CHALLENGES IN SECONDARY ANALYSIS OF HIGH THROUGHPUT SCREENING DATA

Abstract

Repurposing an existing drug for an alternative use is not only a cost effective method of development, but also a faster process due to the drug's previous clinical testing and established pharmacokinetic profiles. A potentially rich resource for computational drug repositioning approaches is publicly available high throughput screening data, available in databases such as PubChem Bioassay and ChemBank. We examine statistical and computational considerations for secondary analysis of publicly available high throughput screening (HTS) data with respect to metadata, data quality, and completeness. We discuss developing methods and best practices that can help to ameliorate these issues.

Introduction

Despite increasing investment in drug research and development in recent years, the pharmaceutical industry has seen limited results in the form of novel marketable drugs[43]. Attention has recently turned to drug repositioning, or finding new uses for already developed drugs. Drug repurposing is particularly attractive due to its simplified timeline; while the traditional drug discovery process can take between ten and seventeen years to bring a drug to production, repurposing a drug can take as little as three to twelve years depending on the

drug's previously established chemical properties.[44] In several cases, repurposing has provided enormous benefit to patients with previously limited treatment options, such as the repositioning of thalidomide to treat multiple myeloma, or bromocriptine for Type 2 diabetes. Other well-known repositioning successes include Wellbutrin as Zyban for a smoking cessation aid, Minoxidil for hair loss, and Viagra (sildenafil) for erectile dysfunction.[43–45]

A potentially valuable resource for drug repositioning efforts is publicly available high throughput screening (HTS) data.[22] A primary strategy for drug discovery, the automated high throughput screening process allows for the activity of hundreds of thousands of chemical compounds to be tested simultaneously.[23] Compounds are screened against a particular target compound, typically a receptor or enzyme implicated in a disease, and are declared active if their results differ from the majority of the test compounds. However, it is well known that there are several common sources of variation within high throughput screens, both technological, such as batch, plate, and positional (row or column) effects, and biological, such as the presence of non-selective binders, which can result in false positives and negative bioactivity results.[22,23,46–48] These problems are can be resolved through pre-processing, standardization and normalization methods, which include the z-score, percent inhibition, and median-based methods among others.[23,49,50]

Results from high throughput screening projects, primarily from academic institutions, are often made available through public databases such as NCBI PubChem Bioassay and ChemBank. [22] The PubChem Bioassay database

contains the results of high throughput screens for the biological activities of molecules cross-listed in PubChem Substance and Compound.[20,51] Each PubChem assay has a unique assay identifier (AID). Assay data sets usually contain compound information, accompanying readout (for example, recorded fluorescence emission), activity score, activity outcome, and the mean values of minimum and maximum control wells for each plate in the assay. Activity scores and outcome are defined in the assay description, which typically explains the threshold used to declare a particular compound active.[20] The actual raw HTS data is not included in PubChem, however, and therefore there is no information on batch, plate, or within-plate position for each screened compound.

The Broad ChemBank database also contains the results of small molecule screens, as well as the raw datasets from screening centers. Each assay in ChemBank therefore contains not only compound information and accompanying readout, but also batch, plate, row, and column annotation for each screened compound. Additionally, each assay is conducted twice, so assay datasets contain replicate fluorescence readings.[21]

Given the common sources of variation known to affect high throughput screening data, it is crucial that the quality of a particular bioassay is evaluated before its results are used in further research efforts. For instance, researchers interested in using bioactivity information from databases such as PubChem and ChemBank for computational repositioning methods must first be convinced of the reliability of the screens in these databases.[47] Issues in assay quality can result in false positive or false negative bioactivity results, affecting which

compounds are considered for potential repositioning. Here, datasets from both PubChem and ChemBank are evaluated to quantify the advantages and limitations of each repository as well as to investigate common sources of variation such as batch, plate, and positional effects. This analysis is representative of a typical investigation of HTS data that would be conducted before utilizing this data in further computational repurposing efforts. Overall, the problems encountered here illustrate some of the key barriers to effective secondary use of publicly available high throughput screening data in order to realize the full potential of these datasets.

Methods

In this study, exploratory analysis was conducted on representative bioassay datasets from PubChem and ChemBank to examine data completeness, particularly in the context of data pre-processing and addressing technical sources of variation. Additional data was obtained directly from the original screeners of the highlighted PubChem study to complete the exploratory data analysis and allow for comparable assessments to the ChemBank study.

PubChem Example

The PubChem CDC25B (AID 368) dataset contains the results from approximately 65,222 compounds and controls of a primary screen against the target CDC25B. CDC25 is a protein tyrosine phosphatase cell cycle regulator, and of three existing isoforms, two are oncogenic and have been found to be

overexpressed in a variety of human tumors. The goal of this screen was to find potential inhibitors for the CDC25B isoform.[52] The CDC25B dataset contained the following attributes: PubChem Substance ID, PubChem Compound ID, activity score, activity outcome, database URL, comment field, raw fluorescence intensity, calculated percent inhibition, mean of minimum control well signals (by plate), mean of maximum control well signals (by plate), calculated z-factor, and assay run date. Exploratory data analysis was conducted to evaluate the overall distribution of fluorescence intensity, percent inhibition, minimum control well means, maximum control well means, and calculated z'-factors. However, no further analysis could be performed for this dataset in the form available from the PubChem database, given the lack of plate level data such as batch number, plate number, and row and column information for each well.

Full PubChem Example

The full CDC25B dataset, including plate-level annotation, was obtained directly from the PMLSC screening center and contained results from approximately 83,711 compounds and controls across 218 384-well microtiter plates. In addition to PubChem Compound ID, raw fluorescence emission, calculated percent inhibition, mean minimum signal, mean maximum signal, calculated z-factor, and run date, this dataset also included assay batch, plate ID, row, column, well number, and well annotation. This information enabled further exploratory data analysis such as evaluation of fluorescence intensity distribution by well type and across plates and batches. Heatmaps were created for

individual plates to check for positional effects. The mean signal to background ratio and percent coefficients of variation for the minimum and maximum control wells were also calculated. Based on the exploratory data analysis, percent inhibition was chosen as the most appropriate normalization method, which was also the method chosen by the original screeners when processing the dataset. [23,52]

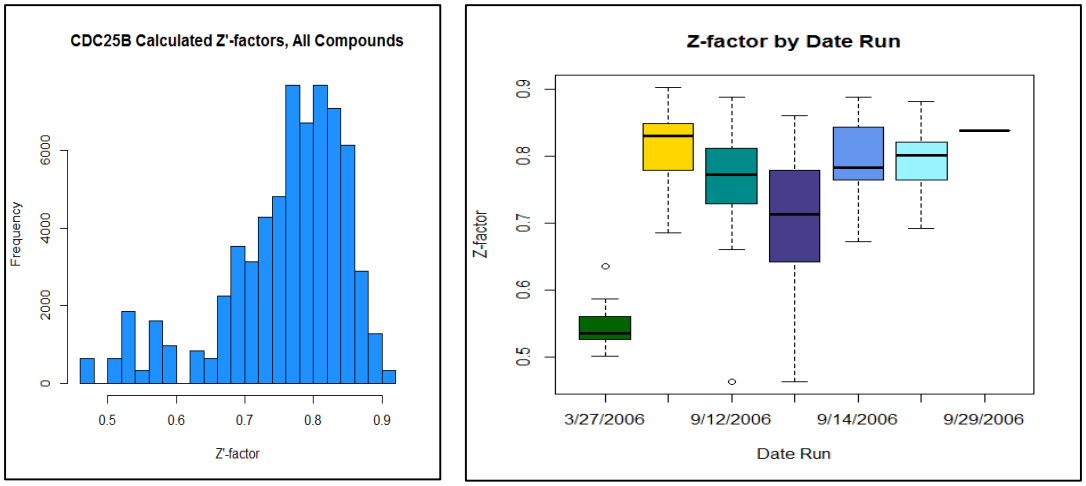
ChemBank Example

The ChemBank BRAF dataset contains the results from approximately 41,088 compounds and controls of a primary screen to find an inhibitor of the $BRAF^{V600E}$ mutant. The BRAF gene plays an important role in the mitogen-activated signaling pathway and in particular, the $BRAF^{V600E}$ mutation has been implicated in melanoma, papillary thyroid carcinoma, and colorectal cancer.[53] The BRAF dataset is composed of seven different assays, each with two replicates. Given limited assay description and annotation provided, each of the seven assays was evaluated separately. First, correlation of raw fluorescence intensity between the two replicates was assessed for each of the seven assays, and if present, any outlying data points were investigated at the plate level. Next, exploratory data analysis was conducted for each assay to assess the overall distribution of fluorescence intensity, background-subtracted values, and calculated z-score. This analysis included histograms, boxplots, and quantile-quantile plots for individual replicates and statistical indices of the combined data, as appropriate.

Results

PubChem Example

Overall, the distribution of fluorescence intensity across all compounds in the CDC25B dataset is strongly skewed right, while the distribution of percent inhibition across all compounds is strongly skewed to the left. The distribution for the range between the mean minimum and mean maximum control wells is slightly skewed bimodal (See Supplementary Material S1) The distribution of z'-factors across all compounds is fairly skewed to the left and appears to be slightly bimodal. Boxplots of z'-factor by run date reveal strong variation by date (Figure 1).



(A) (B)

Figure 1. Distribution of Z'-factors for PubChem CDC25B dataset. (A) Histogram depicting distribution of calculated z'-factors. (B) Boxplots by run date for calculated z'-factors.

It is noted that the compounds run in March 2006 have much lower z'-factors than the remaining compounds, run in August and September 2006. Additionally, the compounds run on September 13th, 2006 exhibit a much wider

range of z'-factors than compounds run on any other dates, while compounds run on September 29th, 2006 exhibit a much narrower range. Given that the z'-factor is a commonly used measure of assay quality, plates with a such divergent z'-factors should be examined for possible errors and batch effects. Here, however, further investigation into the sources of this variation could not be conducted due to the lack of plate level annotation available through the PubChem Bioassay database. If the metadata had been available, it would then be possible to attempt to correct for batch and technical sources of variation.

Full PubChem CDC25B example

Histograms of fluorescence intensity by well type (compound, 50% inhibition, minimum, and maximum) for the full CDC25B dataset show that the distribution of fluorescence intensity across all wells is somewhat normal with a strong peak. The distributions of fluorescence intensities for compound wells and maximum control wells are slightly skewed right, while the distributions of fluorescence intensities for minimum and 50% inhibition control wells are more strongly skewed to the right (See Supplementary Material S2 Fig 1 and 2). Fluorescence intensity appears to vary widely by both batch and run date as well as by plate within respective batches (See Supplementary Material S2 Fig 3-8). No apparent positional effects were detected by visual examination of heatmaps for each of the 218 plates in the dataset.

Following a recently proposed decision process for HTS data processing, percent inhibition was chosen as the most appropriate method of normalization,

due to the fairly normal distribution of fluorescence intensity, lack of row and column biases, a mean signal to background ratio greater than 3.5, and percent coefficients of variation for both the minimum and maximum controls wells less than 20% [23] (See Supplementary Material S2 Table 1). This appeared to successfully normalize the data by batch, date, and across plates within each batch and reproduced the original analysis (See Supplementary Material S2 Fig 9-16). It is important to note that it would not be possible to successfully evaluate this data set with regard to pre-processing and normalization without the plate level annotation.

ChemBank Example

There was a large range with regard to correlation of fluorescence intensity between replicates: 0.436-0.910 (Table 1). Scatterplots further illustrate the high variability among some replicates (Figure 2). This allows easy identification of signal discrepancies. For example, the bottom of the scatterplot for assay 1110.0002, it is easy to detect a set of mock treatment wells (in red) where signal was present in replicate A, but not in replicate B. Similarly, the upper left-hand corner of the scatterplot for assay 1110.0003 shows a replicate specific cluster of compound treatment wells. The outlying data points in assay 1110.0002 were found to be confined to one plate, 1110.0002.Base. The outlying data points in assay 1110.003 were similarly located on a single plate, 1110.0003.2340.

Table 1. Correlation Coefficients for Fluorescence Intensity Replicate A vs Fluorescence Intensity Replicate B, by Assay, ChemBank BRAF dataset.

Assay Number	1110.0001	1110.0002	1110.0003	1110.0004	1110.0005	1110.0006	1110.0007
Correlation	0.436	0.536	0.906	0.910	0.902	0.869	0.846

Examination of the well-plate layout for 1110.0002 allowed identification of an obvious positional effect in the upper six rows of the plate (Figure 3). Similarly for 1110.0003, the corresponding well-plate layout illustrated a clear positional effect along the bottom two rows of the plate.

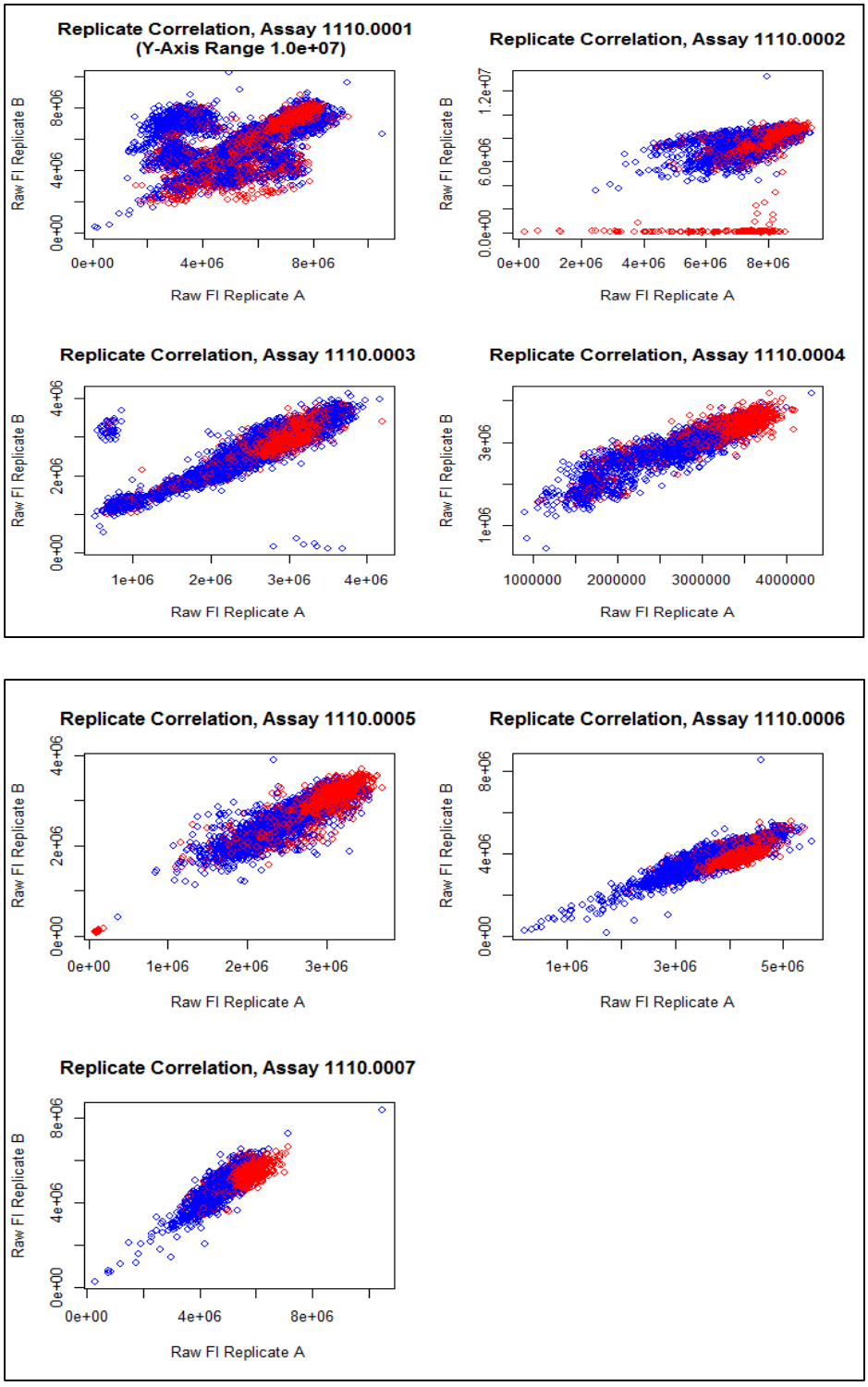


Figure 2. Scatterplots for Correlation of Fluorescence Intensity Between Replicates A and B. Correlation between replicates of Assay 1110.0001- 1110.0007. Blue indicates compound-treatment wells, red indicates control wells.

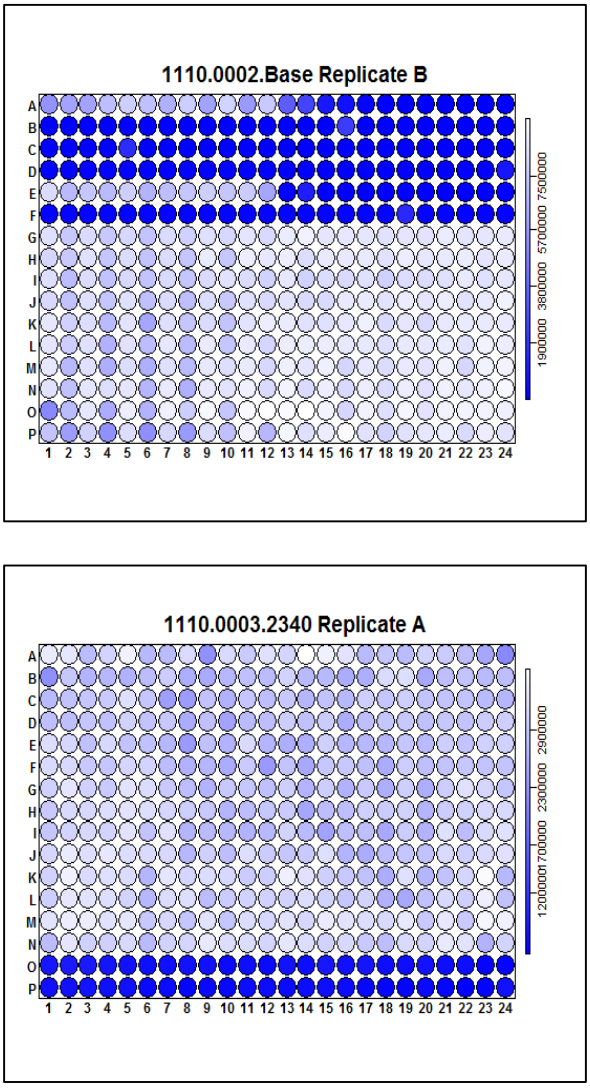


Figure 3. Well Plate Layouts for Selected BRAF Assays. (Left) Replicate B of Base Plate for Assay 1110.0003. (Right) Replicate A of Plate 2340 for Assay 1110.0003. Darker wells indicate decreased fluorescence.

Overall, each of the seven assays in the BRAF dataset showed fairly different distributions for fluorescence intensity, background-subtracted values, and calculated z-scores (See Supplementary Material S3), further reiterating the role of exploratory data analysis to examine model assumptions prior to downstream analysis.

Boxplots of the fluorescence intensity by plate were then examined. It was noted that the signal varies considerably across plates, both within and across each of the seven assays. (Replicate A shown in Figure 4). Beginning with assay 1110.0003 in replicate A, it is apparent that within each assay, fluorescence intensities steadily increase with each successive plate that is run before dropping down at the beginning of the next assay. In the absence of timestamps for each plate, it was assumed that increasing plate numbers indicate passage of time. However, without that appropriate metadata, it is not possible to determine the actual source of variation, again limiting the ability to correctly model batch or temporal effects.

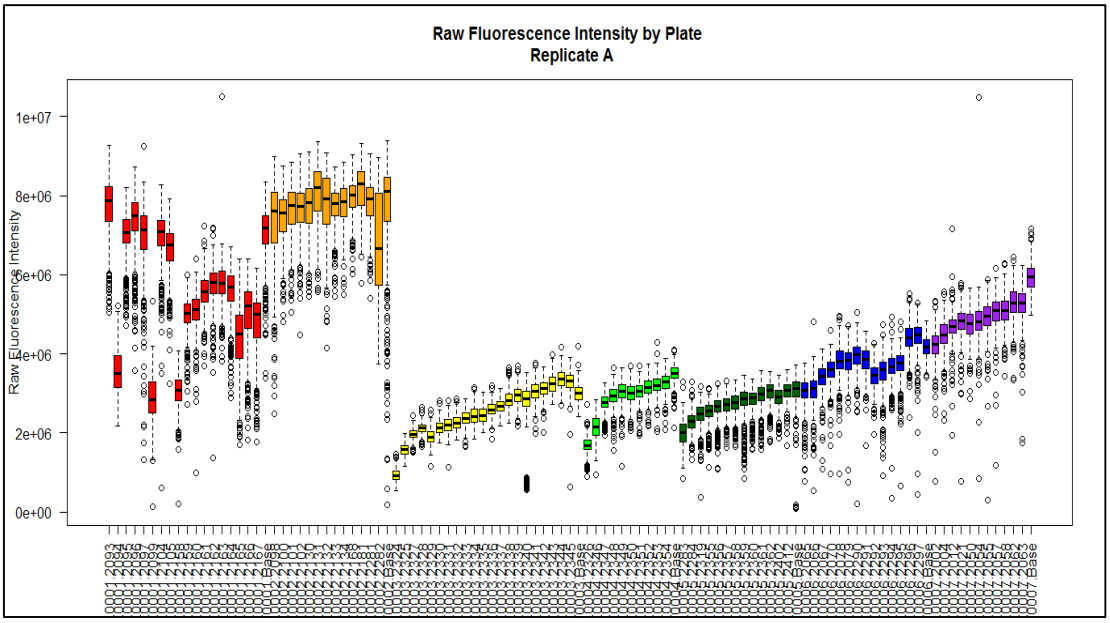


Figure 4. Raw Fluorescence Intensity by Plate, Across All Assays, Replicate A, ChemBank BRAF dataset. Each boxplot depicts the fluorescence values of the wells of one plate. Colors indicate assay "Name", which may or may not be synonymous with batch.

Discussion

Both repositories examined provide excellent opportunities for secondary analysis of public HTS data. However, we have noted several issues that need to be addressed in order to realize their full potential. Most notably, the lack of actual raw data, and therefore plate level annotation for bioassays in PubChem BioAssay prevents rigorous analysis of data quality. As illustrated above, initial exploratory analysis of the limited CDC25B dataset (as obtained from PubChem) reveals potential quality issues, such as variation by run date. These issues cannot be fully investigated, however, without knowledge of batch and plate numbers and row and column positioning for each tested compound. The complete CDC25B dataset, obtained directly from the screeners, allowed for more in-depth investigation of sources of variation, which in turn allowed for more appropriate pre-processing and normalization recommendations to be made. It would not have been possible to evaluate the dataset solely from the data and annotation made available through the PubChem database.

Another issue for researchers seeking to extract assay information from PubChem is the lack of description for the particular readouts used in assays. While the PubChem assay discussed in this paper provided a full description of the fluorescence emission readout, many assays do not necessarily include this level of information. It is also important to note that the issues discussed here are likely extensible to other databases, such as ChEMBL, which contain bioactivity information from selected PubChem Bioassays.[54]

The ChemBank database is currently the only publicly available bioassay database that requires the inclusion of plate level annotation in their datasets. While this information is crucial for secondary analysis, the value of the datasets in ChemBank is negatively impacted by the lack of assay annotation and description. For instance, the BRAF dataset was composed of seven different assays, but it was unclear how these differed from one another, if at all. From the assay descriptions, it appeared that only the first assay differs in its biological components, but there was no additional information as to why the remaining six assays were conducted separately. Additionally, while we might expect strong correlation between replicates for each assay, several assays exhibited exceptionally poor correlation, which casts doubt on the overall quality of the screening data. Furthermore, the lack of date or timestamps for the ChemBank data makes it impossible to confirm temporal batch effects, limiting one to data visualization by plate, with an assumption that plate order corresponds with time, as done in Figure 4.

Correspondence with PubChem confirmed that PubChem Bioassay does not require plate level annotation in uploaded datasets to the BioAssay database. It is also noted that there is no way to query for which, if any, datasets include this level of annotation (Personal communication with PubChem). ChemBank also confirmed that the “AssayName” field is used by depositors in different ways: it can be used for biologically different assays or batches of similar assays. Currently, there is no method of querying for datasets to identify those for which particular descriptive information/metadata are included (Personal

Communication with ChemBank). These issues affect not only the general usability of the databases, but in particular hinder a larger-scale systematic quality analysis of HTS assays. The analysis presented here was restricted to one assay from each database primarily due to difficulties in accessibility and poor annotation.

Issues such as these in turn stymie the usage of high throughput screening data in further research efforts such as computational repositioning efforts requiring bioactivity information. There is the potential for improved data standards and development of best practices for data dissemination to improve the quality and reusability of the data in these repositories. At a minimum, the inclusion of metadata such as plate and well-level annotation will enable a more thorough secondary analysis of HTS data. Additional oversight to ensure descriptor fields for assays are completed may also encourage assay re-use. With respect to cost-benefit analysis, the potential for re-use of the data via secondary analysis far outweighs any costs due to additional data standards or metadata requirements, as the metadata has already been generated. Further impact in time/resources for depositing additional metadata can easily be mitigated by automation. One example of methods to facilitate the reporting of this metadata is a recently proposed method to first extract workflows directly from screening data in PubChem and then use the workflows to organize data within screening projects. [55]

Addressing these issues in the research community and in the requirements for submission to these repositories could improve the re-use of

these data sets. A PubMed search for “PubChem” results in only 263 articles, and the more specific “PubChem BioAssay” pulls up only 51 articles. Querying for “ChemBank” returns even fewer articles, with only 17 results. For perspective, searching “GEO” brings up approximately 8480 results for Gene Expression Omnibus. While both PubChem BioAssay and ChemBank are fairly young databases and more expansive mining efforts using their datasets may still be yet to come, the annotation and data quality issues in both databases cannot be ignored as a potential barrier to dissemination. Expanded datasets as well as more rigorous quality standards are necessary to ensure the public data is truly accessible and re-usable.

Author Contributions

This chapter originally published as ‘Challenges in Secondary Analysis of High Throughput Screening Data’, by Aurora S. Blucher and Dr. Shannon K. McWeeney for the Pacific Symposium on Biocomputing, PMCID PMC3976302.

CHAPTER 3. AIM 1: EVIDENCE-BASED PRECISION ONCOLOGY WITH THE CANCER TARGETOME

Abstract

A core tenet of precision oncology is the rational choice of drugs to interact with patient-specific biological targets of interest, but it is currently difficult for researchers to obtain consistent and well-supported target information for pharmaceutical drugs. We review current drug–target interaction resources and critically assess how supporting evidence is handled. We introduce the concept of a unified Cancer Targetome to aggregate drug–target interactions in an evidence-based framework. We discuss current unmet needs and the implications for evidence-based clinical omics. The focus of this review is precision oncology but the discussion is highly relevant to targeted therapies in any area.

Introduction

Precision Oncology Requires Rigorous Drug–Target Information

The advent of **precision oncology** (see Glossary) is often hallmarked with the development of the **targeted therapy** imatinib to treat *BCR–ABL1*-positive chronic myeloid leukemia (CML) [3]. Over time, the term precision oncology has evolved to include the use of genetic **biomarkers** to guide treatment selection as well as to refer to the emerging paradigm of treating cancer in a mutation-centric manner over a histology-centric manner [56–58]. However, the promise of precision oncology has been dimmed with the realization that only a small number of genetic variants in cancer are currently

actionable with approved drugs [56,59]. Much of the work focused on expanding what is considered to be actionable in cancer genomics has focused on the characterization of cancer-associated and driver genes and prioritization of these candidates for therapeutic intervention [60–63]. However, any endeavor to expand the actionable space and thereby expand patient treatment options requires that we have a working knowledge of the interactions between drugs and their biological **targets**.

As illustrated in Figure 1 (Key Figure), **drug–target interactions** play an integral role in many different precision oncology applications. Clinical trials for cancer therapies are at the forefront of design and methodology development [64]. Newly emerging trial designs include umbrella trials, in which patients with the same type of cancer are assigned to different treatment arms according to key genetic variants [65,66], and basket trials, in which patients are assigned to treatment based on genetic variant but irrespective of cancer type [67,68]. Both of these trial designs rely on drug–target interaction information. Computational and predictive modeling approaches to predict drug response or anticipate adverse drug reactions require both primary and secondary target information for a complete picture [11,14,69–71]. Drug repurposing, or finding alternative uses for existing drugs, often makes use of secondary or so-called ‘off-target’ binding, where a drug binds to a target other than the one it was designed for [43,72]. Lastly, designing **combinatorial drug treatment** for a patient based on multiple genetic variants requires knowledge of drugs interacting with targets affected by each of those genetic variants [8]. Each of these examples requires knowledge of

the biological targets that a drug may potentially interact with, but the specific context of a precision medicine application will dictate more or less rigorous requirements for the strength of supporting evidence for a drug–target interaction. Because this information can directly impact drug or target prioritization decisions and ultimately affect treatment options for patients, it is imperative that researchers have access to drug–target interaction information with clear literature and experimental evidence.

Historically, the scope of the approved drug–target interaction space has been difficult to pin down precisely. Since the first characterization of the **druggable genome** [73,74] nearly 20 years ago, estimates of the number of biological targets for approved drugs have varied with both the definition of the target and the scope of data collection [7,75–79]. The realization that many currently approved drugs display **polypharmacological** or non-selective behavior [15] has added another layer of complexity to characterization of the drug–target interaction space.

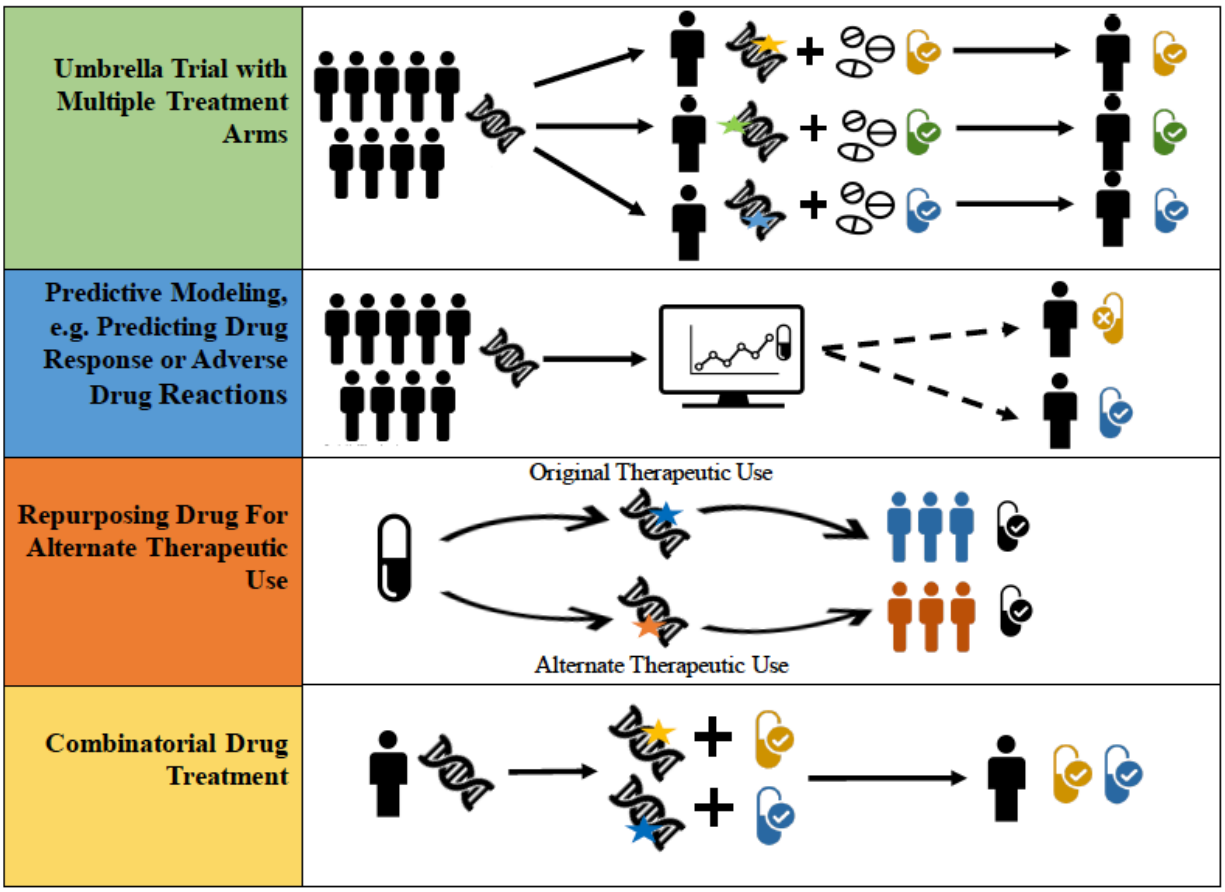


Figure 1. Key Figure. Precision Oncology Applications Rely on Drug–Target Interaction Information. (A) Umbrella clinical trial with multiple treatment arms. Patients are assigned to different treatment arms in a clinical trial according to their genetic alterations. Drug treatments in each arm are determined according to interactions between drugs and priority genomic variants in tumors. (B) Predictive modeling (e.g., predicting drug response or adverse drug reactions). Patient genomic data is used with *in silico* drug treatment simulation to predict which patients will respond beneficially (or adversely) to particular drug treatments. (C) Repurposing a drug for alternative therapeutic use. Drug binding information at additional or 'secondary' targets can be used to repurpose a drug for a secondary therapeutic indication. (D) Combinatorial drug treatment. A combination of drugs is chosen for a patient using his or her genetic variant information and prioritization of variant-related targets according to known interactions with existing drugs.

Current public informatics resources for drug–target interaction information do not reflect a strong and consistent understanding of cancer drug binding across multiple targets. While the broader drug–target interaction space in the public domain faces the limitation of sparsity (only so many drug–target interaction pairs have been tested), there is a plethora of drug–target interaction

and bioactivity information that is available but currently underutilized by the precision oncology research community. Hurdles to using this information include the need for aggregation across resources, unclear reference lineage, and differing types of supporting evidence. These challenges pose significant barriers to researchers looking to critically assess existing target annotations for a particular drug and this task quickly becomes intractable as the number of drugs of interest increases.

Current Resources for Drug–Target Interactions

Here we briefly review resources and databases for drug–target interaction information (Table 1). Current resources for drug–target interaction data can be broadly categorized into two types: drug centric and bioactivity centric. Resources such as DrugBank, the Therapeutic Targets Database, and KEGG Drug contain drug–target annotations supported by literature evidence and are subject to manual curation but currently do not incorporate experimental binding activity evidence [80–84]. Other resources, such as the International Union of Basic and Clinical Pharmacology/British Pharmacological Society (IUPHAR/BPS) Guide to PHARMACOLOGY, include manually curated experimental binding activities with drug–target annotations [85]. The Drug Gene Interaction Database aggregates drug–target annotations across multiple sources, allowing the user to see the parent sources and total literature reference count per drug–target interaction, but does not currently include binding activity evidence [86]. Other resources that provide experimental binding evidence for

target annotations for approved drugs and/or clinical trial drugs include DrugCentral, Pharos, SuperTarget, and STITCH [63,87–89]. Open Targets is a recently released academic–industry collaborative resource that includes drug–target interaction information, but is currently more focused on enabling target validation efforts [90]. While all of these resources allow multiple targets per drug, differing standards for target inclusion can result in discrepant target annotations across resources [16].

Bioactivity databases such as ChEMBL, BindingDB, and PubChem Bioassay aggregate chemical compound experimental binding activity information through manual extraction or text mining from the literature and other bioactivity databases [91–95]. These resources offer differing coverage with respect to compounds, targets, and interactions due to differences in data scope, collection methods, and curation [96–98].

While bioactivity databases offer a wealth of potential compound–target information due to large-scale collection of high-throughput screening results [51], they do not directly provide drug–target interaction annotation and it is therefore up to the user to determine an appropriate binding activity threshold when collecting and assessing experimental binding activity data. This presents its own challenge, as the choice of an appropriate activity threshold depends on the biological context of the problem. To determine the bioactivity of compounds, the threshold of 10 000 nM (10 μM) is often used, but a much stricter threshold of 100 nM or under is more appropriate when requiring interactions to be relevant to drug binding [15,99]. Paolini et al. (2006) required the best activity across assay

types (IC_{50}, EC_{50}, K_i, and K_d) to be less than 10 000 nM in their analysis of global pharmacological space [15]. Similarly, Koutsakas et al. (2013) used a bioactivity threshold of 10 000 nM to obtain a balance between chemical space coverage and the inclusion of weakly active compounds [100]. This bioactivity threshold has been used by others in target prediction methods [101,102] or the analysis of drug–target annotations [103], while other groups have used more conservative bioactivity thresholds across assay types (1000 nM) [104] in target prediction or used only a single assay type (K_d < 3000 nM) in calculating selectivity measures [105]. Finan et al. (2017) used a threshold of 100 nM on ChEMBL bioactivity data (across all assay types) to supplement target annotation found in company pipelines and the literature for approved and clinical trial drugs [106]. The Pharos platform, which presents data from the Target Central Resource Database (and uses a Target Development Level scheme to group targets based on level of study and association with small-molecule bioactivity), uses bioactivity thresholds based on target family-specific cutoffs [63].

Table 1. Databases for Drug-Target Interactions[a]

Database (URL)	*Description*	*Year est.*	*Version (release date)*	*Drug/ compound*	*Target inclusion criteria*	*Supporting evidence*	*License*	*Refs*
DrugBank https://www.drugbank.ca/	Drug database including FDA-approved small molecules, FDA-approved biotech drugs, nutraceuticals, and experimental drugs	2006	5.0.7 (07/06/17)	A, Cl, N, W	Manual	PubMed	Public for non-commercial use[b]	[79,80]
Therapeutic Targets Database	Includes therapeutic protein and nucleic acid targets,	2002	4.3.02 (08/25/11)	A, Cl	Manual	PubMed	No license indicated[c]	[81]

http://bidd.nus.edu.sg/BIDD-Databases/TTD	targeted disease condition, pathway information, and drug interactions		Site updated 09/10/15					
IUPHAR/BPS[d] *http://www.guidetopharmacology.org/*	Expert-curated resource for pharmacological, chemical, genetic, functional, and pathophysiological data on targets of approved and experimental drugs	2013	2017.4 (05/23/17)	A, CI	Manual	PubMed, US patent, Binding$_{selected}$	Database ODbl[e], contents are CC-BY SA 3.0[f]	[84]
DGIdb[g] *http://dgidb.genome.wustl.edu/*	Integrated from 13 primary sources for 'druggable genome', genes with known drug interactions or genes that are potentially druggable	2013	3.0 (06/30/17)	A, CI	Manual	External DB, PubMed	GNU General Public License V3	[85]
Open Targets https://www.targetvalidation.org/	Informatics platform for target validation with extensive evidence associating targets and diseases	2016	3.2.0 (07/27/17)	A, CI	All	External DB (ChEMBL)	Public	[89]
Pharos https://pharos.nih.gov/idg	Contains druggable human protein targets as part of the NIH Illuminating the Druggable Genome project; also includes target information for full human proteome	2016	Uses 4.6.2 of Target Central Resource	A, CI, L	Threshold	Binding$_{selected}$	CC BY-SA 4.0[h] International	[62]
DrugCentral http://drugcentral.org/	Drug compendium of structure, bioactivity, regulator, pharmacological action, and indication information for active pharmaceutical ingredients approved by FDA, EMA, and PMDA	2016	9.4 (04/25/17)	A[i]	Threshold	Binding$_{selected}$	CC BY-SA 4.0 International	[86]

SuperTarget http://insilico.charite.de/supertarget/	Web-based warehouse that integrates drug-related information with indication, adverse effect, metabolism, and gene ontology terms for target proteins	2008	NA	A, L	All	External DB, PubMed, $Binding_{all}$	CC BY-NC-SA 3.0[j] US	[87]
STITCH http://stitch.embl.de/	Aggregates protein–chemical interactions from both experimental and manually curated evidence	2008	5.0 (2016)	A, L	All	External DB, PubMed, $Binding_{all}$	CC BY-NC-SA 4.0[k,l]	[88]
KEGG Drug http://www.genome.jp/kegg/drug/	Drug information for approved drugs from Japan, USA, and Europe	2010	04/03/17	A[i]	Manual	NA	KEGG Medicus publicly available[m]	[82,83]
BindingDB https://www.bindingdb.org/	Experimentally determined binding affinities for protein–ligand complexes, extracted from scientific literature, selected databases, and US patents	2000	08/01/17	A, Cl, L	All	External DB, PubMed, $Binding_{all}$	CC BY 3.0[n,o] USA	[91]
ChEMBL https://www.ebi.ac.uk/chembl/	Large-scale bioactivity database with information manually extracted from medicinal chemistry literature	2011	23 (05/17)	A, Cl, L	All	External DB, PubMed, $Binding_{all}$	CC BY-SA 3.0	[90]
PubChem BioAssay https://pubchem.ncbi.nlm.nih.gov/#	Repository for results of high-throughput screening experiments for small molecules and RNAi	2004	NA	A, Cl, L	All	External DB, PubMed, $Binding_{all}$	Public domain[p]	[92,93]

[a]Database name and URL, year established, current version (release date, as available on August 4, 2017), drug/compound (A, approved drug; Cl, clinical trial/investigational drug; N, nutraceutical/natural product; L, ligand; W, withdrawn), target inclusion criteria [Manual, manual curation involved; Threshold, activity threshold applied; All, all targets listed (applicable to bioactivity databases)], supporting evidence (PubMed ID; $Binding_{selected}$, selected binding values for drug–target interaction; $Binding_{all}$, all binding values for drug–target interaction reported;

External DB, parent source database as applicable), and database license. License information reflects most recent information in publication or from database websites on August 4, 2017.
[b]DrugBank Open Data datasets (subset of full database) available under Creative Commons CCO 1.0 International License.
[c]All rights reserved.
[d]IUPHAR/BPS Guide to PHARMACOLOGY.
[e]Open Data Commons Open Database License.
[f]Creative Commons Attribution-ShareAlike 3.0.
[g]Drug–Gene Interaction Database.
[h]Creative Commons Attribution-ShareAlike 4.0.
[i]Approved drug in Japan (PMDA), USA (FDA), and Europe (EMA).
[j]Creative Commons Attribution-NonCommercial-ShareAlike 3.0.
[k]Creative Commons Attribution-NonCommercial-ShareAlike 4.0.
[l]For chemical–protein interaction information.
[m]Academic users require subscription for FTP access; commercial users require subscription through Pathway Solutions.
[n]Creative Commons Attribution 3.0.
[o]Data sourced from ChEMBL is CC BY-SA 3.0.
[p]Data supplied by contractors or non-federal government entities or employees may be subject to copyright. However, PubChem listed as (ODbL) at https://www.healthdata.gov/dataset/pubchem.

Methods

The Need for a Unified Cancer Targetome

While there are many resources for drug–target interaction and compound bioactivity data, it remains an enormous task to collect, assess, potentially reconcile, and make informed decisions about putative drug–target interactions. This challenge is illustrated by a recent comprehensive analysis of all FDA-approved drugs that curated all efficacy drug targets (as defined by Santos et al. (2016)) through an extensive search of both prescribing information and the scientific literature [7]. There is a critical need for aggregation of drug–target information in a framework that allows assessment of the supporting evidence for each interaction.

We aggregate drug–target interaction and bioactivity data for FDA-approved antineoplastic drugs from four publicly available resources and introduce a framework for categorization of the type of evidence supporting each

interaction to create a unified **Cancer Targetome**. Briefly, we selected these four resources in an effort to obtain representative coverage of the drug–target interaction space that is both publicly available and widely used by the research community. DrugBank is a popular resource for drug and drug–target data that is used widely by pharmacy and medicinal researchers, clinicians, educators, and the public [81]. The Therapeutic Targets Database offers extended coverage of biological targets [107]. The IUPHAR/BPS Guide to Pharmacology utilizes expert manual curation and rigorously requires experimental binding evidence from a primary source for all drug–target interactions [85]. However, the IUPHAR/BPS Guide to Pharmacology typically provides only one experimental binding assay value for each drug–target interaction, so we also included an aggregated bioactivity database (BindingDB) in our collection efforts for the Cancer Targetome. BindingDB provides wide coverage of binding assay data by aggregating across the scientific literature as well as from other bioactivity resources such as ChEMBL and PubChem [92]. Across four resources (DrugBank, the Therapeutic Targets Database, the IUPHAR/BPS Guide to Pharmacology, and BindingDB), we retrieved a total of 137 drugs and 658 targets participating in a total of 6385 unique drug–target relationships. We emphasize that the number of unique drug–target relationships should not be regarded as an estimate of actual drug–target binding space, as many of these relationships are supported by experimental binding values that reflect very weak binding. DrugBank provided the highest coverage of drugs participating in drug–target relationships while BindingDB provided the highest coverage of targets

participating in drug–target relationships (Figure S1 in the supplementary information online). BindingDB also provided the highest coverage of unique drug–target relationships, which can be interpreted as drug–target interactions that have been experimentally tested but are not necessarily 'true' drug–target binding events.

To assess the strength of supporting evidence for collected drug–target interactions, we develop a three-level evidence scale. Evidence Level I, II, or III is assigned to drug–target relationships retrieved from a database with no additional supporting information, with supporting literature information, or with supporting literature information and at least one reported experimental binding value, respectively. Experimental binding values may be reported as K_d, K_i, IC_{50}, or EC_{50} assay values. Because drug–target information is aggregated across multiple databases, each unique drug–target relationship may have different types of supporting evidence reported across all four databases and therefore can be associated with multiple evidence levels.

Results

As we require increasing levels of supporting evidence for drug–target relationships, we see an overall decrease in the coverage of drugs, targets, and unique relationships, as expected (Figure 2). In the Level III evidence tier, we can further threshold according to the numeric value of reported experimental binding activities. This allows us to triage experimentally tested drug–target relationships

to those that have been reported with a binding value that is potentially relevant for drug and target binding having clinical impact.

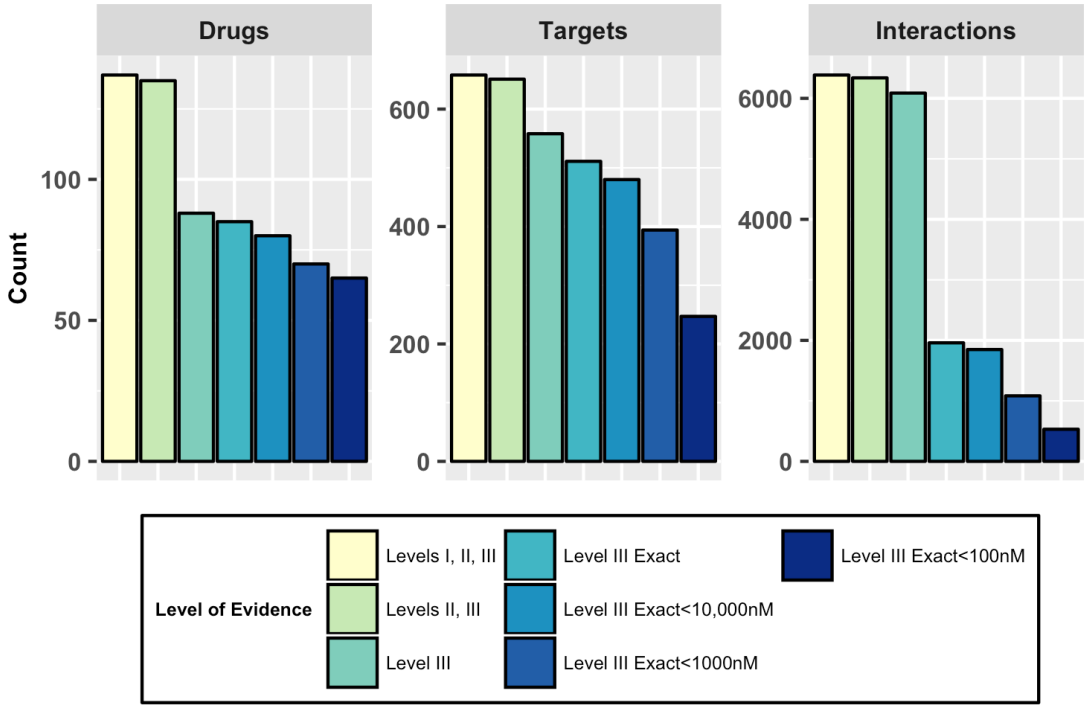


Figure 2. Cancer Targetome Aggregated Counts for Drugs, Targets, and Unique Interactions by Evidence Level. Drugs are FDA-approved antineoplastic drugs (total 137), uniquely identified using the National Cancer Institute Thesaurus [108] targets are unique human UniProt Accession IDs (total 658), and interactions are unique relationships (total 6385) between one drug and one target. Counts are colored by supporting evidence level: Level I indicates database annotation only; Level II indicates database and literature reference annotation; Level III indicates database, literature, and experimental binding value annotation. Level III Exact refers to annotation of a binding value with an exact ("=") binding value reported rather than "<" or ">". Thresholds on Level III binding activities were applied at 10 000 nM, 1000 nM, and 100 nM.

We demonstrate how the Cancer Targetome framework allows the filtering of aggregated drug–target relationships to obtain those meeting particular evidence criteria. For instance, to obtain an estimate of drug–target interaction space for which there is strong experimental evidence to support nanomolar binding interactions (e.g., relevant to clinically achievable doses for a given drug) we can require Level III evidence and further threshold to reported binding affinities of

less than 100 nM, which produces a total of 529 unique drug–target interactions. Interestingly, of these 529 putative drug–target binding interactions, the majority are reported by only one database, with only a quarter of these putative binding interactions reported by two or more databases (Figure S2A in the supplementary information online). Within this set of putative interactions, we can also examine the 'best' or minimum experimental binding affinity value reported for each unique drug–target interaction and the database that is responsible for contributing this value. While the majority of such minimum assay values are contributed by BindingDB, the IUPHAR contributes the minimum assay value for approximately 50 interactions (over 10%) (Figure 2B in the supplementary information online). This example highlights the benefit of aggregation across multiple sources to provide the research community with a more comprehensive resource for precision oncology.

Protein Kinase Inhibitors Are Highly Experimentally Tested Against Targets

The majority of antineoplastic drugs have been experimentally tested against fewer than 20 protein targets. This sparsity of the publicly available drug–target interaction space has been discussed by others [6] and represents a key limitation for efforts by the research community to assess drug promiscuity, or binding to 'secondary' targets. However, a small set of drugs (all protein kinase inhibitors) have been experimentally tested with more than 300 targets (Figure S3 in the supplementary information online), providing us with several examples of drugs with extensive binding data with which we may assess potential target

interactions and provide recommendations for future drug–target interaction curation efforts.

This meets expectations given the enormous resource commitment to targeting kinases in oncology following the breakthrough drug imatinib [2]. For instance, Davis et al. (2011) performed an extensive and comprehensive analysis of kinase inhibitor selectivity including both approved and investigational-stage drugs [105,109]. Experimental binding results for select approved cancer drugs from their analysis are included in our aggregated resource due to our data collection from the bioactivity database BindingDB. Among this set of highly tested kinase inhibitors, we see variation in the number of interacting targets for each drug (Figure S4 in the supplementary information online). As we threshold the experimental binding evidence to stronger binding affinities (10 000 nM, 1000 nM, 100 nM), we see that some drugs have a small number of targets meeting strong binding affinity criteria, such as afatinib, imatinib, and lapatinib, while other drugs have a seemingly high number of targets, such as bosutinib, crizotinib, dasatinib, and sunitinib. Due to the high number of experimentally tested targets for this subset of drugs, we can perform deeper data-quality analysis and, in particular, investigate the contributions of different experimental binding activity types.

Imatinib and Vandetanib Use Cases

We highlight two use cases for the drugs imatinib and vandetanib. Both of these drugs are protein kinase inhibitors and have extensive binding activity

information available across a large number of targets. Using the Cancer Targetome evidence framework, we assess the experimental evidence supporting target binding for imatinib and vandetanib at the strict threshold of 100 nM. In Figure 3A we show all targets for imatinib with experimental binding evidence under 100 nM. While there are a total of 14 targets with assay evidence under 100 nM, the tyrosine protein kinase ABL1, the canonical target of imatinib [2,31,110], notably has low-nanomolar assay evidence across all four binding assay types (K_d, K_i, IC_{50}, and EC_{50}). For K_d, K_i, and IC_{50} assay evidence, ABL1 has multiple low-nanomolar assay values, which lends more confidence to ABL1 being a biological target of the drug imatinib. Furthermore, for each of the four binding assay types, ABL1 has either the lowest or second-lowest assay value for target interactions with imatinib (Figure 3A). The case of imatinib serves as an example where evidence for the canonical 'primary' target can be seen in experimental binding data. In the cases where a target other than ABL1 occupies the best or close to the best assay value [epithelial discoidin domain-containing receptor 1 (DDR1), platelet-derived growth factor alpha (PDGFRA), and platelet-derived growth factor beta (PDGFRB)], there is binding assay support from only one or two of the binding assay types rather than all four binding types as in the case of ABL1.

In Figure 3B we show all targets for vandetanib with experimental binding evidence under 100 nM. In total there are 26 unique targets meeting these criteria but we see a striking discordance in the type of binding assay support

available for these targets. Experimental K_d values indicate that ABL1, the mast/stem cell growth factor receptor Kit (KIT), receptor-interacting serine/threonine protein kinase 2 (RIPK2), the epidermal growth factor receptor (EGFR), and the proto-oncogene tyrosine protein kinase receptor Ret (RET) have very-low-nanomolar experimental evidence. Experimental IC_{50} values indicate that vascular endothelial growth factor receptor 2 (KDR or VEGFR2), EGFR, vascular endothelial growth factor receptor 1 (FLT1), and the proto-oncogene tyrosine protein kinase Src (SRC) all have evidence for interaction at a very-low-nanomolar level. According to K_d assay evidence, EGFR is strongly supported as a target (multiple low-nanomolar assay values), while according to IC_{50} assay evidence KDR is strongly supported as a target. For vandetanib, no EC_{50} binding assay values and no K_i binding assay values under 100 000 nM were available.

Interestingly, vandetanib is considered to be a dual KDR and EGFR inhibitor or, in some cases, a multiple kinase inhibitor for EGFR, KDR, and RET [2,111,112]. A literature search reveals that, while originally designed to inhibit KDR, vandetanib exhibited additional activity with EGFR in preliminary lead candidate stages [112]. These results prompted further testing that established vandetanib as inhibiting EGFR in mouse cells, human cancer cells, and seven human cell lines lacking the target KDR [113]. This example highlights the rich contextual information for drug–target interactions that is currently not captured in drug–target interaction or bioactivity resources.

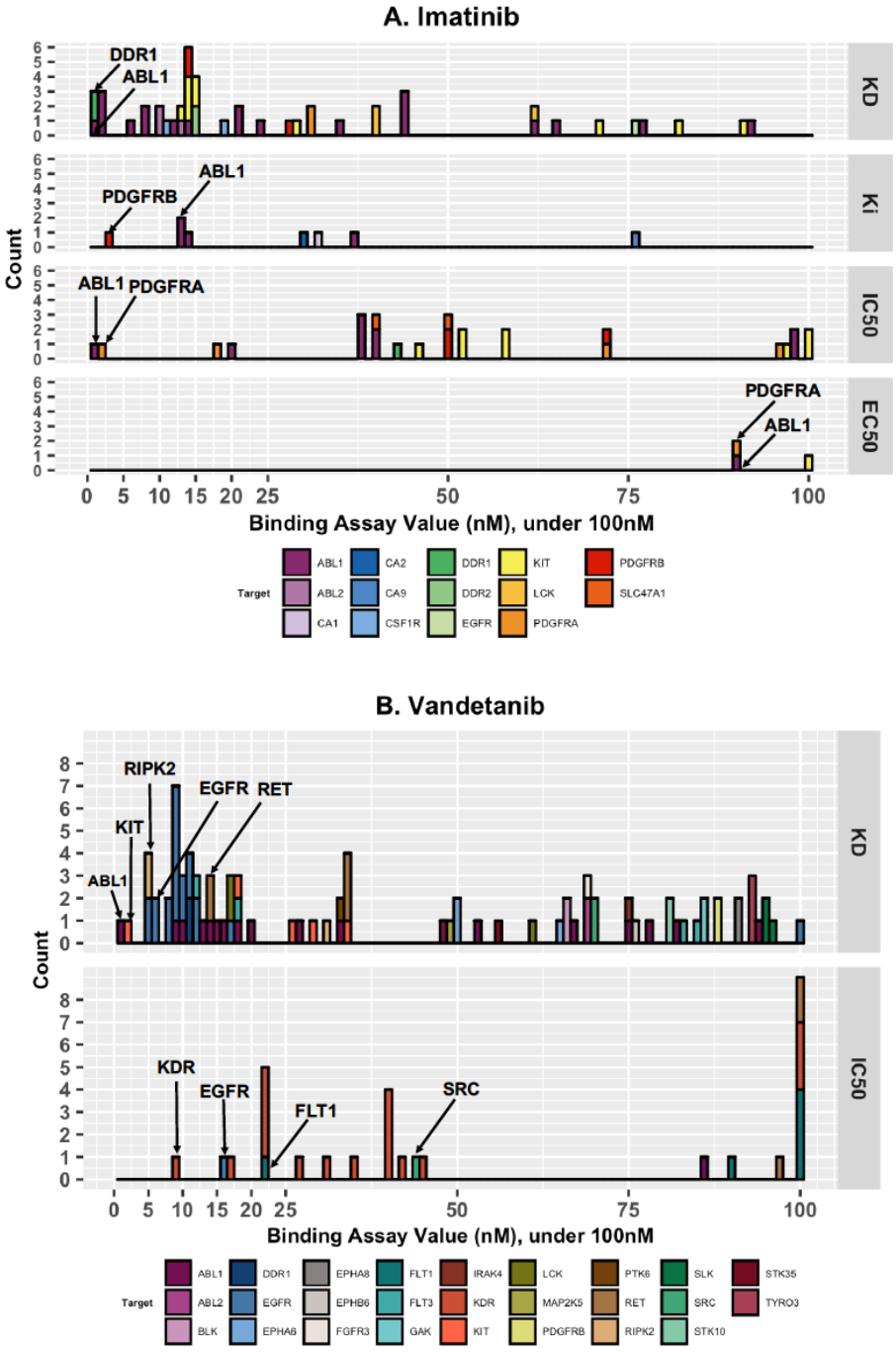


Figure 3. (A) Imatinib target interactions under 100 nM. Colored by target; bin width, 1 nM. Imatinib has experimental binding evidence for 14 different targets under 100 nM. ABL1 stands out as it has many low-nanomolar assay results and occupies the best or second-best assay value for each binding assay type. **(B) Vandetanib target interactions under 100 nM**. Colored by target; bin width, 1 nM. Vandetanib has experimental binding evidence for interactions with 26 different targets under 100 nM. K_d and IC_{50} assay evidence provides strong support for different targets for vandetanib [epidermal growth factor receptor (EGFR) and KDR, respectively]. Vandetanib does not have any targets supported by EC_{50} assay evidence and does not have any targets supported by <100 000-nM K_i assay evidence.

Next Steps for Drug–Target Interaction Evidence Curation

A unified Cancer Targetome framework provides researchers with access to cancer drug–target relationships from the public domain that are accompanied by a transparent literature and experimental binding evidence lineage. The proposed evidence framework allows researchers to prioritize drug–target relationships according to the evidence criteria that are best suited to their research aims. Transparent and well-evidenced drug–target interactions will enable higher confidence and more informed decision making in the prioritization of drugs and targets in precision oncology efforts.

However, examining the factors needed for the creation of the Cancer Targetome reveals critical unmet needs. In particular, the vandetanib use case highlights the need for binding assay metadata. While we were able to retrieve and assess experimental binding affinities between vandetanib and many biological targets, we must also consider the information that is not captured in this process. Namely, we are currently unable to capture metadata such as the cell line used in experimental binding assays, the tumor or non-tumor status of the cell line, and whether the cell line is derived from patient cells. The availability of this metadata would allow further tiering of drug–target binding evidence to aid target prioritization. For instance, the category we have proposed for experimental binding evidence (Level III) could be further subdivided into tiers indicating whether the interaction has been tested in non-cancer cells, cancer cells, or cells that are patient derived. Further tiering could be used to capture

metadata indicating whether other targets were knocked down or remained functional during the experimental binding assay for the target of interest. This metadata is invaluable for the prioritization of drug–target binding information in precision oncology, where it is critical to know whether experimental evidence was obtained using cancer or non-cancer cell lines.

Mapping Drug–Target Interactions to Pathways

Given the dysregulation that can occur in multiple pathways in cancer, there has been increasing attention and effort dedicated to targeting cellular pathways, particularly through the use of combination drug therapies [8,114]. We conducted a simple pathway analysis to assess the targeted pathway coverage of approved cancer drugs. Briefly, we mapped all targets participating in drug–target relationships to Reactome pathways using increasingly strict supporting evidence requirements. Reactome is a comprehensive open-source pathway resource widely used by the research community [24]. Cellular pathways in Reactome are organized in a hierarchical manner allowing smooth pathway navigation and improved integration with external data resources. We designate those biological pathways containing one or more drug targets as ‘light’ or potentially targetable by approved antineoplastic drugs. Conversely, biological pathways containing no drug targets are ‘dark’ or currently out of scope for approved antineoplastic drugs. While a considerable portion of pathways (approximately 60%) are light to antineoplastic drugs when we consider any type of supporting evidence for drug–target interactions, this should be considered the

most liberal estimate of potentially targetable pathways (Table 2). A more reasonable estimate is obtained when we require drug–target relationships to be supported by experimental binding evidence with a reported assay value of less than 100 nM. This estimate indicates that there is strong evidence for approved antineoplastic drugs targeting approximately 39% of Reactome pathways. Depending on the distribution of key molecular aberrations for a given patient among the light and dark pathways, the evidence-based curation as presented and envisioned herein will refine the selection of therapeutics and in some cases could dramatically limit therapeutic options. We highlight the NOTCH signaling pathway in Box 1, which contains several dark child pathways. Dark pathways that are currently out of scope for FDA-approved cancer drugs represent areas for future cancer therapeutics development.

Table 2. Light Pathways of the Cancer Targetome, by Supporting Evidence Level

Evidence level required	*Number of unique targets*	*Number of light pathways*	*Total number of pathways*	*Percentage light pathways*
Levels I, II, III	658	1214	2008	60.46%
*Levels II, III**	651	1213	2008	60.40%
Level III	558	1139	2008	56.72%
Level III Exact	511	1091	2008	54.33%
Level III Exact, threshold < 100 nM	246	790	2008	39.34%

All unique targets involved in drug–target relationships (supported by indicated evidence levels) were mapped to Reactome pathways. The number of unique targets in the set of drug–target relationships is shown in the second column. A pathway is considered light if it contains at least one drug target.

Box 1. Pathway Example: NOTCH Signaling

Dark pathways are of particular interest for future drug discovery and development efforts as they are currently outside the scope of approved cancer agents. We highlight the NOTCH signaling pathway, which is light at the topmost hierarchical level but contains several dark child pathways (Figure I). Dysregulated NOTCH signaling has been implicated in breast, prostate, lung, head and neck, and central nervous system cancers as well as T cell leukemia and has thus been identified as a therapeutic target of interest [115–117]. Three of the five child pathways of NOTCH signaling are currently dark to cancer drugs (Signaling by NOTCH2, Signaling by NOTCH3, and Signaling by NOTCH4). We highlight the light child pathway Pre-NOTCH Expression and Processing in the figure inset, which shows that there are two drugs potentially interacting with two targets in this pathway. Arsenic trioxide putatively interacts with the transcription factor AP-1 (JUN) (UniProt P05412) and G1/S-specific cyclin D1 (CCND1) (UniProt P24385) while vinblastine sulfate putatively interacts with JUN.

However, all three of these drug–target relationships have Level II evidence only, as there is no accompanying experimental binding evidence. Therefore, if we assess light pathway coverage while requiring at least experimental binding evidence for drug–target interactions, this nested pathway goes dark. We use this example to illustrate that the classification of a particular pathway as light or dark to approved cancer drugs is directly impacted by the

strength of supporting evidence for the drug–target interactions involving the pathway of interest.

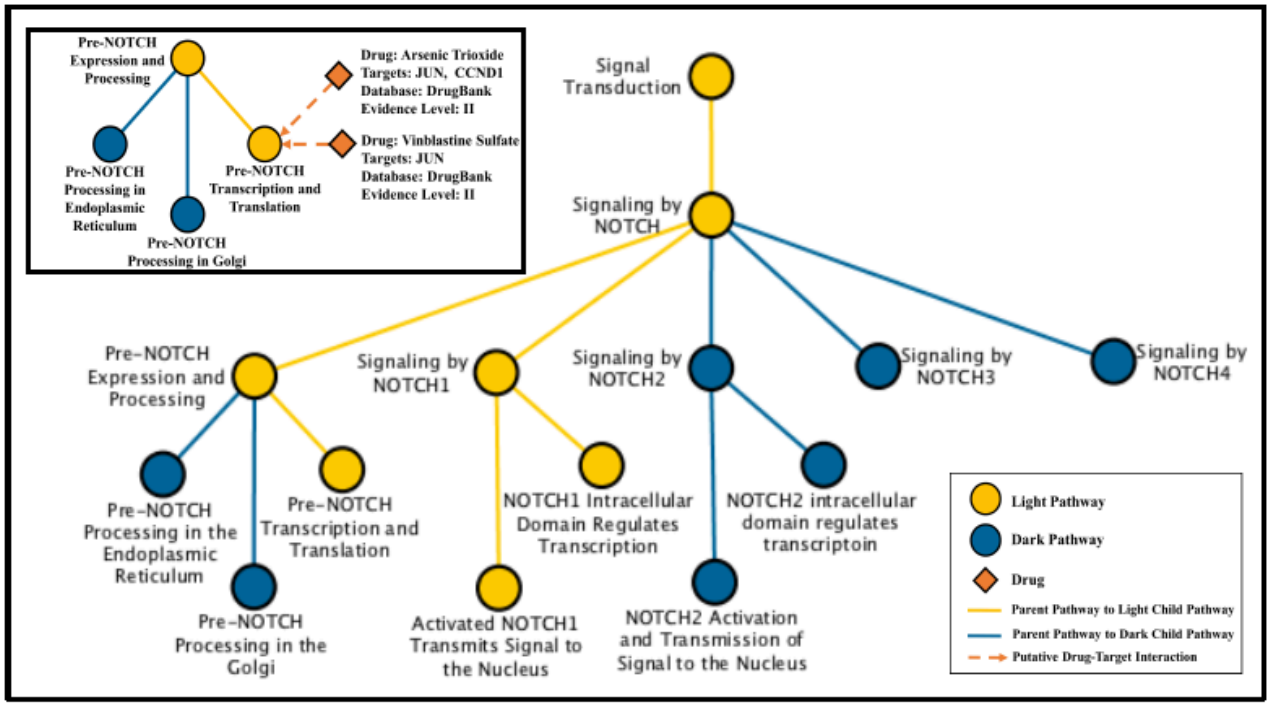


Figure I. Signaling by the NOTCH Pathway. Signaling by NOTCH is a light or potentially targetable pathway when considering drug–target relationships supported by any level of evidence. Main figure: Of the five child pathways in Signaling by NOTCH, two are light (gold) and three are dark (blue) to current approved antineoplastic drugs. Inset: The light child pathway Pre-NOTCH Expression and Processing contains two targets, JUN and CCND1, that are putatively targeted by antineoplastic drugs. This pathway is light when drug–target interactions of Level II evidence are included but goes dark when Level III evidence is required.

Discussion

We foresee the possibility that these analyses will allow weighting of the level, extent, and type of evidence to guide the prioritization of drugs moving to the clinic, for better synchronization of preclinical promise and patient benefit. Recently, attention has been drawn to the need for evidence quantification of patient-specific alterations in tumors to guide decisions about actionable therapies [5]. A similar characterization of evidence is also needed for drug–

target and drug–pathway interactions if we hope to unite drug–target information with patient-specific information and develop targeted therapies (Box 2; see Outstanding Questions). In particular, evidence characterization frameworks accommodate the inherent uncertainty in the targetome space due to multiple types of supporting evidence.

Given the recent attention and dedication of resources to the investigation of understudied areas of the druggable genome by the NIH Illuminating the Druggable Genome Consortium, we believe this work will be of current interest to the larger precision medicine community. This has implications for other therapeutics areas of interest with respect to guided investigation into understudied and underdeveloped therapeutic drugs, targets, and pathways.

Box 2. Mapping Evidence Levels to Precision Oncology Applications

The appropriate level of evidence to require when including drug–target interactions in precision oncology applications will be heavily context dependent. In Figure 1 we detailed several examples of precision oncology applications that have a key dependency on drug–target interactions. For applications that are exploratory or hypothesis generating in nature, such as computational and predictive modeling (Figure 1B), the use of drug–target interactions supported by Level I or Level II will often be appropriate. Such applications would benefit from casting a wider net of drug–target interactions so that all options can be explored. Similarly, exploratory work geared towards drug repurposing (Figure 1C), such as the inclusion of FDA-approved drugs on a screening panel for an indication

other than the drug's primary one, may also benefit from liberal evidence requirements that allow the investigation of all possibly relevant drug–target interactions. While additional Level III experimental binding evidence would lend support to these interactions being potentially relevant for human physiology, this will always not be necessary at the discovery stage. Applications involving the planned use of a drug in a patient, however, will require (at a minimum) rigorous Level III experimental binding assay evidence. These applications could include off-label use of a drug, design of combination therapies (Figure 1D), or the inclusion of an already-approved drug in a clinical trial for an alternative indication (Figure 1A). In these examples choice of therapy may be driven by a patient's particular molecular aberrations if there is substantial evidence that those aberrations can be targeted by an existing pharmaceutical therapy. The requirements for evidence supporting such drug–target interactions must be very rigorous – meaning very-low-nanomolar binding evidence for a drug–target interaction, ideally across binding assay types and from multiple, independent sources. As mentioned in the text, experimental metadata (such as cell line information) will also be necessary for rigorous evaluation and prioritization of drug–target interactions. For clinical applications of drugs in patients, we emphasize that drug–target interaction evidence (even rigorously supported evidence) is intended to supplement but never to replace an oncologist's or tumor board's expertise and recommendations. We envision the use of this information as one line of evidence among the many that are evaluated by medical experts when deciding the best course of action for a patient.

Author Contributions

This chapter was originally published as ‘Evidence-Based Precision Oncology with the Cancer Targetome’ by Aurora S. Blucher, Gabrielle Choonoo, Dr. Molly Kulesz-Martin, Dr. Guanming Wu, and Dr. Shannon K. McWeeney in Trends in Pharmacological Sciences on Sep 27, 2017. DOI 10.1016/j.tips.2017.08.006.

CHAPTER 4. AIM 2: A TARGETOME-PATHWAY PERSPECTIVE ON DRUG RESPONSE FOR TARGETED THERAPIES IN ACUTE MYELOID LEUKEMIA

Abstract

To meet current challenges in precision cancer treatment, we require a mechanistic understanding of drug targeted therapies and their interaction with patient genetic alterations. Investigating the association between aberrational pathways and drug response in de novo acute myeloid leukemia patient samples from the Beat AML Consortium reveals many significant pathway-level associations with drug sensitivity or resistance. We note these are driven by mutations in a spectrum of genes within the pathway, and therefore are potentially missed when considering only single gene interactions with drug response. To further understand how these intrinsic mutational perturbations result in drug sensitivity or resistance, we use a probabilistic graphical modeling framework to model pathway impact. Complementary to this, we model extrinsic drug perturbations on these pathways using quantitative drug-target binding information from the Cancer Targetome to model impact downstream. We will discuss our developments on the development of a unified framework of intrinsic and extrinsic perturbation modeling for rigorous *in silico* hypothesis generation and testing to facilitate future drug combination screening recommendations.

Introduction

While targeted therapies have revolutionized cancer treatment with the ability to target specific genetic aberrations, challenges such as tumor heterogeneity and intrinsic drug resistance often result in lack of patient drug response or cancer relapse [1,33]. In particular, response to single-agent targeted therapies is often temporary and soon eclipsed by drug resistance [29]. In light of these challenges, combination therapies offer the potential of targeting multiple genetic aberrations at once to tackle tumor subclonal populations or overcome resistance mechanisms [33]. However, in order to rationally design combination therapies, we require a mechanistic understanding of single agent targeted therapies and their interaction with patient genetic alterations.

Traditionally, pharmacogenomics approaches investigate associations between gene-level mutations and drug sensitivity or resistance [34]. To capture the effect of mutations across multiple genes, we pursue a pathway-based perspective on mutational aberrations. Pathway-based analysis and modeling frameworks enable us to aggregate multiple mutations to signaling pathways and construct sets of shared dysregulated pathways across patients [37]. Aggregation of mutations to pathways provides a perspective on intrinsic or mutational perturbation that considers multiple aberrations per patient sample as well as shared pathway-level dysregulation across patient samples.

In this aim, we focus on an application to understanding patient response to targeted therapies in acute myeloid leukemia. Acute myeloid leukemia (AML) has seen little improvement in treatment and overall survival in the last forty

years [26,29,118]. Notably, AML exhibits considerable heterogeneity across patients, where some cases are characterized by large chromosomal rearrangements but the majority have mutations in individual genes [26,27]. AML is also characterized by a fairly low number of gene mutations per patient when compared to other cancers [28]. Given the heterogeneity of mutations observed in AML patient, especially for low frequency mutations, there is a need for improved understanding of how patient mutation profiles affect overall response to targeted therapies [28,119]. Pathway-based approaches are therefore an attractive potential avenue for understanding more about AML mutational architecture and how it affects drug response. In particular, improved understanding will help to facilitate the development of combination strategies for treatment of AML.

Here, we take a pathway perspective on patient mutational aberrations in de novo acute myeloid leukemia. We show that aggregation of mutations to functionally defined pathways in Reactome [24,25] allows us to capture associations between pathways and drug response. This pathway view reveals more information about targeted therapy response than what is necessarily obtained from drug-single gene relationships. Building on these findings, we introduce a modeling framework, built on probabilistic graphical modeling principles, to model mutational pathway impact [37,38]. We demonstrate that this framework models pathway impact in a manner that reflects both functional impact of a mutated gene and the topology of the impacted pathway. In parallel, we use this framework to also model drug pathway impact. Previous work with

the Cancer Targetome leverages information about cancer drugs and their targets in an evidence-based empirical manner and allows us to consider drug binding across multiple targets [120] **(Chapter 3).**

With this pathway modeling framework, we introduce a way to interrogate the effect of intrinsic mutation perturbation and extrinsic drug perturbation on biological pathways. We can use this information to begin to tease apart the role these perturbations play in overall patient drug response and ask targeted questions for further analysis and bench testing. We explore modeling coverage and provide recommendations for adapting this modeling framework to handle both mutation and drug pathway impact in the future. We then explore future directions for joining mutation impacted and drug-impacted pathways in the modeling framework to guide concerted efforts to understand drug response in AML.

Methods

BeatAML Patient Samples

This analysis uses patient mutation and drug screen data from the BeatAML Consortium [30]. For a total of 187 samples for patients with de novo AML, mutations from whole exome sequencing were used (paired samples, tumor/normal). The following thresholds on mutation functional impact were required: PolyPhen >0.25, SIFT <0.50, or an indication of 'HIGH' from Variant Effect Predictor (VEP). If there was no PolyPhen, SIFT, or 'HIGH' indication, then 'moderate' or 'modifier' were also allowed for mutations. Accompanying drug

screening data was available for 108 of 187 AML de novo patient samples [30,31].

Mapping to Reactome Pathways

Gene mutations for 187 AML De Novo patient samples were mapped to Reactome [25] pathways using gene symbols for entities in each pathway. A total of 461 Reactome pathways were used, of which 363 contained AML de novo mutated genes.

Pathway Enrichment by Gene

To assess pathway enrichment by gene, a hypergeometric test was used to assess whether the number of genes mutated in each pathway (across all 187 AML de novo samples) was greater than the number of genes expected to be mutated in each pathway at random. The resulting p-values were corrected using the Benjamini-Hochberg or False Discovery Rate multiple testing correction method [122].

Pathway Enrichment by Sample

To assess pathway enrichment by number of samples, we conducted a series of permutation tests. First, for each sample, we tallied the observed number of mutated genes. Then, we randomly drew that same number of genes from the set of all possible pathway genes. Assigning these genes to their respective pathways created a sample-to-gene-to-pathway mapping. This was repeated

10,000 times to create a distribution of the expected number of samples with genes mutated for each pathway. Then, for each pathway, we assessed how extreme the observed number of samples mapped was when compared to the distribution of expected number of samples mapped. P-values were corrected using the Benjamini-Hochberg method for correcting for False Discovery Rate [122].

ANOVA Analysis for Pathway Mutational Status and Drug AUC

For each pathway and drug interaction with at least 2 mutated samples, ANOVA analysis was conducted to assess if there was a difference in mean drug AUC between two groups: samples with mutation(s) in the pathway, and samples without mutation(s) in the pathway. Resulting P-values were corrected for multiple testing using the Benjamini-Hochberg [122] method. For effect size, Cohen's D was calculated as follows:

$$Cohen's\ D = (mean1 - mean2) / (sqrt\ (sd1\text{^}2 + sd2\text{^}2\ )$$

Creation of Factor Graphs from Reactome Pathways

For a total of 431 biochemical Reactome pathways, a factor graph model was constructed for each pathway. Factor graphs are bipartite graphs with two types of nodes: variable nodes and factor nodes [38,39]. Pathways are converted in the following manner, pathway entities become 'variable' nodes, and factor functions (or factor nodes) designate the relationship between variable nodes [37]. Factor

functions can also be interpreted as ‘compatibility’ functions, which indicate the respective state probabilities for a variable based on its neighbors. Then, within each factor graph model, every variable node where a gene is specified (‘gene’ nodes) are extended with an additional factor function and a ‘protein’ node [37,41,123]. These ‘protein’ nodes become the anchors for which we will attach either mutation or drug information.

Belief Propagation (Message Passing) on Factor Graphs

Briefly, we explain how we conduct message passing using loopy belief propagation [38,39] on each pathway factor graph with no observed sample or drug information. This is to establish the baseline, or null perturbation states for nodes in the factor graph in the absence of any patient mutation or drug information. For each factor graph, we first initialize messages (of 1.0) for all factor nodes. Then, we randomize the order of factors and variables for message propagation throughout the graph. When every node has both an incoming and an outgoing message, then at each node we calculate the ‘beliefs’, or associated probabilities that the node is in either state 0 or state 1. This is a vector that specifies the belief or probability that the node is in either state, where the vector contents sum to 1 [38,39]. For instance, a node with the vector [0.33, 0.67] would mean that there is a 33% probability that the node should be in the 0 (unperturbed) state and a 67% probability that the node should be in the 1 (perturbed) state. We conduct message passing iteratively until ‘convergence’ of

entity belief states (no change in the estimated final probabilities for each node), or we reach the maximum iteration (set to 250) without reaching convergence.

Pathway Stability Assessment

Because loopy belief propagation is not guaranteed to converge, and may converge to different state configurations, we conducted extensive stability analysis. For each pathway factor graph, we ran the null perturbation run 1000 times. Then, we clustered the null perturbation run results into groups, where group membership requires the same state configuration across all nodes in the pathway. Pathways were designated as ‘stable’ if we achieved the same configuration state at least 950/1000 runs. A ‘stable’ configuration means that for multiple runs, every node in the pathway reached the same state assignment probabilities (<0.001 difference). Pathways were considered ‘not stable’ if they did not meet this criteria, or alternatively, if no convergence at all could be reached for the pathway, as in the case of pathways with a very large number of loops in the graph, they were designated as ‘no convergence’. Out of 431 pathway factor graphs, 291 were found to be stable according to these criteria and used for further analysis in mutation and drug impact analysis.

Probabilistic Graphical Model with Mutation Data and Calculation of Pathway Impact

For each sample, mutation data is loaded onto the set of pathway factor graph. For every mutated gene, the functional impact score is encoded into the factor

function specifying the impact at the protein node. The table below specifies the conversion between functional impact score and factor function. The factor function for each protein node must specify the belief that the protein is in state 0 (no impact) and the belief that the protein is in state 1). For example, a mutated gene with a PolyPhen functional impact score of 0.99 would have a corresponding factor function encoded as [0.01, 0.99]. This indicates that our belief that the protein is in the perturbed state is 0.99 while our belief that the protein is in the unperturbed state is 0.01 (Table 1).

Table 1. Factor node encoding of mutational impact according to functional impact (PolyPhen, SIFT, or Variant Effect Predictor (VEP) scores).

Functional Impact Type	***Functional Impact Score***	***Factor Function [State 0, State 1]***
PolyPhen	X	[1-X, X]
SIFT	Y	[Y, 1-Y]
VEP = High	NA	[0.0002, 0.9998]
VEP = Moderate	NA	[0.50000,0.50000]

Mutation information is loaded on the appropriate protein nodes, in all pathways where the protein node is a member. Thus, one mutated gene may actually have impact on several pathways. Message passing (loopy belief propagation) is then run for each pathway, independently, to calculate the beliefs at each node. Overall pathway impact is calculated by comparing the beliefs at each node in the pathway factor graph to the beliefs calculated from the null perturbation run.

For each entity, we calculate the odds of impact for perturbation as follows:

Log Odds = Log10([Belief State 1 Mutation/ (1- Belief State 1 Mutation)] / [Belief State 1 Null / (1- Belief State 1 Null)])

Thus for each pathway, we have an odds score for each entity. Then, for each pathway, we summarize overall pathway impact by averaging the log odds score. Lastly, we average over only designated 'output' entities in the pathway, where 'output' entities are designated as such by Reactome biochemical pathway annotation.

Probabilistic Graphical Model with Drug-Target Interaction Data

Here, instead of calculating pathway impact for each sample, we calculate pathway impact for each drug. For each of the 32 drugs on the BeatAML screening panel and in the Cancer Targetome, we load drug-target interaction information onto the pathway graphs. Similar to the methods used in the pathway modeling for mutation data, we specify factor functions for 'protein' nodes wherever we load drug-target information. Factor functions are specified according to the strength of the drug-target binding assay data obtained in the Targetome (Table 2). For this analysis, these relationships were naively set so that lower nanomolar interactions results in higher impact on target nodes. For each drug-target interaction pair, we used the best (lowest) possible assay binding value present in Targetome. Future sensitivity analysis is needed to tune these factor functions and assess more quantitatively the relationship between factor node impact and overall drug impact on the pathway.

Table 2. Factor node encoding of drug impact according to drug-target interaction binding assay evidence from the Cancer Targetome.

Drug-Target Interaction Best Assay Evidence	*Factor Function [State 0, State 1]*
>10,000nM	[0.999, 0.001]
1000-10000nM	[0.250, 0.750]
100-1000nM	[0.150, 0.850]
10-100nM	[0.050, 0.950]
1-10nM	[0.025, 0.975]
<1nM	[0.001, 0.999]

We conduct message passing for each pathway factor graph with the drug-target interaction information loaded into factor functions, and then calculate the overall pathway impact score in the same manner as for mutation pathway impact.

PGM Modeling Coverage and Assessment

For modeling coverage assessment, we took the set of all significant drug-pathway associations from the ANOVA analysis conducted with the simple pathway mapping (see Section 'ANOVA Analysis for Pathway Mutational Status and Drug AUC above'*)*. For these drug-pathway interactions, we assessed the mutational pathway impact score from the modeling approach. Note that for each drug-pathway interaction, we used the average mutation pathway impact score across all mutated samples. For those drug-pathway interactions where we were able to conduct drug-impact pathway modeling, we also assessed the drug pathway impact score. We created coverage graphs that show the ANOVA

significant drug-pathway interactions in light of both the mutation impact pathway scores and the drug impact pathway scores.

Results

AML De Novo Mutations and Pathway Coverage

This analysis uses acute myeloid leukemia patient samples collected as part of the BeatAML Consortium [30]. All patient samples used in this analysis are paired tumor and normal samples for de novo AML. For a total of 187 de novo acute myeloid leukemia patient samples, gene mutation information was mapped to Reactome pathways. We required mutations to have a functional impact score meeting at least one of the following thresholds: PolyPhen > 0.25, SIFT<0.50, a Variant Effect Predictor category of 'High', or either 'Moderate' if no other fields were provided. Out of 431 total Reactome pathways subject to analysis, 363 included mapped de novo AML mutations. Table 3 shows the number of mutated genes in each functional impact category and in total. The term 'high prevalence' refers to those genes mutated in more than 5 samples in this cohort.

Table 3. Gene Mutations for 187 AML De Novo Patient Samples, by Variant Impact and Pathway Coverage to Reactome Pathways.

	Samples (n=187)		*AML Mutated Genes (n=1654)*						*Pathways (n=431)*
Variant Category	**# Samples with Variants**	**# Samples Mapped**	**# Mutated Genes Total**	**# Mutated Genes Mapped**	**% Mutated Genes Mapped in Category**	**% Mutated Genes Mapped of Total (1654)**	**# High Prevalence Genes**	**# High Prevalence Genes Mapped**	**# Pathways with Mutations**
PolyPhen >0.25	177	166	822	438	53.28	26.48	14	14	318
SIFT <0.50	182	174	1237	644	52.06	38.93	14	14	351
PolyPhen OR SIFT	182	174	1280	663	51.80	40.08	14	14	255
HIGH IMPACT **No PolyPhen or SIFT*	172	153	308	161	52.27	09.73	12	9	199
MODERATE **Overlaps PolyPhen and SIFT*	186	179	1405	727	51.74	43.95	14	14	359
All Variants	**187**	**184**	**1654**	**850**	**51.39**	**51.39**	**25**	**21**	**363**

From a pathway perspective, we find that the majority of pathways have approximately 1-20 samples mapped, with a small number of pathways having between 20 and 50 samples mapped, and a few outlying pathways with over 50 samples (Figure 1A). Within each pathway, between 1 and 15 genes are mutated

for the majority of pathways, with a few pathways containing over 30 mutated genes across the cohort (Figure 1B). From a sample perspective, we find that most samples map to between 1-45 pathways, with a few samples mapping to more than 45 pathways.

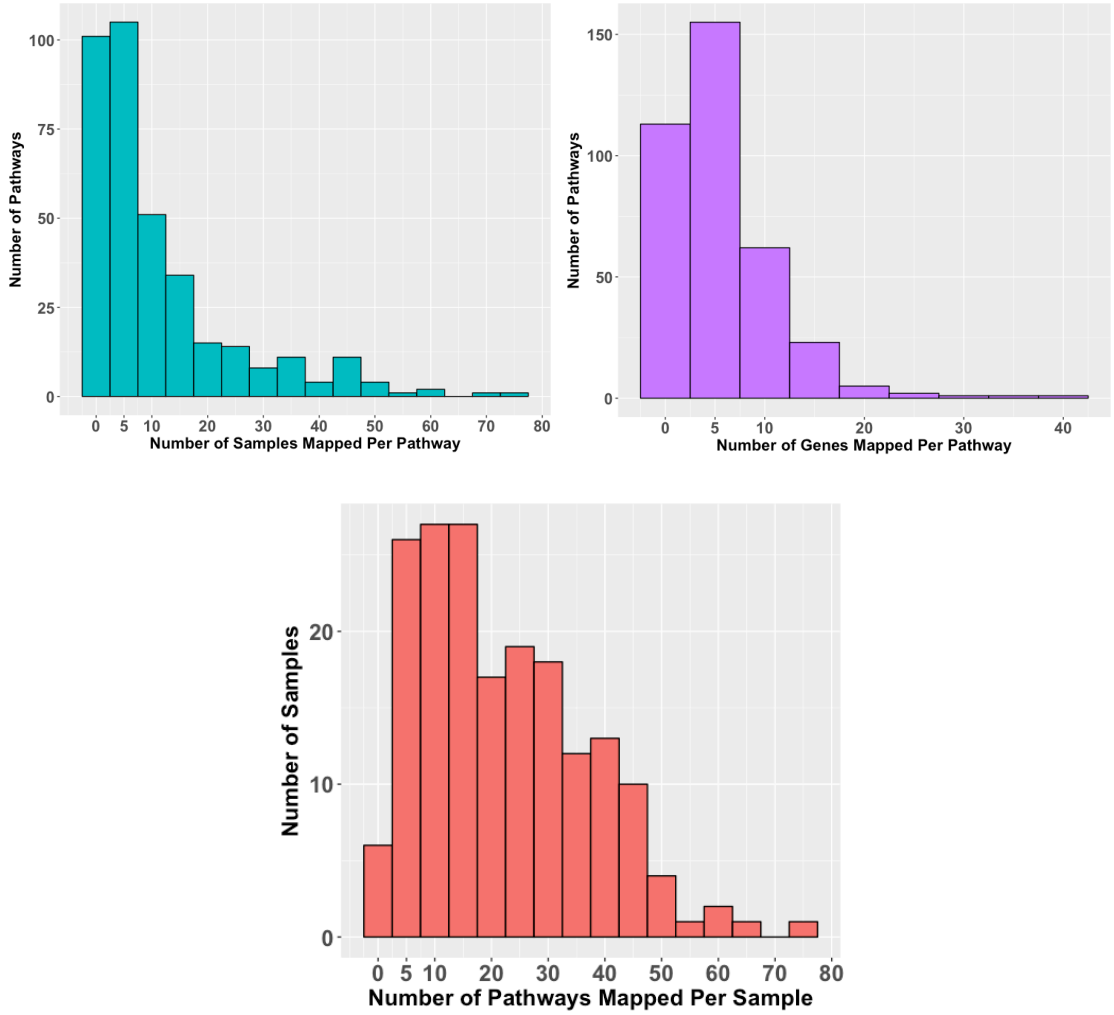


Figure 1. A. Pathway Coverage by Number of Samples Per Pathway, for 187 AML De Novo Patient Samples. B. Pathway Coverage by Number of Mutated Genes Per Pathway, for 187 AML De Novo Patient Samples (187). C. Number of Pathways Per AML Per Sample, for 187 AML De Novo Patient Samples. Gene mutations for 187 AML de novo samples mapped to 363 total Reactome pathways.

AML is well established as being extremely heterogeneous across patients. Known gene mutations of impact for this disease include FLT3, NPM1, DNMT3A, IDH2, and TET2 among others [29]. When taking a pathway mapping approach, it is critical to consider coverage of key mutations by pathways. While Reactome is the most comprehensive pathway database, not all cancer-related genes will currently be curated due to limitations in our pathway understanding of these entities [25]. We assessed the top mutated genes in our cohort of de novo AML patient samples for coverage by Reactome pathways. Out of the 25 genes mutated in five or more patient samples, we mapped 21 to Reactome pathways (Figure 2). The four unmapped genes are WT1, BCOR, PHF6 and MXRA5. Overall we achieve a very high coverage of key mutations for our cohort of de novo AML patient samples.

By aggregating gene mutations to pathways, we are grouping multiple mutations together into sets of functionally interacting entities. This approach allows us to aggregate mutations for a single patient to pathways as well as across patients. In particular, it allows us to assess AML mutations from a perspective of shared dysregulated pathways. Given the heterogeneity in AML, this may allow us to collect together mutations in seemingly disparate genes into a smaller number of functionally connected pathways or sets of pathways.

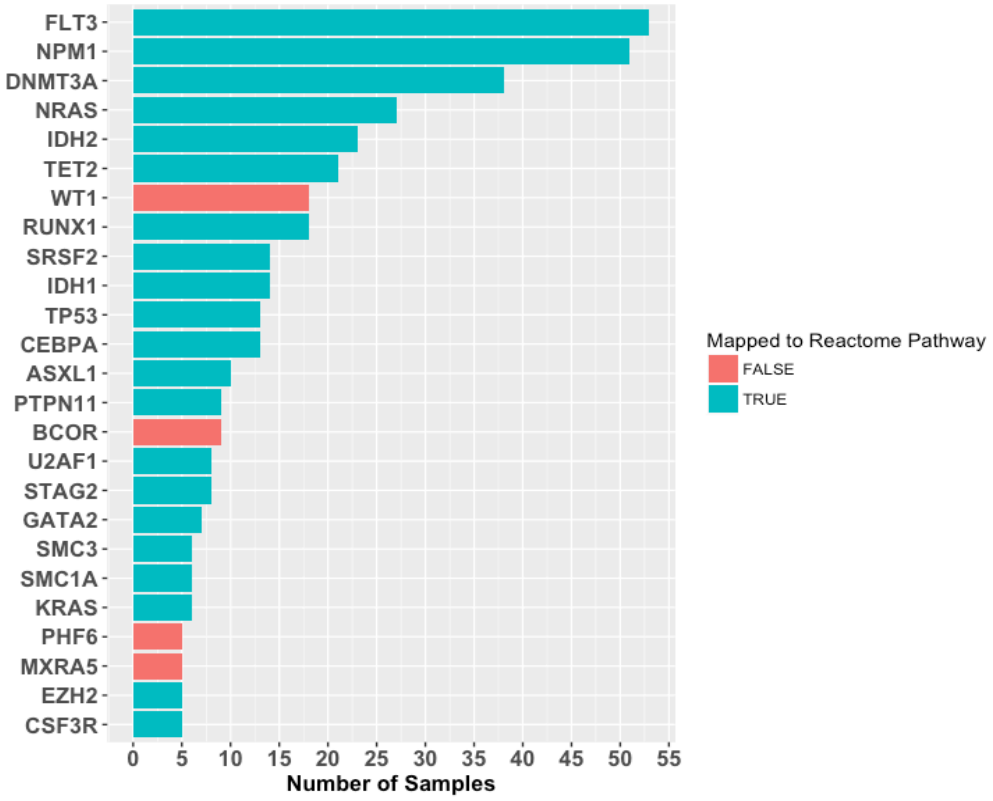


Figure 2. Pathway Coverage for Top Mutated Genes in 187 AML De Novo Patient Samples. Genes shown in figure are mutated in at least five AML patient samples. Blue indicates genes mapped to one or more Reactome pathways, red indicates genes not mapped to Reactome

We examine pathway coverage for AML de novo mutations by number of samples and genes simultaneously (Figure 3). The bubble plot in Figure 3 shows that top pathway with respect to number of samples mapped include: 'Chromatin organization', 'TP53 Regulates Transcription of Cell Cycle Genes', 'Other Interleukin Signaling', 'Transcriptional regulation by the AP-2 (TFAP2) family of transcription factors', 'Neutrophil degranulation', 'Nucleosome assembly, 'Transcriptional Regulation by RUNX3', 'Signaling by SCF-KIT', 'Signaling by FGFR1', 'Signaling by MET', and 'Signaling by PDGF'. Top pathways with respect to number of mutated genes mapped include 'Chromatin organization', 'Neutrophil degranulation', 'Generic Transcription Pathway', 'Deubiquitination',

'Ion channel transport', 'Keratinization', and 'Class I MHC mediated antigen processing and presentation'. While such assessment provides an estimate of coverage, we note that some of the pathways with a large number of mutated AML de novo genes are very large pathways at baseline (for instance, Generic Transcription pathway), and would likely have a large number of genes mapped for most disease states. For this reason, we are interested in investigating pathways that show enrichment for AML de novo-specific mutations.

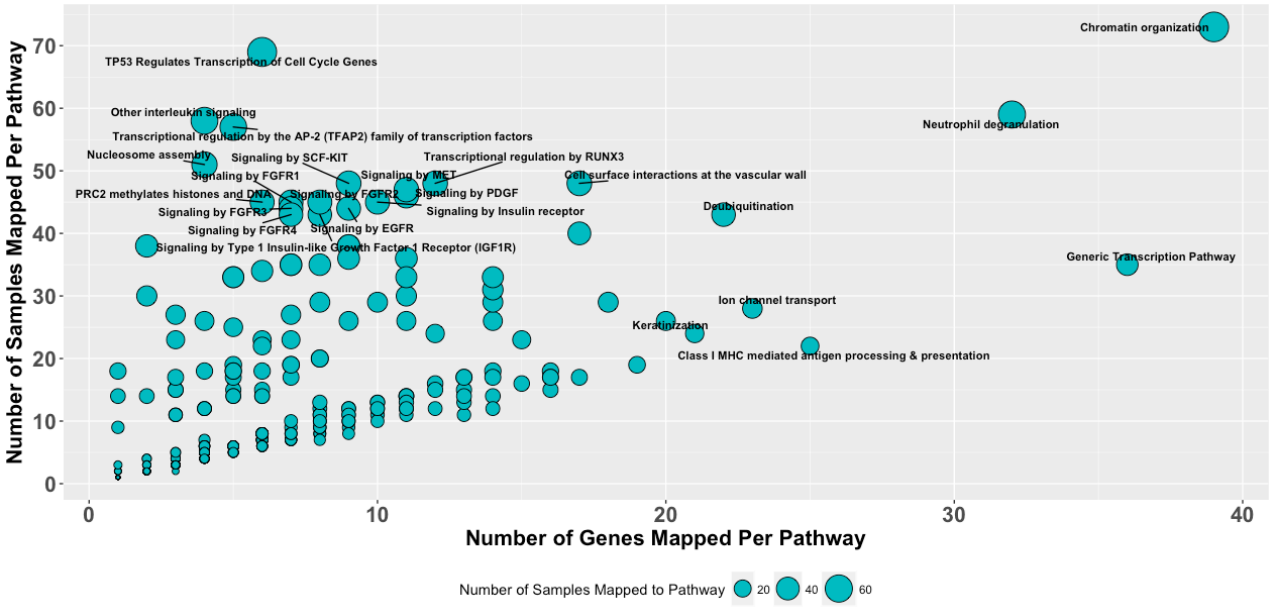


Figure 3. Pathway Coverage for Gene Mutations in 187 AML De Novo Patient Samples. Size of circle indicates number of samples in cohort mapped to pathway.

Pathway Enrichment Analysis

Pathway Enrichment by Gene

We conducted pathway enrichment for the mutated genes of the AML de novo patient samples cohort. Pathways enriched for mutated genes are those pathways with a higher number of mutated genes than would be expected by

chance. Overall, we found three pathways to be significantly enriched for mutated genes, after correcting for multiple testing: ‘Chromatin organization’ (FDR < 0.05) ‘Collagen biosynthesis and modifying enzymes’ (FDR < 0.10), ‘and ‘Sumolyation of DNA damage response and repair proteins’ (FDR < 0.10) (Figure 4). Thus, for only three pathways do we find that the number of genes mutated in the pathway is higher than what we would expect to find by chance.

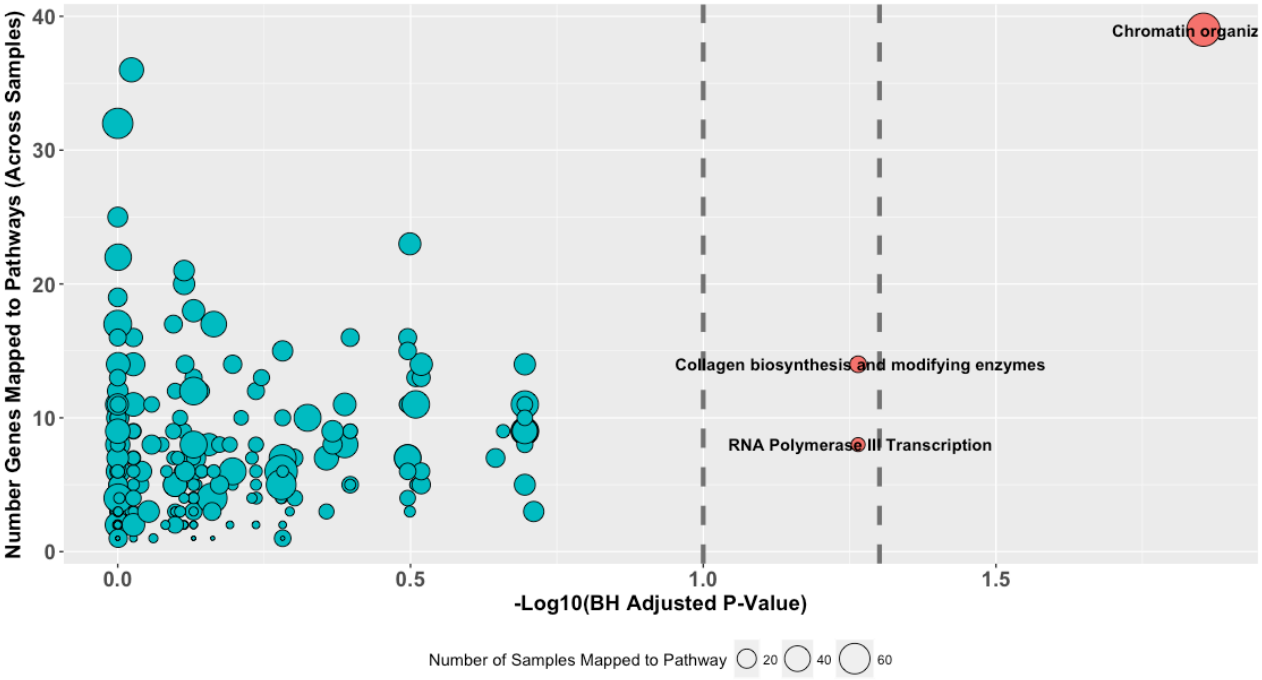


Figure 4. Pathway Enrichment by Gene for 187 AML De Novo Patient Samples. Enriched pathways passing multiple testing (Benjamini-Hochberg (BH) procedure, FDR<0.10 or FDR<0.05) are shown in red and labeled by name.

Pathway Enrichment by Sample

Alternatively, we can also conduct pathway enrichment from a sample-centric perspective. For this analysis, we assess which pathways have a higher number of samples mapped than we would expect at random. By taking a sample-centric

perspective for pathway enrichment, we can identify pathways with many samples who may share a small subset of mutated genes or pathways with many samples that do not share mutated genes. In both of these cases, such pathways would be of interest for understanding AML de novo dysregulated pathways, but may not have a large enough number of overall mutated genes to qualify as enriched from the gene perspective. This analysis reveals over 40 pathways that are significantly enriched with respect to the number of samples per pathway (Figure 5). This is likely due a small number of genes that are mutated with high frequency in the cohort of patient samples.

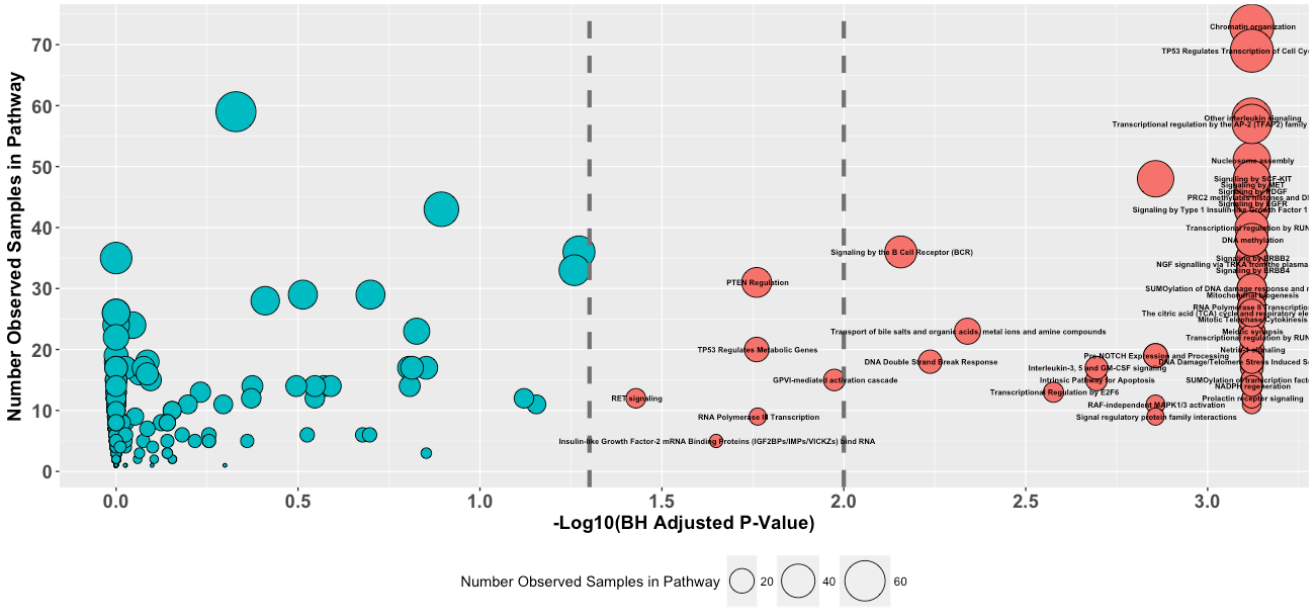


Figure 5. Pathway Enrichment by Sample for 187 AML De Novo Patient Samples. Enriched pathways passing multiple testing (FDR<0.05) are shown in red and labeled by name. Pathways stacked to the right-most part of the graph all have the same adjusted p-value due to the correction method used.

Assessing the Relationship Between Pathway Mutational Status and Drug Response

Grouping mutations into functionally related pathways may shed insight into shared dysregulated pathways among AML de novo patients. In particular, we are interested in whether shared dysregulation at the pathway level can inform our understanding of drug sensitivity and resistance to targeted therapies. Here, we investigate the relationship between pathway mutation status and drug sensitivity or analysis. For pathways and drug combinations, we conduct an ANOVA analysis between pathway mutation status and drug AUC. To do so, we group samples according to pathway mutation status (yes or no), and then assess whether there is a difference in the mean drug AUC between the two groups. If the mean drug AUC is higher for the group with pathway mutations, then that pathway is associated with resistance to the drug, while if the mean drug AUC is lower for the group with pathway mutations, then that pathway is associated with sensitivity to the drug.

Our analysis finds many significant pathway and drug associations for drug sensitivity and resistance. In Figure 6 each point on the volcano plot represents one pathway and drug interaction, where size indicates the number of samples with a pathway mutation, and color indicates significance after multiple testing correction (blue indicates FDR<0.05). On the right-hand side of the pot, significant associations represent relationships between pathway mutational status and higher drug AUC, or drug resistance. On the left-hand side of the plot,

significant associations represent a relationship between pathway mutational status and lower drug AUC, or drug sensitivity.

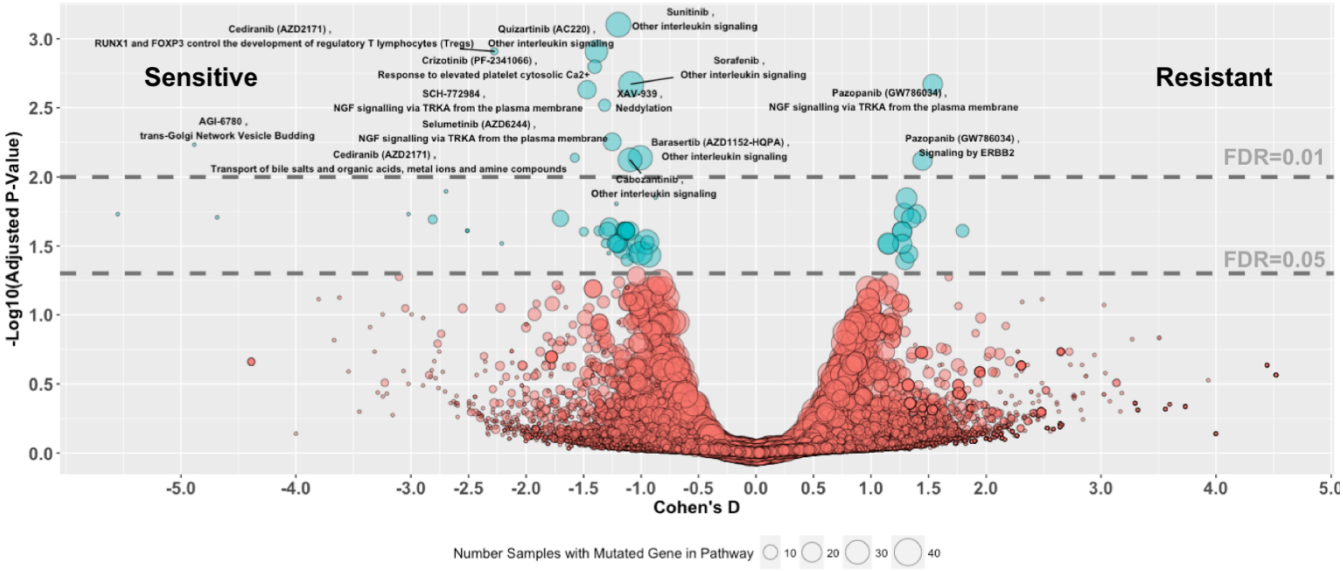


Figure 6. Pathway-Level Analysis Reveals Significant Associations with Drug Sensitivity or Resistance in 108 De Novo Acute Myeloid Leukemia Patient Samples. For each drug and pathway combination, Analysis of Variance (ANOVA) was conducted to compare the mean drug AUC between samples with mutations in pathway and samples without mutations in pathway. A negative Cohen's D indicates samples with mutation in pathway have a lower mean drug AUC and are more sensitive, while a positive Cohen's D value indicates samples with mutation in pathway have a higher mean drug AUC and are more resistant. Adjusted P-values are FDR corrected. Blue indicates adjusted p-value is significant at FDR<0.05. Circle size indicates the number of samples with one or more gene mutations in pathway.

All significant pathway and drug interactions shown in blue in the above plot are collected in Table 4 and Table 5, for sensitive and resistance interactions respectively. A total of 22 pathways were found to have sensitive interactions with 29 different drugs. Notably, for sensitive interactions we found sunitinib, quizartinib, sorafenib, barasertib, and cabozantinib as high-ranking drugs to interact with the pathway 'Other Interleukin Signaling'. These drugs are all inhibitors for the FLT3 protein, a receptor tyrosine kinase, which is contained in the pathway 'Other Interleukin Signaling' and has critical mutations for AML

[26,29]. Given that these are established drug and gene relationships, we would expect that we would still be able to detect this association through the pathway which contains FLT3. Thus, these associations serve as a form of positive control, indicating that our analysis of pathway and drug interactions is on target.

For resistance pathway and drug interactions, we found a total of 11 pathways to have resistant interactions with 4 different drugs: pazopanib, NVP-ADW742, vatalanib, and the combination of inhibitors sorafenib and panobinostat. In particular, we will highlight the interaction between the inhibitor NVP-ADW742 and the pathway ‘Signaling by SCF-Kit’ (Figures 7-8). We note that 6 pathways found to be sensitive to drugs were also involved in resistance interactions (with other drugs).

Table 4. Significant Sensitive Pathway and Drug interactions from ANOVA for Pathway Mutational Status and Drug AUC.

No.	*Pathway*	*Drug*	*P-Value*	*Adjusted P-Value*	*Cohen's D*	*N Samples Without Pathway Mutation*	*N Samples With Pathway Mutation*
1	Other interleukin signaling	Sunitinib	8.49E-08	0.000790933	-1.197337354	68	33
2	Other interleukin signaling	Quizartinib (AC220)	1.82E-07	0.001233017	-1.389363685	62	27
3	RUNX1 and FOXP3 control the development of regulatory T lymphocytes (Tregs)	Cediranib (AZD2171)	1.55E-07	0.001233017	-2.27580135	86	3
4	Response to elevated platelet cytosolic Ca2+	Crizotinib (PF-2341066)	2.83E-07	0.001595138	-1.4037121	91	9
5	Other interleukin signaling	Sorafenib	4.70E-07	0.002129022	-1.086528905	68	34
6	NGF signalling via TRKA from the plasma membrane	SCH-772984	1.00E-06	0.002336443	-1.469648831	58	17
7	Neddylation	XAV-939	1.42E-06	0.003025776	-1.317386852	84	7
8	NGF signalling via TRKA from the plasma membrane	Selumetinib (AZD6244)	2.99E-06	0.005569156	-1.251232736	74	16
9	trans-Golgi Network Vesicle Budding	AGI-6780	3.22E-06	0.005846631	-4.884025013	60	2
10	Other interleukin signaling	Barasertib (AZD1152-HQPA)	4.19E-06	0.007265732	-1.007475474	59	32
11	Transport of bile salts and organic acids, metal ions and amine compounds	Cediranib (AZD2171)	4.17E-06	0.007265732	-1.577362731	85	4
12	Other interleukin signaling	Cabozantinib	4.58E-06	0.007521686	-1.097564784	58	30
13	trans-Golgi Network Vesicle Budding	AUS-131	8.70E-06	0.012714891	-2.694703306	66	2
14	Mitochondrial translation	Vismodegib (GDC-0449)	1.01E-05	0.014130477	-0.873878647	89	2
15	MyD88 cascade initiated on plasma membrane	PHT-427	1.15E-05	0.015630027	-1.216068452	86	2
16	The citric acid (TCA) cycle and respiratory electron transport	SJB-66	1.45E-05	0.018537723	-5.549692943	21	2
17	trans-Golgi Network Vesicle Budding	AGI-5198	1.47E-05	0.018537723	-3.021425285	58	2
18	trans-Golgi Network Vesicle Budding	Azacytidine	1.58E-05	0.01960412	-4.684935281	65	2
19	Other interleukin signaling	AZD2811(Aurora)	1.68E-05	0.019940597	-1.701590469	22	13

20	Response to elevated platelet cytosolic Ca2+	AZD6378 (ATR) - Cytarabine	1.74E-05	0.02028672	-2.810100932	31	4
21	Fc epsilon receptor (FCERI) signaling	SCH-772984	2.07E-05	0.022973975	-1.274693654	58	17
22	B-WICH complex positively regulates rRNA expression	KU-55933	3.38E-05	0.024530131	-1.365876592	88	5
23	DAP12 interactions	CI-1040 (PD184352)	2.68E-05	0.024530131	-1.135491645	76	16
24	DAP12 interactions	Selumetinib (AZD6244)	2.25E-05	0.024530131	-1.152269676	74	16
25	Gastrin-CREB signalling pathway via PKC and MAPK	SCH-772984	2.76E-05	0.024530131	-1.29186204	60	15
26	Gastrin-CREB signalling pathway via PKC and MAPK	Selumetinib (AZD6244)	3.66E-05	0.024530131	-1.13103076	75	15
27	NGF signalling via TRKA from the plasma membrane	CI-1040 (PD184352)	2.87E-05	0.024530131	-1.132874844	76	16
28	Signaling by ERBB4	Selumetinib (AZD6244)	3.66E-05	0.024530131	-1.13103076	75	15
29	SIRT1 negatively regulates rRNA expression	KU-55933	3.21E-05	0.024530131	-2.510439983	91	2
30	Striated Muscle Contraction	KU-55933	3.21E-05	0.024530131	-2.510439983	91	2
31	Signaling by ERBB2	Selumetinib (AZD6244)	3.69E-05	0.024575577	-1.101780723	73	17
32	Signaling by NODAL	H-89	3.80E-05	0.024877089	-1.498419899	89	4
33	Other interleukin signaling	Foretinib (XL880)	4.57E-05	0.029593817	-0.951578827	60	32
34	Response to elevated platelet cytosolic Ca2+	Roscovitine (CYC-202)	4.68E-05	0.029823878	-0.944915185	84	9
35	DAP12 interactions	Trametinib (GSK1120212)	5.00E-05	0.030286586	-1.202634091	80	17
36	NoRC negatively regulates rRNA expression	KU-55933	4.85E-05	0.030286586	-1.30689706	89	4
37	Signaling by ERBB2	SCH-772984	4.95E-05	0.030286586	-1.217110007	59	16
38	TAK1 activates NFkB by phosphorylation and activation of IKKs complex	Motesanib (AMG-706)	4.97E-05	0.030286586	-2.210843772	90	2
39	NGF signalling via TRKA from the plasma membrane	Trametinib (GSK1120212)	5.07E-05	0.030431519	-1.2019734	80	17
40	NCAM signaling for neurite out-growth	Selumetinib (AZD6244)	5.25E-05	0.0305911	-1.040831622	73	17
41	Fc epsilon receptor	Trametinib (GSK1120212)	5.80E-05	0.033258662	-1.161293816	78	19

	(FCERI) signaling						
42	Fc epsilon receptor (FCERI) signaling	Selumetinib (AZD6244)	6.41E-05	0.035670112	-1.045382639	72	18
43	Other interleukin signaling	Pelitinib (EKB-569)	6.41E-05	0.035670112	-1.002252862	60	29
44	trans-Golgi Network Vesicle Budding	XAV-939	6.34E-05	0.035670112	-1.281467926	89	2
45	Other interleukin signaling	17-AAG (Tanespimycin)	6.74E-05	0.036946074	-0.932015658	61	30
46	Neddylation	H-89	7.45E-05	0.03993543	-1.122094255	86	7

Table 5. Significant Resistant Pathway and Drug Interactions from ANOVA Analysis for Pathway Mutational Status and Drug AUC.

No.	*Pathway*	*Drug*	*P-Value*	*Adjusted P-Value*	*Cohen's D*	*N Samples without Pathway Mutation*	*N Samples with Pathway Mutation*
1	NGF signalling via TRKA from the plasma membrane	Pazopanib (GW786034)	8.00E-07	0.002129022	1.535636359	76	20
2	Signaling by ERBB2	Pazopanib (GW786034)	4.90E-06	0.007624316	1.446350125	77	19
3	Signaling by PDGF	NVP-ADW742	1.03E-05	0.014219381	1.308613728	65	22
4	NCAM signaling for neurite out-growth	Pazopanib (GW786034)	1.37E-05	0.018254285	1.285165561	76	20
5	Gastrin-CREB signalling pathway via PKC and MAPK	Pazopanib (GW786034)	1.45E-05	0.018537723	1.395731366	78	18
6	Signaling by ERBB4	Pazopanib (GW786034)	1.69E-05	0.019940597	1.347271878	77	19
7	NGF signalling via TRKA from the plasma membrane	Sorafenib - Panobinostat	2.35E-05	0.024530131	1.793740632	28	8
8	DAP12 interactions	Pazopanib (GW786034)	3.55E-05	0.024530131	1.269104134	77	19
9	Signaling by FGFR1	NVP-ADW742	3.81E-05	0.024877089	1.269215689	67	20
10	Signaling by SCF-KIT	NVP-ADW742	4.85E-05	0.030286586	1.148568855	64	23
11	C-type lectin receptors (CLRs)	Pazopanib (GW786034)	5.17E-05	0.03054905	1.151816387	75	21
12	Signaling by FGFR4	NVP-ADW742	5.16E-05	0.03054905	1.270037362	68	19
13	NGF signalling via TRKA from the plasma membrane	Vatalanib (PTK787)	6.53E-05	0.036059593	1.32873851	77	16
14	NCAM signaling for neurite out-growth	Vatalanib (PTK787)	7.58E-05	0.040313089	1.293988723	76	17

When we examine individual significant pathway and drug associations, we find that they are driven by a spectrum of genes that are mutated within the pathway. For instance, the interaction between ‘Signaling by SCF-KIT’ and the inhibitor NVP-AD1742 is a significant pathway association (adjusted p-value = 0.03) with drug resistance (Figure 7). Examining the 23 samples with mutated genes in this pathway reveals a total of 8 different genes that are mutated. Genes mutated in multiple samples include NRAS (9 samples NRAS alone, 2 with KRAS), PTPN11 (5 samples) and KRAS (2 samples KRAS alone, 2 with NRAS). Additionally, we also observe mutations in CBL, JAK2, KIT, and STAT5B. In Figure 7 we can see that the AUC for samples with these mutations are grouped together fairly closely and are much higher compared to the bulk of the AUC values for samples with no mutations in this pathway. Notably, for each of these genes there is a small number of samples mutated, and therefore we would not necessarily be able to assess whether there is an association between gene mutational status and drug sensitivity or resistance, as described below.

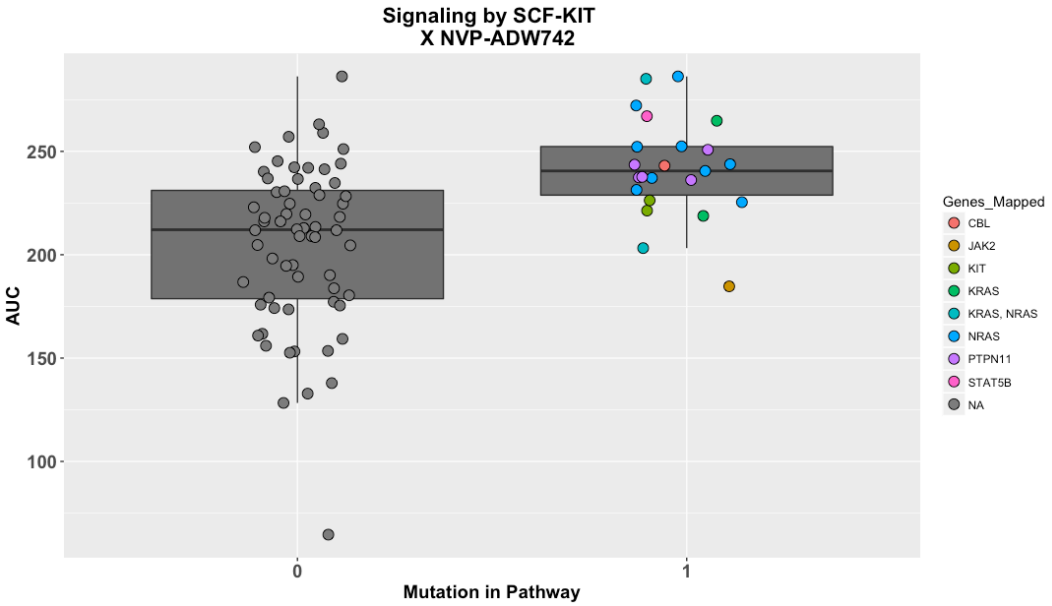


Figure 7. Aberrational Pathway Status of Signaling by SCF-KIT is Associated with Increased Resistance to the Inhibitor NVP-ADW742. Samples are plotted according to whether they have a mutation in the pathway 'Signaling by SCF-KIT' and corresponding area under the curve (AUC) for the investigational stage drug NVP-ADW742. Color in right-hand boxplot indicates mutated gene(s) for sample.

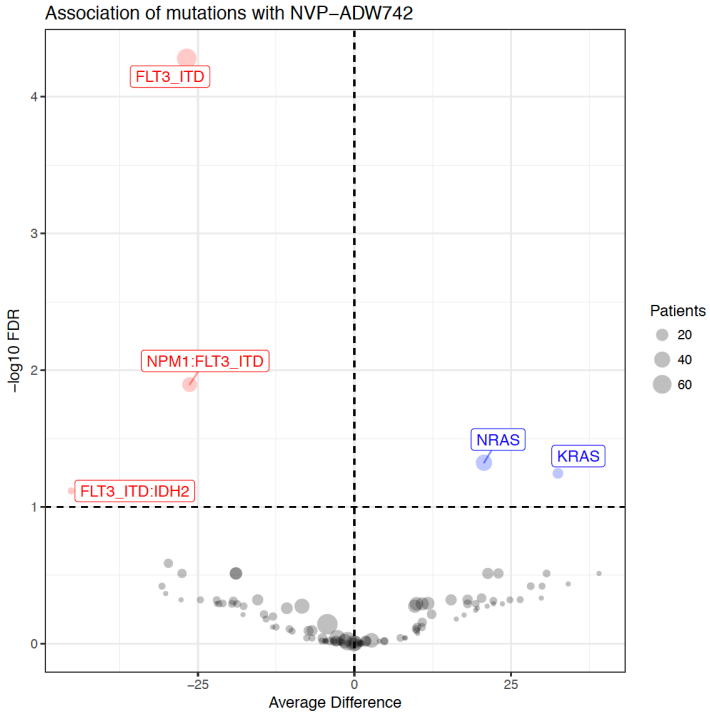


Figure 8. NVP-ADW742 and Gene Associations. Inhibitor by gene associations for full cohort of AML patient samples, paired and tumor only mutation samples, not restricted to AML de novo. Volcano Plot generated in Vizome, by Daniel Bottomly, Dr. Beth Wilmot, and Dr. Shannon McWeeney.

Associations between drug AUC for the inhibitor NVP-ADW742 and mutated genes are shown in Figure 8. While NRAS and KRAS have an association with drug resistance for this inhibitor, we can see that the other genes belonging to the pathway 'Signaling by SCF-KIT' do not appear to be significantly associated with resistance to this inhibitor. For example, the gene PTPN11 was mutated in 5 patient samples and does not appear to be significantly associated with resistance when considered as a single gene.

The pathway perspective allows us to aggregate multiple mutations together and find associations between drug response and the pathway unit. Such associations may be missed when considering only single gene interactions and drug response.

Modeling Pathway Impact with Probabilistic Graphical Modeling Framework

Given that we have identified significant interactions between pathway-level dysregulation and drug sensitivity or resistance, we now want to investigate the relationship in a more mechanistic manner. We want to be able to consider the impact of individual gene mutations, and develop a corresponding measure of impact. In this way, we can start to assess how mutations in the same pathway are similar or different, and what the respective effects will be on drug response.

Mutational Impact on Pathways

Summary of Mutation Impacted Pathways, Cohort Level

To further investigate how individual gene-level mutations impact pathway signaling, we leverage a probabilistic graphical modeling framework to model pathways from Reactome and the resulting impact of gene-level mutations. This work is an adaptation of pathway impact modeling as proposed in the PARADIGM computational framework [37] and leverages a suite of graphical model construction from Reactome pathways and accompanying message passing algorithms developed in Java by Wu et al. (2014) [121].

For 187 AML de novo samples, we modeled the mutation impact on 291 Reactome pathways [25]. Each pathway in Reactome is converted into a factor graph representation, where nodes are biological entities (e.g. proteins) and factor functions describe the relationships between nodes. Each node can be in one of two states: (0) not impacted or (1) impacted. For each of these states, the node has an associated probability of occupying the state, hence the term probabilistic graphical modeling. For each pathway factor graph, a null perturbation run of message passing is conducted with no observed mutations. This establishes the baseline belief states for each entity in the pathway in the absence of any individual patient sample information. Then, for each patient sample, mutation information is loaded onto the corresponding protein node and factor functions designates the impact on the nodes in relation to the mutations' functional impact score. The message passing algorithm 'loopy belief propagation' [38], is used to propagate the effect of the impact throughout the

rest of the pathway. For each pathway entity, we can then calculate the log odds the entity being in the impacted state for the mutation run compared to being in the impacted state for the null run. This provides a per-entity impact score. Averaging the entity-level scores across all entities in the pathway produces a pathway-impact score, or average log odds of mutational impact. This score is used in following figures and discussion of PGM pathway impact. Here we note that while a total of 431 pathways were first transformed into factor graph representations, only 291 of these pathways were found to be “stable” during the null perturbation run, i.e., convergence of belief states was reached during message passing on the pathway factor graph without patient mutation information. See Methods section on PGM Pathway Stability Assessment for further discussion of pathway stability analysis.

Top Mutation Impacted Pathways in De Novo AML

In Figure 9, we show the top mutation impacted pathways by degree of impact and number of samples (Figure 9). Across the cohort of AML de novo samples, we see that some pathways are impacted in high number of samples, while other pathways are highly impacted but for a small number of samples. For instance, the pathway ‘Neutrophil degranulation’ is impacted in approximately 60 samples, and on average samples have a log odds mutation impact of about 4.0. Other pathways with a high number of samples mutated include: ‘Signaling by Type 1 Insulin-like Growth Factor 1 Receptor (IGF1R)’, ‘Generic Transcription Pathway’, ‘Other Interleukin Signaling’, ‘Ion Channel Transport’, and ‘Regulation of RUNX1 Expression and Activity’. Pathways with a high average log odds

impact (across samples) include: ‘G alpha (q) signaling events’, ‘G alpha iii signaling events’, ‘Post translational protein phosphorylation’, and ‘Transcription of DNA Repair Genes’.

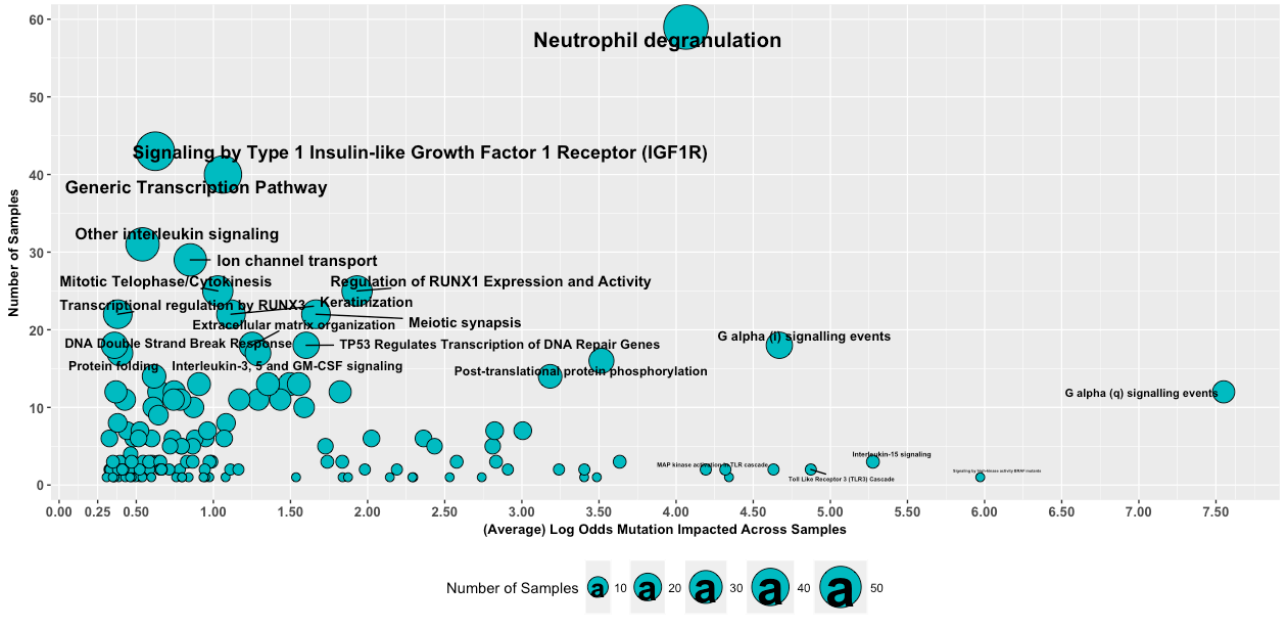


Figure 9. PGM Mutation Impacted Pathways for 187 AML De Novo Samples. Circle size indicates number of samples with pathway impacted.

Variability in Sample Mutational Pathway Impact

However, representing pathway impact as an average across samples does not reflect all the information available. In particular, we would expect that samples harboring different gene mutations would have different pathway impact scores for the sample pathway. In Figure 10, we show the top mutation-impacted pathways (requiring at least 5 samples with impact in the corresponding pathway) ordered by the variability across impacted samples. Here we can see that pathways such as ‘Interleukin-2 family signaling’, ‘Signaling by SCF-KIT’, ‘NCAM signaling for neurite out-growth’, ‘Amyloid fiber formation’, ‘Interleukin-6

signaling', and 'RET Signaling' show a large range between the minimum and maximum pathway impact calculated across samples. In contrast, the pathway 'Neutrophil degranulation', which is impacted in almost 60 samples, shows very low variability, as most samples exhibit pathway impact between a log odds impact of 3.75 and 4.5.

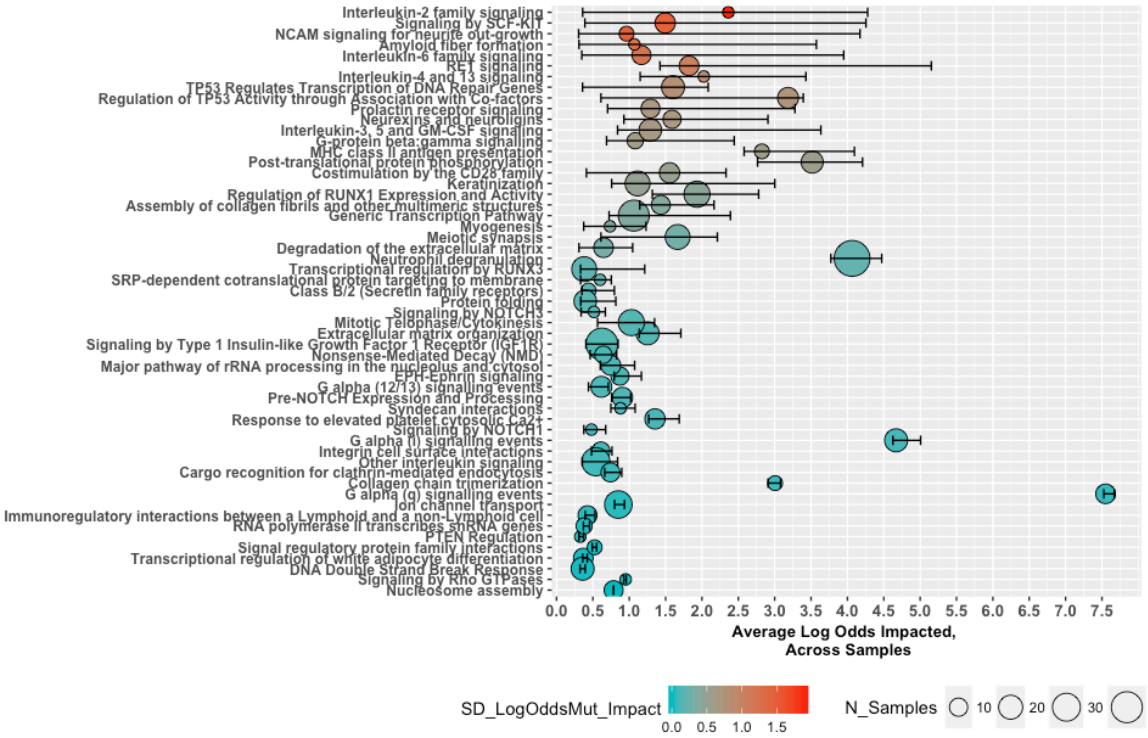


Figure 10. PGM Mutation Impacted Pathways for 187 AML De Novo Patient Samples, by Variability in Pathway Impact. Ordering pathways by variability highlights the range of pathway impact inferred within sets of samples with dysregulation in the same pathway. Pathways shown are impacted for at least 5 AML de novo patient samples.

Pathway Impact by Mutated Gene

Examining mutational impacted pathways from a by-gene perspective allows us to assess the contribution of mutations in individual genes to overall pathway impact. Assessing the pathways impacted in at least 10 samples from this perspective reveals three key trends.

First, in some pathways, we see that there are one or two genes that (when mutated) result in a much higher pathway impact than other genes. By capturing this, our modeling framework allows us the potential for investigating the impact of mutations in different genes on the same pathway. In Figure 11, we show the mutational impact for 48 samples with mutations in the pathway 'Signaling by SCF-KIT'. Three samples with mutations in the gene KIT stand out as having a much higher overall pathway impact (3.5-4.0). KIT is much further upstream in this pathway than many of the other mutated genes (JAK2, STAT5B, NRAS, KRAS, etc.) and thus a mutation affecting KIT is propagated throughout the rest of the pathway. Mutations in JAK2, STAT5B, and PTPN11 (with the exception of one sample) have moderate impact when compared to KIT (log odds: .75-1.25). By comparison, the genes KRAS and NRAS are some of the most downstream genes in the pathway, so mutations in these genes have a much smaller impact (less than 0.25) due to a small number of downstream entities being affected.

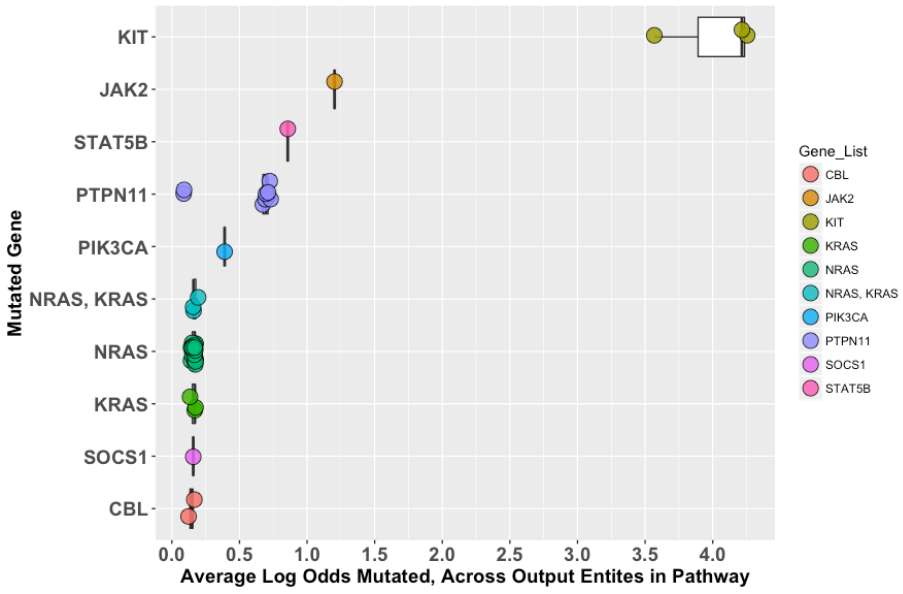


Figure 11. SCF-KIT Pathway Impact Ascross 48 AML De Novo Samples. Each circle represents one AML de novo sample with the respective pathway impacted. Color indicates genes mutated for that sample in pathway.

In another example (Figure 12), we see that mutations in TP53 have a high impact on the pathway 'TP53 Regulates Transcription of DNA Repair Genes' compared to other mutated genes. In this pathway, TP53 is a central component, and thus mutations in the pathway propagate to many other entities. By comparison, the genes CCNH, ATM, and GTF2F1 are much more distal entities and have a small impact on the pathway when mutated.

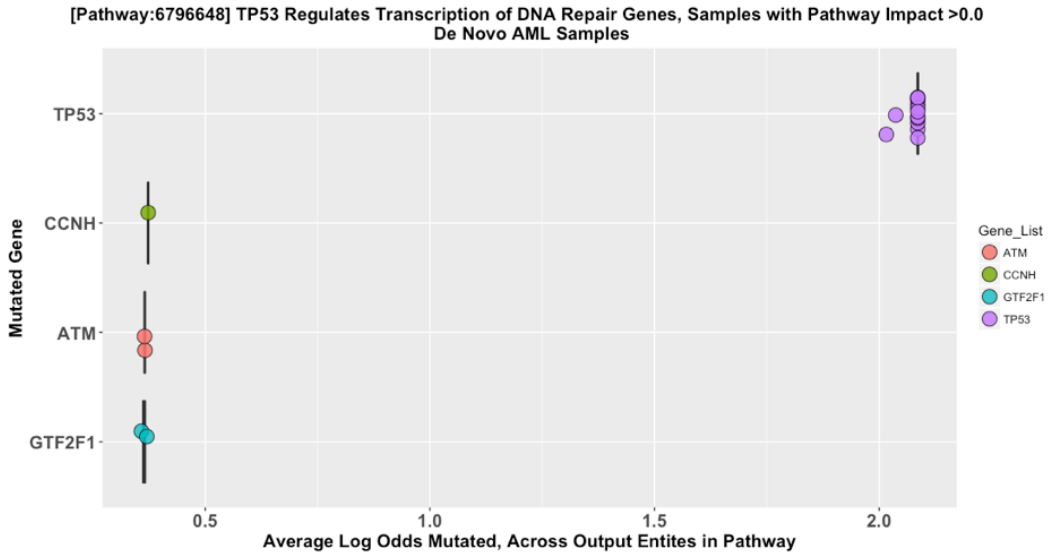
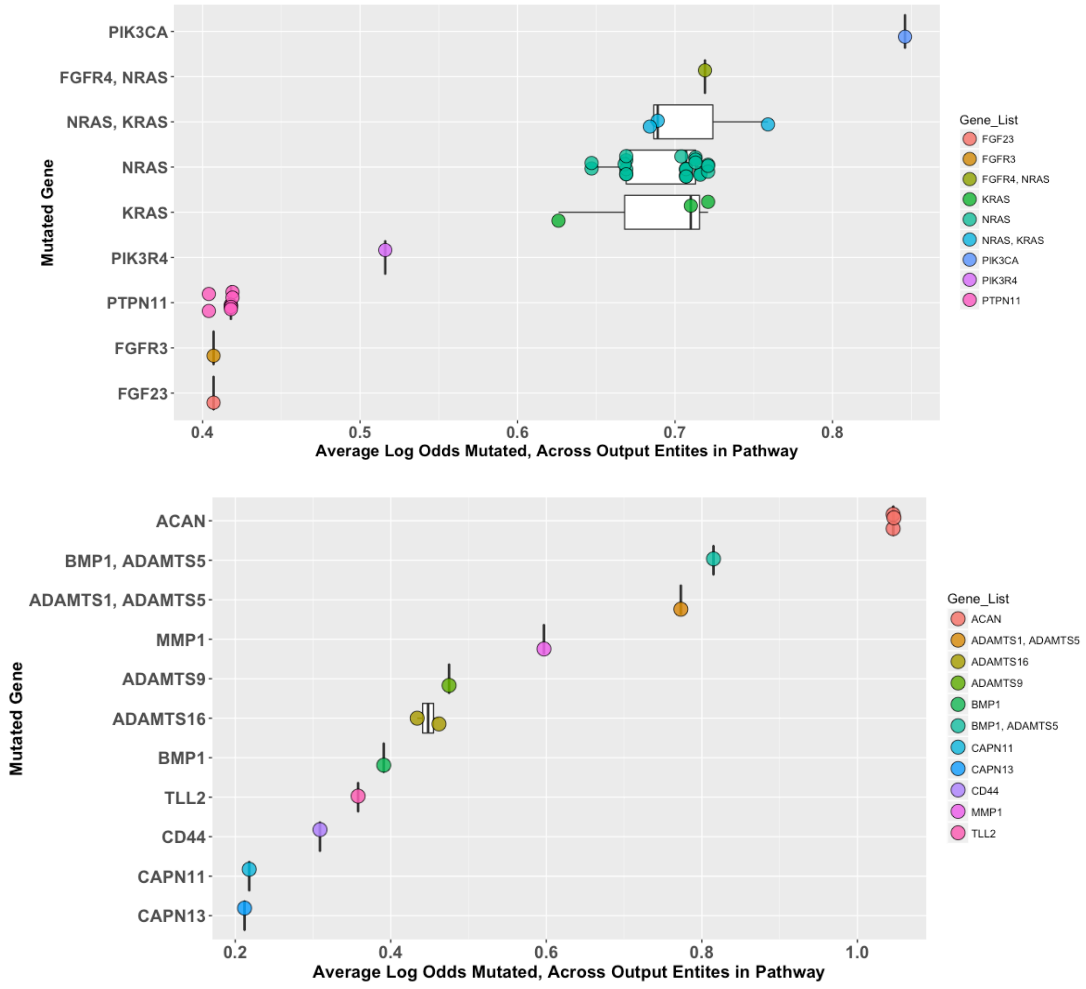


Figure 12. TP53 Regulates Transcription of DNA Repair Genes Across AML De Novo Samples. Each circle represents one AML de novo sample with the respective pathway impacted. Color indicates genes mutated for that sample in pathway.

Second, in some pathways, based on observed mutations in AML de novo samples, there is a slight difference in the impact incurred by mutations in different genes, but no single gene stands out as more strongly impactful on the pathway than others. For instance, in Figure 13A, the pathway 'Signaling by Type1 Insulin-like growth factor 1 receptor (IFG1R)', there is a slight increase in impact caused by genes but no genes that when mutated cause a much larger

pathway impact. This same trend is repeated for the pathways ‘Degradation of the extracellular matrix’, and ‘Neutrophil degranulation (Figure 13 B and C)’. We note that it is possible that other genes may be more impactful (to the pathway) when mutated, but we do not observe any such gene mutations in our particular cohort of AML de novo paired samples.



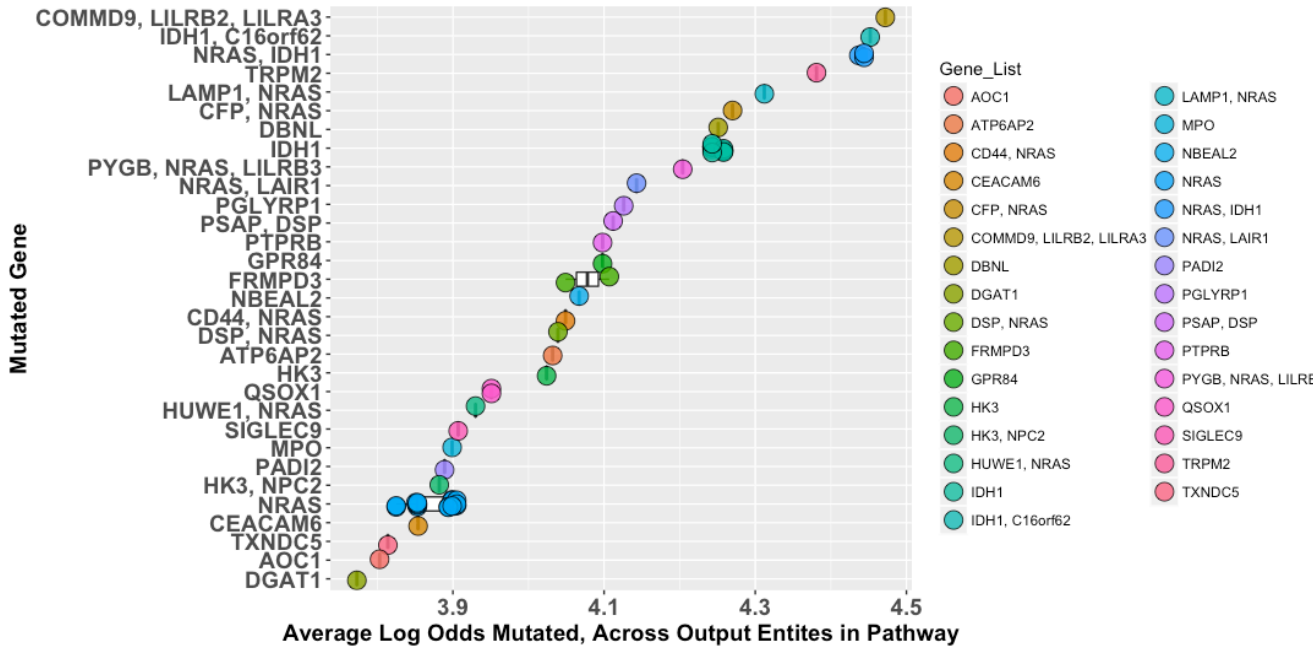


Figure 13. A. Pathway Impact for Samples with Dysregulation in 'Signaling by Type 1 Insulin-like Growth Factor 1 Receptor (IGF1R). B. Pathway Impact for Samples with Dysregulation in 'Degradation of the extracellular matrix'. C. Pathway Impact for Samples with Dysregulation in 'Neutrophil degranulation'. Each circle represents one AML de novo patent sample. Color indicates mutated genes for that sample in the pathway. Co-occurring mutations for the same sample are indicated.

Third, in some cases we are able to see evidence that the probabilistic graphical modeling framework can potentially capture cumulative mutational impact of multiple co-occurring mutations. For a select few impacted pathways, where a few samples actually have co-occurring mutations, we can see that the resulting pathway impact is higher than for the samples with only one of the mutations alone. For instance, for the pathway 'Regulation of RUNX1 Expression and Activity' (Figure 14A), we have a sample with both RUNX1 and CCND3 mutated that shows much higher pathway impact than samples with only RUNX1 mutated. However, we note that our cohort does not currently include samples with only CCND3 mutated, which would allow us to assess if the co-occurrence of mutations in RUNX1 and CCND3 truly result in a higher impact that either

gene mutation alone. In another pathway, ‘Mitotic Telophase/Cytokinesis’ (Figure 14B), samples with mutations in PDS5B, SMC1A, RAD21, STAG2, or SMC all have an overall pathway impact around 1.0. We have one sample for each of the following co-occurring mutations sets: PDS5B and SMC3, SMC3 and STAG2, RAD21 and SMC3, which in all cases have a higher pathway impact than samples with only one of the mutations in the set.

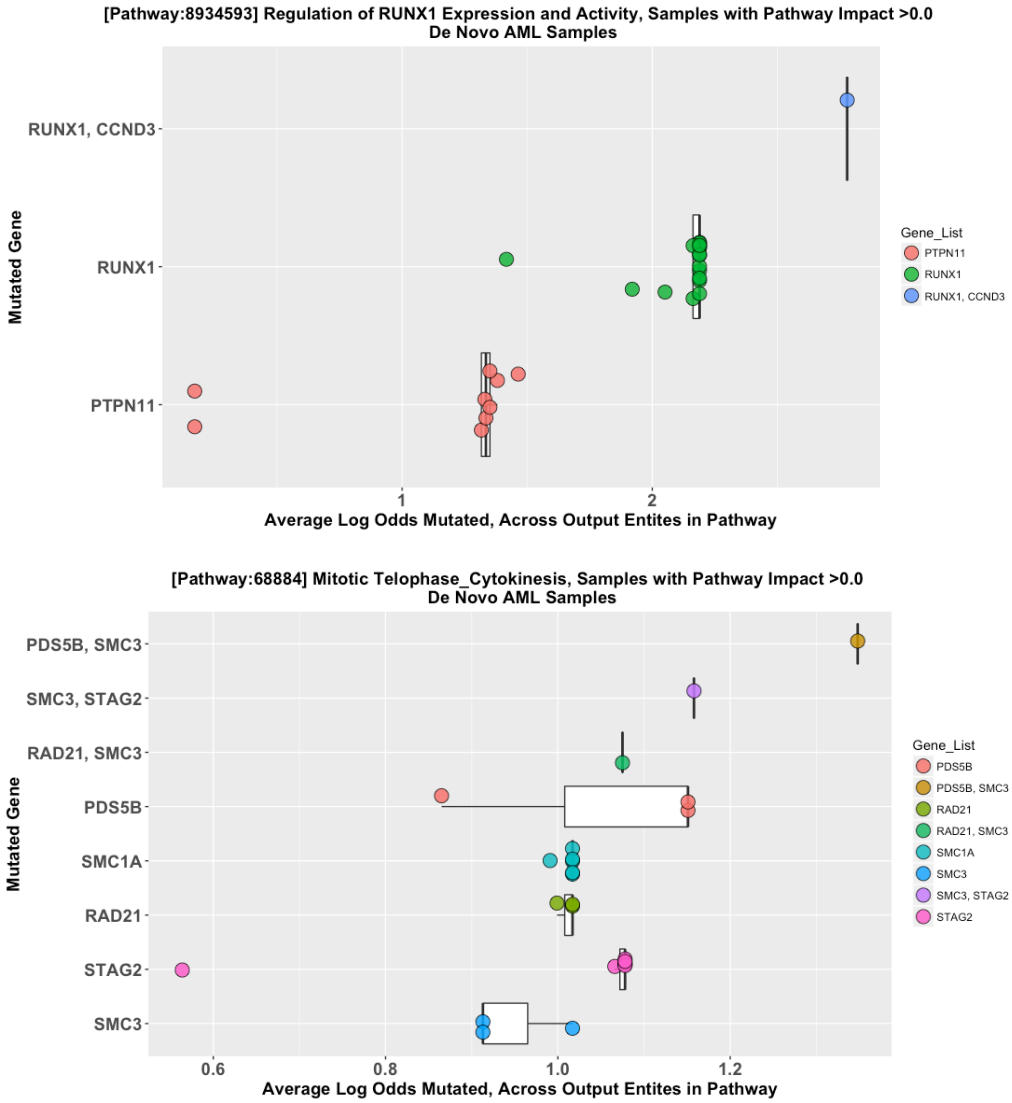


Figure 14. A. Pathway Impact for Samples with Dysregulation in ‘Regulation of RUNX1 Expression and Activity’. B. Pathway Impact for Samples with Dysregulation in ‘Mitotic Telophase and Cytokinesis. Each circle represents one AML de novo patent sample. Color indicates mutated genes for that sample in the pathway.

However, we note that the examples discussed here for increased pathway impact for co-occurring mutations are each for only one sample and require further investigation with larger sample sizes. In particular, we note the opportunity here for further computational modeling, especially simulation of gene mutations to assess cumulative impact of more than one gene mutation in the same pathway. Given both the sparsity of mutations in AML de novo samples and the overall heterogeneity across cohort samples, this is encouraging for future efforts with this type of modeling. In particular, this showcases the potential for using this framework for use cases where samples have multiple co-occurring mutations in the same pathway, or cases where co-occurring mutations are known to confer a difference in drug response.

Mutated Gene Impact – Across Pathways

We can also look at how the impact of a mutated gene on a pathway differs by pathway, as many of the mutated AML genes belong to multiple curated pathways in Reactome. Here we required that genes be mutated in at least 3 samples, for a total of 29 genes. For each gene, we assess pathway impact across all pathways in which that gene is a member. We highlight several examples where the gene belongs to more than one pathway but has very different impact on each pathway. We note that many of the high frequency mutated genes are curated to only one pathway.

In Figure 15, we show a TP53-centric plot which highlights the pathway impact of gene mutations in TP53 across all pathways in which TP53 has

membership. For the nine total pathways, 6 pathways show relatively low impact when TP53 is mutated: ‘Regulation of TP53 Activity through Phosphorylation’, ‘Transcriptional regulation by RUNX3’, ‘PTEN Regulation’, TP53 Regulated Transcription of Cell Death Genes’, ‘Protein Folding’, and ‘DNA Double Strand Break Response’. By comparison, the pathways ‘Pre-Notch Expression and Processing’, ‘TP53 Regulates Transcription of DNA repair Genes’, and ‘Regulation of TP53 Activity through Association with Co-factors’ show much higher impact.

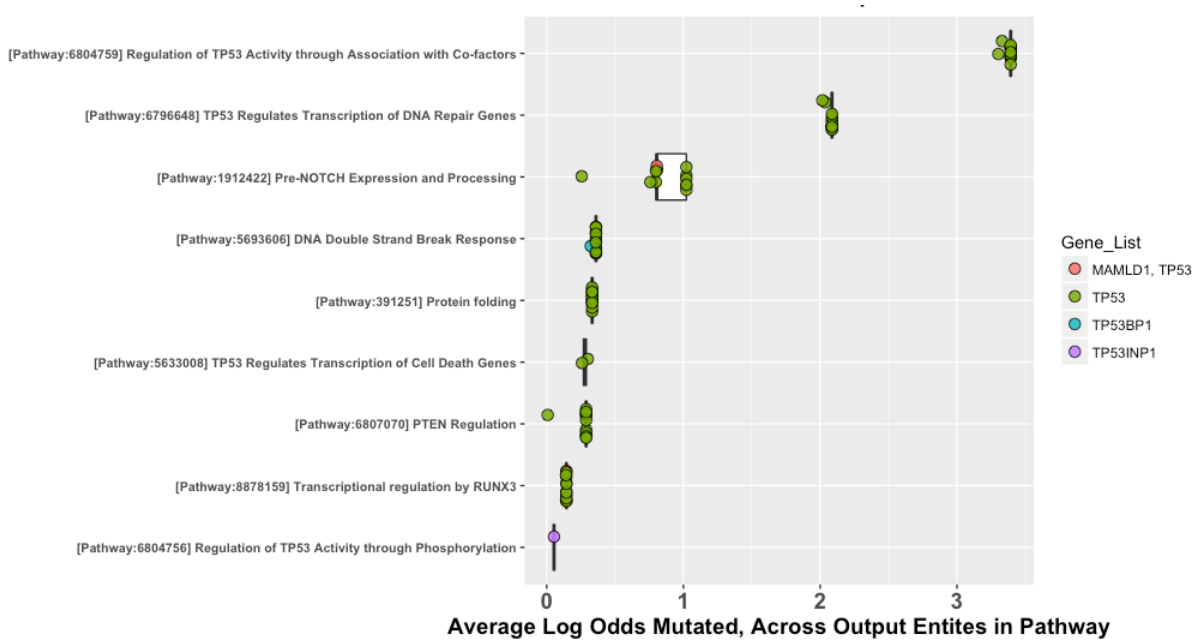


Figure 15. Impact Across Pathways for the Gene TP53. Each circle represents one AML de novo patent sample. Color indicates mutated genes for that sample in the pathway. Co-occurring mutations (in same pathway) indicated when applicable.

In another example (Figure 16), we show a PTPN11-centric plot which highlights the impact of a PTPN11 gene mutation on each of the 13 different pathways containing PTPN11. When PTPN11 is mutated, the pathways ‘RET Signaling’ and ‘Costimulation by the CD28 family’, ‘Prolactin receptor signaling’

and ‘Regulation of RUNX1 Expression and Activity, and ‘Interleukin-3,5 and GM-CSF signaling’ are more highly impacted that pathways such as ‘Netrin-1 signaling’ or ‘Interferon gamma signaling’.

Figure 16. Impact Across Pathways for the Gene PTPN11. Each circle represents one AML de novo patent sample. Color indicates mutated genes for that sample in the pathway. Co-occurring mutations indicated when applicable.

There are several reasons why we might expect the pathway impact of a gene to differ according to pathways. If the gene is one of the primary functioning entities of that pathway (i.e. the pathway is actually curated around relationships that all involve the gene), then we would expect a higher overall pathway impact to result from mutations in the gene. If the gene is less involved in the pathway (for example, it is involved in one reaction very downstream in the pathway), then we would expect an overall much lower resulting pathway impact. For instance, for the gene TP53, it’s involvement in the pathway ‘Transcriptional Regulation by RUNX3’ is fairly minor, and thus overall has a very low impact on the pathway. Additional reasons include connectedness of the pathway, as curation in

Reactome is an ongoing process and some pathways may have entities that are very well connected to each other while some pathways may be much more linear in their signal processing. Future work in this area can include a more robust assessment of a gene's impact on pathway and its location in pathway. This could include assessment of a node's connectivity, for instance, and correlation with pathway impact.

Drug Impact on Pathways

Summary of Drug Impacted Pathways, Across 32 BeatAML and Cancer Targetome Drugs

In addition to modeling the impact of individual patient mutations on pathways, we also want to be able to model the impact of drug application on pathways. To this end, we use the same modeling framework to model a set of drug-impacted pathways. We then discuss considerations for uniting drug-impacted pathways with mutation-impacted pathways to better understand patient-specific drug sensitivity and resistance.

For 32 drugs included on the BeatAML drug screening panel, we had target information in the Cancer Targetome [120]. We note that the majority of drugs used on the drug screening panel are investigational-stage and therefore not included in the Cancer Targetome, which currently covers only FDA-approved cancer drugs. For each drug, target binding assay information from the Targetome was used to model drug-target binding relationships. For this modeling work, we required drug-target evidence to be Level III and have a

minimum reported binding assay value <10,000nM, across any type of binding assay (IC50, EC50, Ki, or KD). For each drug-target relationship, the best or minimum reported assay value was used to represent the relationship. In the factor graph, drug-targeted nodes were set to a level of impact in accordance to the strength of minimum binding assay value retrieved from the Cancer Targetome (Table 5, Methods). Modeling drug impact on pathways in such a manner allows us to investigate in a more mechanistic manner the downstream activity of drug-target binding. Furthermore, by explicitly modeling drug-binding across multiple targets, we are able to take a truly polypharmacological, or multi-target perspective on drug binding. Additionally, we represent drug-target relationships in a way that reflects the strength of binding assay evidence. Thus, we are representing drug-target interactions within the pathway model in an empirical-driven manner.

Similar to the summary measure used for mutation-impacted pathways, we summarize each drug-pathway relationship with a single pathway impact score, which is the average log odds of pathway impact across all pathway entities. In Figure 17, we show the drug-impacted pathways, across the full set of drugs. Towards the top of the figure, we can see those pathways that are targeted by a large number of drugs, such as RAF/MAP kinase cascade (22), Signaling by ERRBB2 (18), NGF signaling via TRKA from the plasma membrane (18), and Signaling by SCF-KIT (18). Towards the bottom of the figure are pathways impacted by only a small number of drugs. Moving right on the graph shows us those pathways which have a high average log odds drug impact

(across drugs). This figure provides a summary overview of impacted pathways when considering the full set (32) of drugs on the BeatAML Panel for which we have information in the Cancer Targetome.

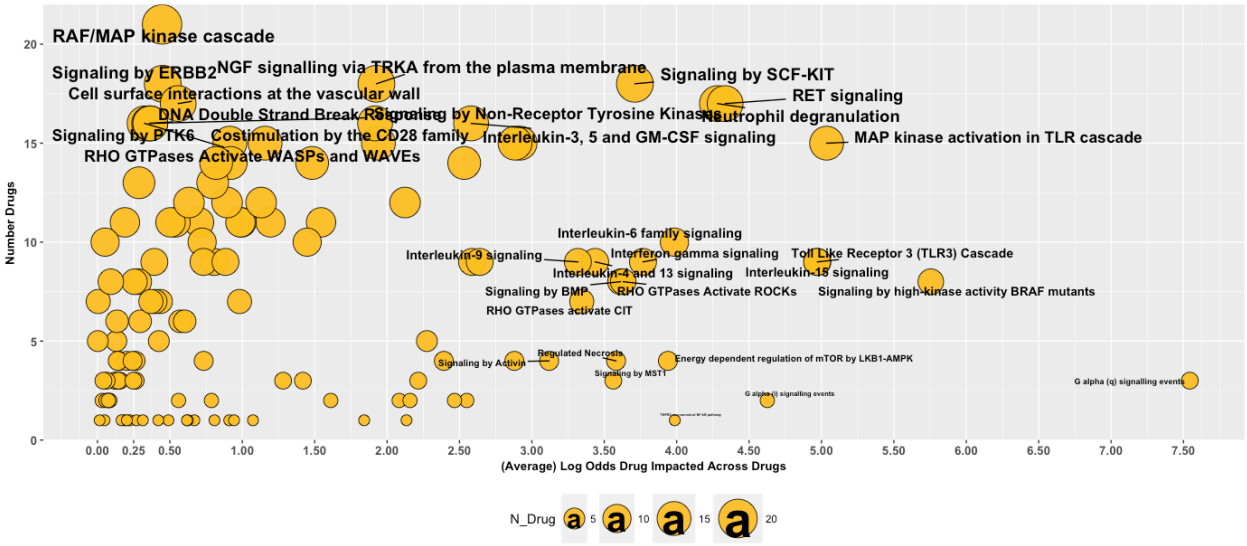


Figure 17. PGM Drug-Impacted Pathways for 32 BeatAML Panel Drugs Found in Cancer Targetome. Each point represents a drug-impacted pathway. Point size indicates the total number of drugs impacting the pathway.

Drug-Impacted Pathways

For each pathway, we can explore which drugs target pathway members and the resulting overall pathway impact. For example, in Figure 18 we show an overview of the pathway 'Interleukin-3,5 and GM-CSF signaling', which has a total of 16 drugs interacting with pathway members. For seven of the drugs, the overall average pathway log odds impact is around 1.0, while the rest of the drugs have a much higher average log odds impact, between 3.5 and 4.0.

To further explore individual drug-target relationships within this pathway, we can plot pathway impact showing the contribution of individual targets. In Figure 19, we show the same pathway, 'Interleukin-3,5 and GM-CSF signaling',

this time with all drug-targeted pathway members shown on the Y-axis. In this plot, drugs are shown across all targets in the pathway with which they interact. For instance, the drug ruxolitinib is shown in magenta and targets JAK3, FYN, HCK, and LYN, and has an overall average log odds impact of about 1.25. The drug idelalsib, shown in teal, targets PIK3CD, PIK3CB, and PIK3CA and has an overall average log odds impact of about 1.35, plotted in Figure 16. The drug dasatinib, shown in olive green, has a much higher impact on the pathway then the previous two drugs and targets a total of 8 targets in the pathway (TEC, JAK2, JAK3, YES1, SYK, FYN HCK, LYN).

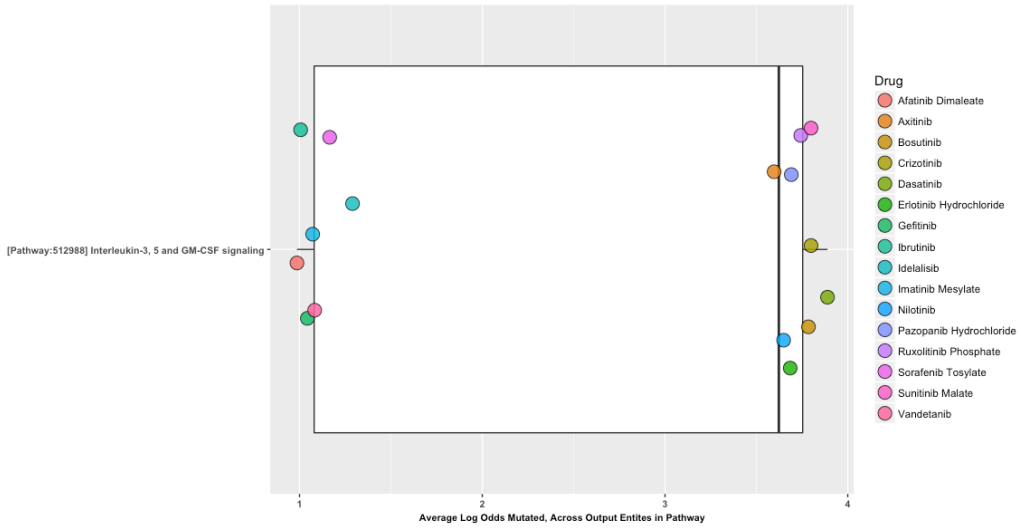


Figure 18. Overview for Drug-Impacted Pathway, 'Interleukin-3,5 and GM-CSF signaling'. Drugs interacting with the pathway are shown as single points, plotted according to the pathway impact score.

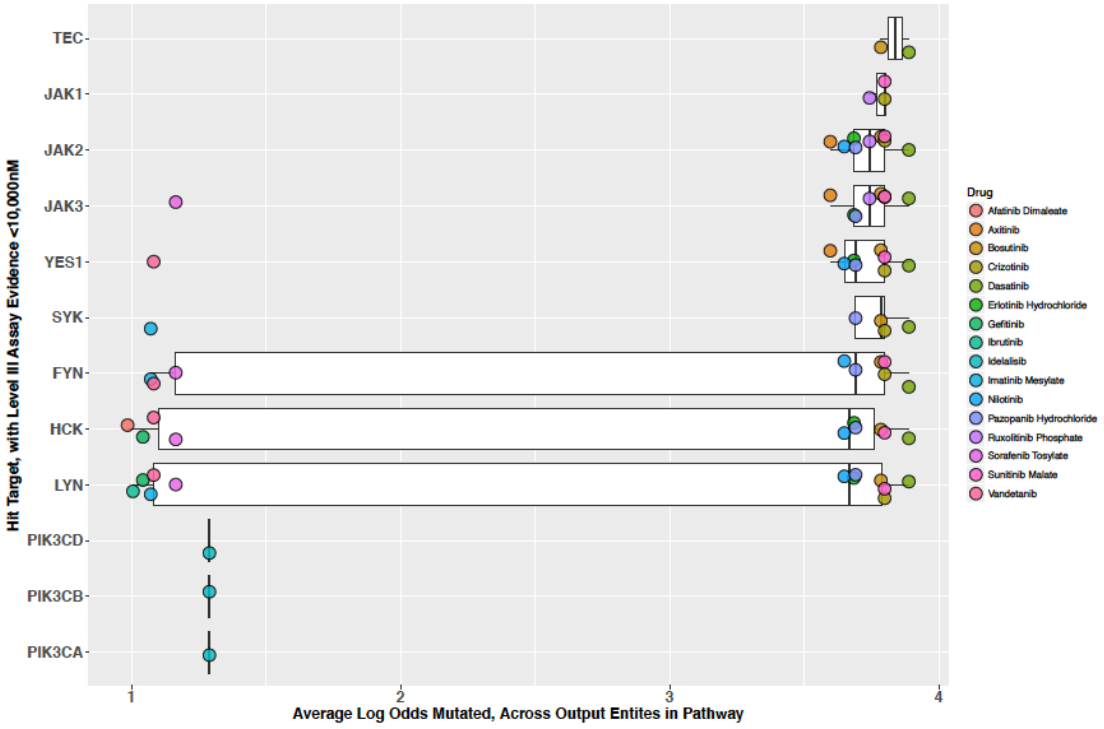


Figure 19. By-Target Plot for Drug-Impacted Pathway, 'Interleulin-3,5 and GM-CSF signaling'. For each drug, all targets genes are shown. Color indicates drug.

Uniting Mutation Pathway Impact and Drug Pathway Impact to Better Understand Patient Drug Response

Our work in parallel arms for modeling mutational impact and modeling drug impact is ultimately directed towards a future unified model of both mutation and drug impact together. In Figure 20, we again show the mutation impacted pathways of the AML de novo cohort, only now we also show the modeled drug impact for each pathway. This allows us to assess both the mutation and the drug impact for important pathways in the AML de novo cohort.

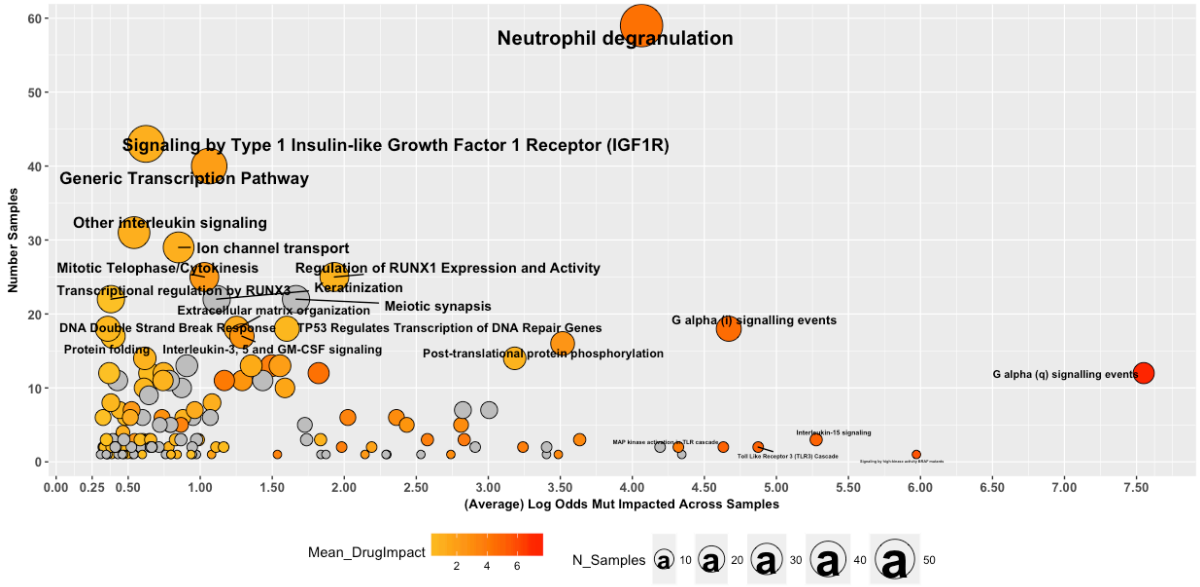


Figure 20. Uniting Mutation Pathway Impact and Drug Pathway Impact. Bubble plot shows mutationally impacted pathways for 187 AML de novo patient samples. Size indicates number of samples with pathway impacted. Color indicates the strength of drug impact for that pathway from the modeling.

Here we will highlight one example, for the pathway 'G alpha (q) signaling events'. This pathway is involved in GPCR signaling, which has important implications for cancer development, in particular with cancer stem cells [121]. A total of 12 AML de novo samples have mutations in this pathway. In Figure 21, we show the resulting pathway impact for each of these samples, all which have very high average log odds pathway impact. We note that all the samples have mutations in different genes in this pathway, again highlighting the strength of the pathway approach to aggregate samples with different gene mutations into potentially functionally related groups (as shown in Figures 7 and 8 in section 'Assessing the Relationship Between Pathway Mutational Status and Drug Response').

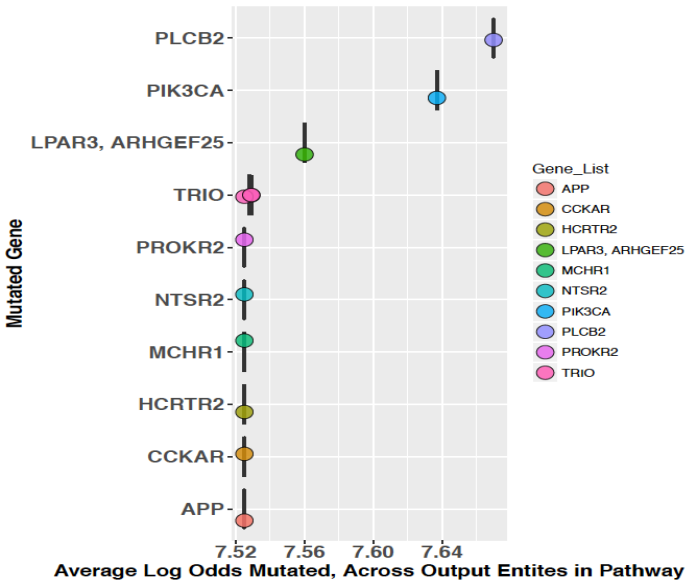


Figure 21. Pathway Mutation Impact for samples with mutations in 'G alpha (q) signaling'. Mutation pathway impact for 12 AML de novo samples with mutations in the pathway 'G alpha (q) signaling). Color indicates mutated gene.

In Figure 22, we show the targets of three drugs putatively interacting with the pathway 'G alpha (q) signaling': bortezomib, idelalisib, and sorafenib. We note that all three drugs have high inferred impact on the pathway according to the modeling.

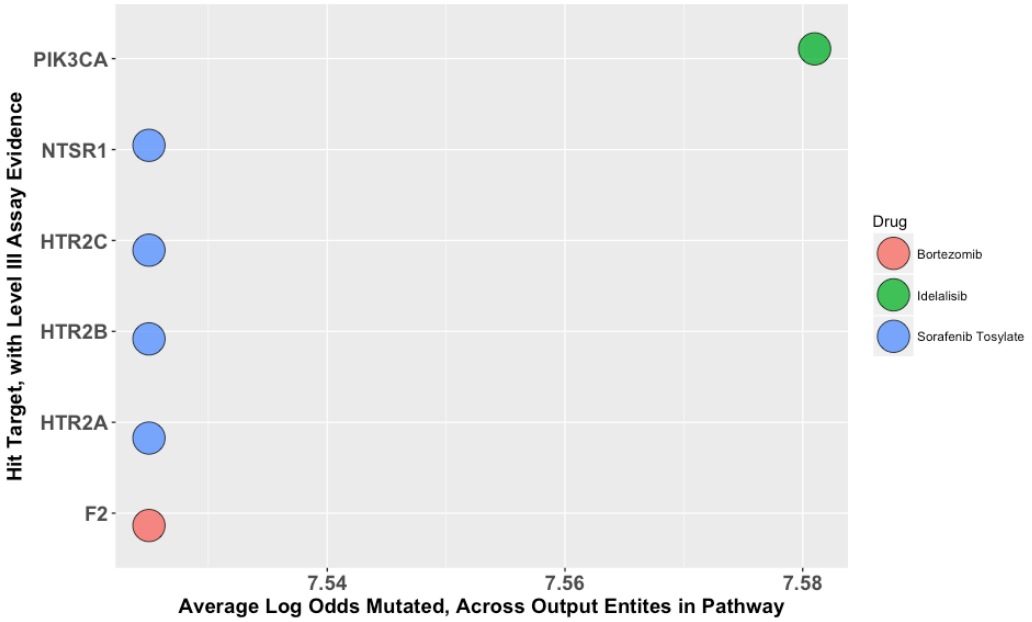


Figure 22. Bortezomib, Idelalisib, and Sorafenib Have High Impact on Pathway 'G alpha (q) signaling events'. Drug targets for each drug are designated by color.

To further explore the relationship between high mutation impact and high drug impact for this pathway, we examine the resulting drug screen data for samples with mutations in ‘G alpha (q) signaling’. In Figure 23, we show the AUC distributions for bortezomib, idelalisib, and sorafenib, with color indicating those samples harboring a mutation in the pathway ‘G alpha (q) signaling events’ and therefore found to have inferred pathway impact in the PGM mutation modeling. Three samples have a very low AUC for Sorafenib, three samples have a very low AUC for Idelalisib, and two samples have a fairly low AUC for Bortezomib. In Figure 24, we show the AUCs for these samples across all three drugs, bortezomib, idelalisib, and sorafenib.

Figure 23. A. Bortezomib AUC distribution for AML de novo samples (6 samples with mutations in 'G alpha (q) signaling events'). B. Idelalisib AUC distribution for AML de novo samples (5 samples with mutations in 'G alpha (q) signaling events'). C. Sorafenib AUC distribution for AML de novo samples (7 samples with mutations in 'G alpha (q) signaling events'). Color indicates samples with a mutation in the pathway 'G alpha (q) signaling' and have inferred pathway impact from PGM mutation impact modeling.

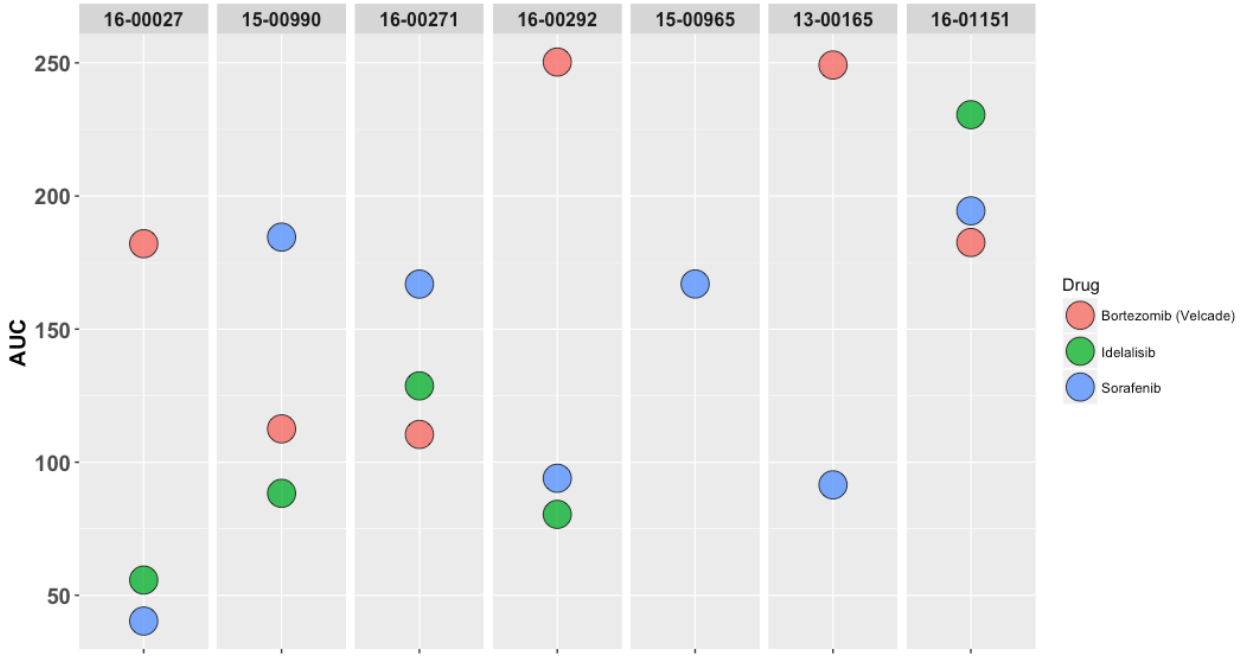


Figure 24. By sample plot of drug AUC for Bortezomib, Idelalisib, Sorafenib. Color indicates drug. All available drug screen data for these patients across these three drugs is shown in plot.

This assessment illustrates how we might begin to unite modeling results for both mutation and drug impacted pathways to better understand individual patient level drug response. Here, we looked at an example of pathway with both high mutation and high drug impact. This presents one possible way to prioritize impacted pathways of interest for the AML de novo cohort. However, additional work in this direction requires several considerations. Firstly, when evaluating patient samples and their drug response, we also need to consider whether well-known mutations (such as FLT3 or NPM1) associated with drug response are co-occurring for the patients. This is especially critical for mutations already known to be associated with patient drug response. This information is not currently shown explicitly in our figures as we are modeling each pathway and its mutated genes independently from other pathways. Secondly, for pathways of particular

interest, we need to more closely examine the pathway diagrams and take into consideration the location of mutated genes and directionality of graph connections. This phase of work would ideally be conducted more closely with experimental biologists to prioritize pathway of interest with strong biological connection to AML.

Assessment of Modeling Coverage with Respect to Significant Pathway and Drug Associations

Here we provide assessment of our impact modeling coverage with respect to both mutation and drug-impacted pathways. First, we examine the pathway and drug interactions found to be significant from the ANOVA analysis (previously described) in light of their inferred mutation pathway impact. These interactions represent relationships in which pathway dysregulation is found to be associated with either drug sensitivity or response. In Figure 25, we show these interactions plotted by their average pathway mutational impact. Overall, many of these associations actually have relatively low mutation pathway impact (under 1.0). We note that many of these associations involve the pathway 'Other Interleukin Signaling' which is home to the gene FLT3. We know FLT3 to be a critical gene mutation for AML, so an internal check for our modeling would be to expect that this gene mutation results in pathway impact. However, the pathway 'Other Interleukin Signaling' is currently poorly connected and thus even when FLT3 is strongly impacted, propagation is minimal to the rest of the pathway factor graph.

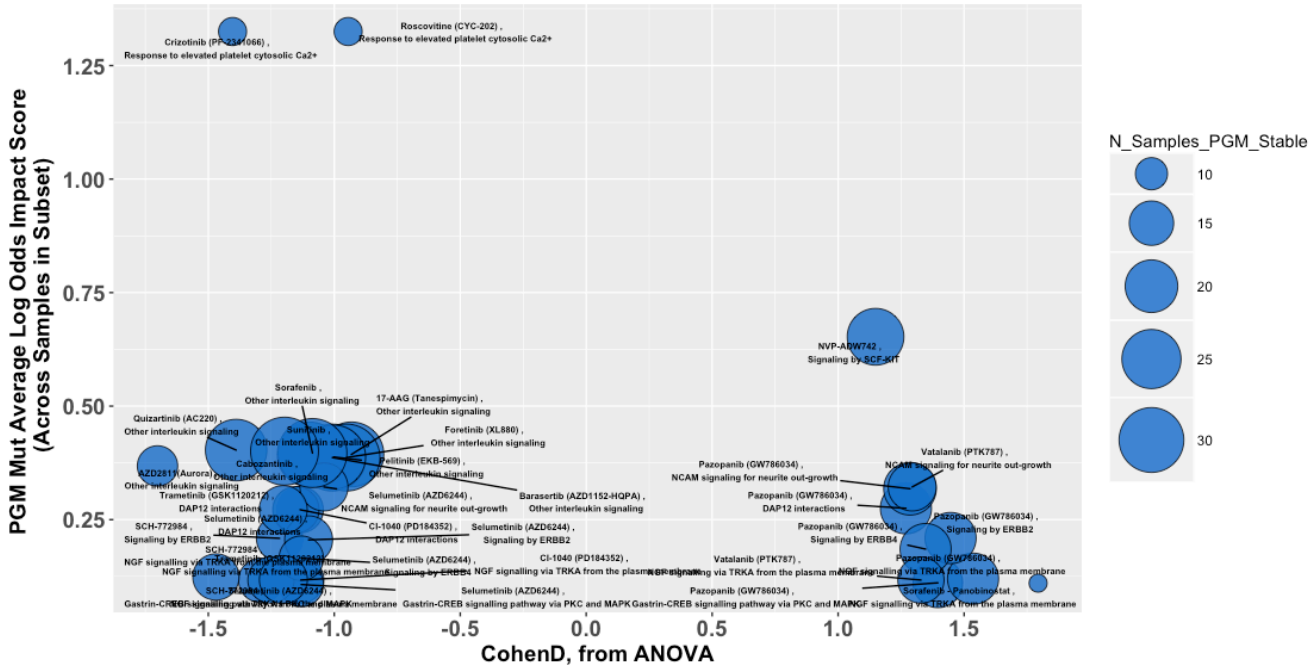


Figure 25. Significant Pathway and Drug Associations from ANOVA Analysis, by Pathway Mutational Impact. All plotted points represent pathway and drug interactions found to be significant in ANOVA described previously in section 'Assessing the Relationship Between Pathway Mutational Status and Drug Response' (FDR<0.05).

Next we consider both mutation impacted and drug impacted pathways. In Figure 26, we show all pathway and drug interactions tested in the ANOVA (described previously) where the drug is included in Cancer Targetome. Significant pathway and drug interactions are colored in blue, and pathways that are inferred to be pgm drug impacted are shown in red. Pathway-drug interactions found to be significant (from ANOVA) and also to be drug-impacted are shown in purple. From this perspective, we can see that several of our ANOVA hits involve pathways that are both PGM mutation-impacted and PGM drug impacted. A total of six ANOVA significant pathway and drug interactions that are also drug-impacted according to pathway modeling. In other words, these drug-pathway associations were found to be significantly associated in our

cohort of AML de novo patient samples and are also found to be impacted according to our pgm drug impacted pathway modeling. For a total of five ANOVA significant interactions, we did not find that they were drug-impacted according to pathway modeling. So, while these drug-pathway associations appear to be significantly associated in our cohort of samples, our modeling work does not currently indicate those pathways to be drug impacted.

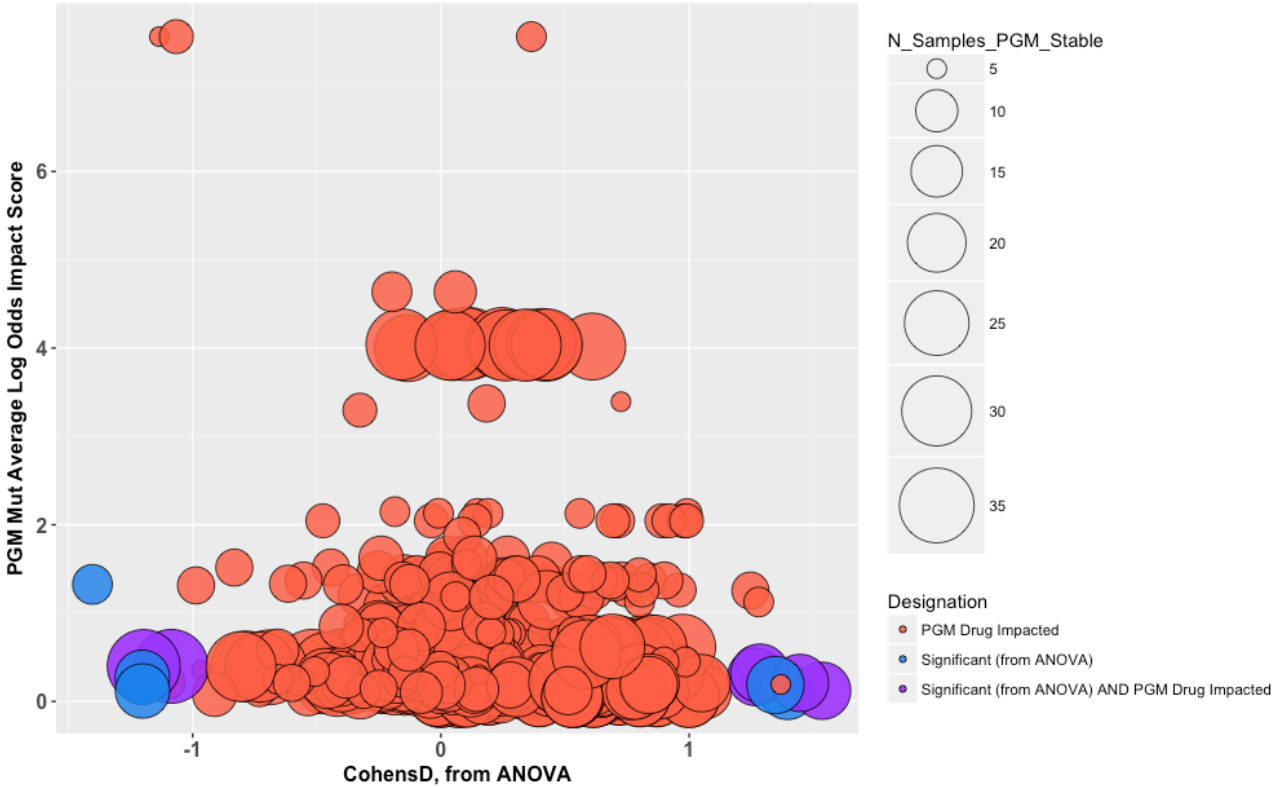


Figure 26. Pathway Impact Modeling Coverage with Respect to Mutational Impact, Drug Impact, and ANOVA Significance for Cancer Targetome Drugs Only. Blue indicates pathway and drug interaction was found to be significant (FDR<0.05). Red indicates pathway and drug interaction was found to be impacted from drug pathway modeling. Purple indicates pathway and drug interaction was both significant and found to be impacted in drug pathway impact modeling. Plot restricted to only those drug-pathway interactions where the drug is found in Cancer Targetome (drug-pathway interactions for 32 drugs in total).

Overall, this assessment provides areas for improvement for our modeling framework. For instance, we should expect to be able to model FLT3 mutations

as having a very high impact on their pathway, ‘Other Interleukin Signaling’, but as curated, this pathway is more of a collection rather than a well-connected set of signaling entities. When we layer on the drug-impacted pathways, we find that approximately half of the significant ANOVA relationships involve pathways that are drug-impacted according to the modeling framework. The drug-pathway interactions that were found to be significant from the ANOVA analysis for the cohort of AML samples but not found to be drug impacted from the modeling work present case examples to investigate further in an effort to understand the underlying mechanism. Additionally, we also note that there are many drug-pathway interactions that were found to have both a high pathway mutation impact and a high pathway drug impact, but did not come up as significant from the ANOVA analysis. Thus, even while mutation and drug impact modeling might indicate that these pathways are impacted, they are not necessarily associated with a meaningful different in drug AUC (in our cohort of AML samples) and thus did not pass significance in the ANOVA analysis. This work could be continued in modeling efforts that integrate both mutation and drug impact pathway modeling together.

Discussion

Targeted therapies for cancer precision treatment face immense challenges, such as drug resistance and decreasing efficacy, commonly resulting in patient cancer relapse [1]. Moving forward, we must integrate both individual patient mutational information and drug mechanistic information to better

understand response to targeted therapies and ultimately select the best treatment option for each patient. In particular, the field is moving to combination therapies as an effective strategy for combating resistance [1], but rationally designing combination therapies necessitates a much deeper mechanistic understanding of patient drug response than what we currently have. By taking a pathway perspective on mutational impact and the associated relationship with drug response, our work here significantly expands upon traditional 'pharmacogenomics' approaches that typically consider relationships between one biomarker gene or gene mutation and one drug.

Mapping mutations in AML de novo samples to pathways allows us to aggregate mutations across multiple genes into sets of functionally interacting entities. This allows us to assess patients from a view of shared dysregulated pathways that encompasses multiple genes, rather than one gene alone. In this way, we can group together more patients than what would have been possible using a single-gene as a biomarker. This approach yielded many significant associations between pathways and drug response. In particular, we recapitulate known associations such as between the FLT3 gene and FLT3 inhibitors (sorafenib, quizartinib, etc.), which serve as an indicator that pathway approaches preserve known signal between one gene and one drug relationships. When we examine these associations individually, we find that a spectrum of genes in the pathway often drives these relationships with drug response. For example, the association between the inhibitor NVP-ADW742 and the pathway 'Signaling by SCF-KIT' is driven by a number of genes, most of

which are present in a small number of samples when considered individually. Because of this small sample size (at the cohort level), an association between a single gene and an inhibitor may not be apparent, or pass multiple testing. Thus, the pathway aggregation approach allows us to group together samples at the pathway level and still capture meaningful information about associations with drug response.

However, with the simple mapping approach and ANOVA assessment, we cannot delve in deeper beyond a gene's status as mapped or not mapped to a particular pathway. In order to investigate the effect of mutations in particular genes on a pathway, we used probabilistic graphical modeling to propagate the effect of a gene mutation to further entities in the pathway. This approach allowed us to model the impact of mutated genes on a pathway, including when multiple genes within a pathway are mutated. Overall, our approach for modeling patient mutational impact on pathway takes into account both the functional impact of gene-level mutations and also the topology of the pathway. For several examples shown, mutations in the same genes result in differing levels of impact on pathways. In some cases, this is because the affected genes may occupy different locations in the pathways to which they belong. Additional future work to address the relationship between pathway location and inferred pathway impact will allow more rigorous assessment of this characteristic. Additionally, the functional impact (captured by variant effect prediction algorithms) dictates the strength of the impact at the mutated gene node in the pathway model. Higher functional impact results in more propagation to the pathway and larger overall

impact. This approach allows us to consider both the location of the mutated gene (within the pathway) and the estimated functional impact of the harbored. Both of these aspects allow us finer grained detail than a simple pathway mapping approach, as conducted earlier.

One of the challenges we faced in evaluating the modeling approach with respect to drug sensitivity was limited sample size. For a given pathway, we have a small number of samples with mutations mapped to that pathway, and an even smaller subset of samples for each gene mutated. This makes assessing (in a statistically rigorous way) difficult. We conducted a correlation analysis (correlation effect size only, no correlation significant test) between pathway mutational impact and drug AUC to investigate if there was a correlation between magnitude of pathway impact and drug AUC. While some associations were found to have a high correlation, the sample sizes were very small (less than 10 samples), and thus we did not proceed further, i.e. conducting correlation significance tests.

In parallel, we also apply our probabilistic graphical modeling framework to model drug impact on pathways, which provides deeper insight beyond a binary mapping of drug-targets to pathways. Mapping drug-targets to pathways in binary manner (mapped/not mapped) essentially treats all targets as equal. This doesn't allow us to distinguish between drugs that may very weakly interact with a pathway and drugs that may very strongly interact with a pathway. Our previous work developing the Cancer Targetome has demonstrated that drug-target interactions are supported by a wide range of strength of binding assay evidence,

and that some drug-target relationships are much stronger (lower nanomolar) than others. Thus it is critical that our framework captures this range of binding evidence in the way we model drug-target interactions. Our modeling framework very uniquely allows us to represent drug-target binding relationships in a quantitative manner (through encoded factors), where strength of a drug-target relationship reflects the strength of its supporting binding assay evidence. We note that the factor functions for our drug-impact modeling were naively set (hard-coded according to thresholds in binding assay values from Cancer Targetome). This area presents an opportunity for further work, i.e. sensitivity analysis of drug-target node factor functions and the resulting effect on pathway impact. Future work could include defining more appropriate functions to map between binding assay values and factor nodes. However, this work presents originality in drug-pathway modeling by representing drug-target relationships in a quantitative and evidence-based manner. In this way, we hope to better represent the nature of drug's pharmacological interaction, or targeting across multiple biological entities, as we know many drugs are highly promiscuous, i.e. kinase inhibitors.

Given that only 32 drugs were in both the BeatAML drug screen and the Cancer Targetome, we were limited in coverage for drugs tested for the AML patient samples. Future work with the Cancer Targetome to expand to investigational drugs will greatly increase coverage for modeling efforts and allow more investigation into the computational modeling efforts for drug pathway impact. Additionally, expansions of the Cancer Targetome to all drugs (i.e. FDA-

approved drugs from therapeutic domains outside cancer) would open opportunities for drug repurposing efforts, where drugs could be matched to patient samples based on pathway impact and repurposed to use as treatment for cancer.

From our assessment of modeling coverage, we can make recommendations for next steps for the probabilistic graphical modeling approach. Three key areas for future modeling efforts are:

1. **Curation and connectedness of pathways.** Not all pathways in Reactome are extensively curated or contain very well-connected biological entities. In some cases, this may represent the current state of our knowledge for particular pathways, while in other cases these pathways may be updated in upcoming releases. In particular, given the importance of FLT3 in AML, having pathways with FLT3 as a member that are extensively connected will greatly aid future modeling efforts. In our data, the pathway ‘Other Interleukin Signaling’ which contains FLT3, came up in a high number of drug-pathway associations, which is to be expected for drugs that are FLT3-inhibitors. However, our pathway-impact modeling for this pathway has a fairly low mutational impact (when compared to all mutation impacted pathways) due to the lack of propagation in the relatively unconnected pathway.
2. **Stability of pathways.** The loopy belief propagation algorithm is not guaranteed to converge, and our modeling efforts reflect that about a third of the Reactome pathways used did not converge to a stable

configuration. For these pathways, we were unable to pursue either mutation or drug-impact modeling. Our stability assessment found that many of these pathways contain a large number of loops. In the future, it may be possible to slightly modify these pathways without invalidating their biological behaviors so that they can converge as represented by factor graphs.

3. **Relationship with drug sensitivity.** Our modeling coverage assessment highlighted that approximately half of the significant drug-pathway relationships found through ANOVA analysis were both mutation pathway impacted and drug pathway impacted. However, overall we saw fairly low pathway mutation impact for these hits, when compared to the full range of mutational pathway impact. This could be addressed in future modeling efforts and we should address whether some pathways have an artificially inflated high mutation impact score due to their relative size.

Future iterations of the modeling framework will explore how we summarize pathway impact. Our approach here for summarizing pathway impact was to average the log odds impact across all entities in the pathway to get an average pathway score. This is a reasonable first pass-measure, but in using the average we suffer a loss of information. In the future, we can explore further ways of quantifying the impact across entities in a pathway. For instance, it may be important to be able to differentiate between when a small number of entities are highly impacted versus when large number of entities are only moderately impacted. Another option would be to break down very large pathways into sub-

pathways or neighborhoods for further analysis. With our current approach, our pathway impact score is reflective of the average impact across all pathway entities. Additional modifications could include adding weights for entities based on prior evidence or knowledge that they are impactful/have an effect on the pathway.

Currently, by creating a separate factor graph representation for each pathway, our modeling approach assumes that pathways are independent of one another. So a pathway that is strongly impacted does not propagate this pathway to another, even if they have shared members. Biologically, we know that this assumption is very weak as biological pathways are highly connected and often share membership of key entities, resulting in pathway crosstalk. A future direction for this work would be to unite pathways into a large model (a network of pathways) that allow pathway impact to actually be propagated to downstream or cross-talking pathways.

One of the most promising future extensions of this work is the application to both mutation pathway impact and drug impact simultaneously. An integrated modeling framework with both types of information (mutation and drug) would allow us to make very specific predictions about patients and drug response based on their mutational profiles. The vision with this type of modeling would be to be able to match patients with the most appropriate drug based on their mutational signatures across pathways. In the future, the modeling framework proposed here could be modified so that pathway factor graphs are loaded with both mutation and drug information, and the resulting pathway impact score

obtained would be for patient-specific drug response. Such extensions will very critically require that we have additional information about the interaction between drug-target binding and specific mutations, and could possibly benefit from formulation of the factor graphs with three states per node rather than two, allow us to capture both up and down-regulated states.

CHAPTER 5. DISCUSSION

Given the challenges facing precision cancer treatment today, such as limited therapy response, drug resistance and cancer relapse, there is a need to improve our mechanistic understanding of both how drugs work and how individual gene alterations affect this process. In this dissertation, we have made several key contributions to these aims. First, we have produced a framework for handling and prioritizing drug-target interaction data that allows for prioritization in an evidence-based manner. Second, we have developed a pathway modeling framework and applied it to understanding the pathway-level impact of somatic mutations and the pathway-level impact of drug application.

Our mechanistic understanding of cancer drugs is often very limited – and while it has become accepted that many drugs bind to multiple targets, characterization of these targets in a rigorous and evidence-backed manner has been lacking [5,6]. Our work in Aim 1 introducing the Cancer Targetome provides a framework for aggregating and evaluating drug-target interactions for cancer drugs [119]. Such a framework is critical for tasks leveraging drug-target interaction information, for as we showed, target designation is not a straightforward task. We demonstrated how to use this framework to assess the strength of evidence supporting drug-target interactions, and in particular looking at all targets supported for a particular drug. We further demonstrated the importance of considering multiple binding assay types (IC50, KD, Ki, EC50), as they often offer complementary information about targets. This work is impactful because it provides a way for researchers (both computational and experimental)

to explore drug-target interaction data in a way where supporting evidence is clear and transparent. In this way, the Cancer Targetome framework is working to elucidate the uncertainty that has up to this point been intertwined with drug-target interaction information.

Given that different binding assay types offer differing information about drug-target relationships, the question arises about how to best prioritize supporting evidence in application. Further work is needed to develop methods of weighting different assay types and evidence backed by multiple sources and to apply these methods computationally in applications such as drug panel development, drug combination design, and in combating drug resistance. Additionally, while the framework developed here allows for evidence-based prioritization of public domain drug-target interaction data, there are very clear biases in the collected data with regards to drugs and targets. For instance, there is a lot of data on kinase inhibitors due to extensive kinase screens, while other drugs have information for only a small number of targets. Applications of this data and framework much therefore keep in mind the information that we do not currently have, for instance – the drug-target interactions not yet tested or available in public resources.

Future expansions for the Cancer Targetome include expansion to all FDA-approved drugs and investigational stage drugs. Drug interactions spanning many therapeutic domains will greatly improve the applicability and usefulness of the Targetome framework for researchers both in the cancer domain and in other therapeutic areas. With regards to data collection, we note that of the resources

used for the Targetome in Aim 1, DrugBank, IUPHAR/BPS, and BindingDB each contributed unique content (drug or target coverage) not contained in the other resources. Therapeutic Targets Database (TTD), however, provided minimal unique content and ultimately only contributed Level I evidence-supported drug-target interactions due to the lack of linkage between interactions and supporting reference. Thus future work will likely not continue to mine TTD for interactions. Expansions to additional resources will likely improve drug coverage, as many of these resources mine information from different sets of journals. In particular resources such as ChEMBL, which houses curated bioactivity data for a wide range of compounds, would likely prove beneficial for increasing Level III-supported interaction coverage. We also note the considerable effort undertaken by the Illuminating the Druggable Genome (IDG) Consortium, an NIH-led effort to characterize understudied yet potentially therapeutically relevant entities in the genome [62,124,125]. Future work for Cancer Targetome could also leverage the extensive work by IDG, in particular their use of target levels, which indicate the level of study associated with a particular target (clinical level, research level, or chemical tool level).

Mapping drug-target interactions and their evidence levels to Reactome pathways allows us to put drug-target relationships into greater biological context. This perspective is critical in our efforts to move toward combination therapies, which need to be strategized at the pathway-level to overcome cancer resistance mechanisms. Light pathways, or potentially druggable pathways, can be leveraged in drug panel development and drug combination design, while

dark pathways pose as areas for future research in developing new compounds. Our work in Aim 1 focused solely on FDA-approved cancer drugs, but future work to bring in all FDA-approved cancer drugs and investigational stage drugs would provide a more accurate perspective of light and dark pathways. In particular, this would open up avenues for repurposing non-oncology drugs into the cancer domain for use.

We build upon the pathway perspective in Aim 2, where we pursue a pathway modeling framework for both patient somatic mutation impact and drug impact. In this aim, we focus on application to the disease acute myeloid leukemia, which is in desperate need of improved therapies and is characterized by extreme heterogeneity of patient mutational profiles. To demonstrate the applicability of pathway approaches, we first show that aggregating patient mutations to pathways captures association between some pathways and drug response. Importantly, these associations are driven by many mutated genes within the pathway. This means that analysis considering only one gene and one drug at a time is likely missing these relationships. This analysis at the pathway level captures important information about mutation and drug response interactions and furthermore offers additional information than what is available from analysis at the gene-level.

We pursue further mechanistic understanding with our probabilistic graphical modeling framework. In parallel, we model both pathway impact of somatic mutations and the pathway impact of drug-target binding. For somatic mutation information, we demonstrated that this framework captures both the

functional impact of a mutation and also the positional impact of the gene based on pathway topology. For drug-target impacted pathways, we also show that this framework captures the strength of the drug-target interaction and the positional impact of the target gene based on pathway topology. The ability to encode drug-target interactions into this modeling framework is powered by the Cancer Targetome, as Level III-supported drug-target interactions are accompanied by quantitative assay binding values. For both the functional mutation impact and drug-target interaction binding values, we are able to incorporate these into the probabilistic graphical modeling framework through the use of factor functions, which describe the impact due to mutation or drug-binding at a particular gene node.

Modeling coverage assessment allows us to make recommendations for the next iteration of modeling. In particular, we assess pathway modeling for both mutation and drug impact with respect to significant pathway and drug associations from the earlier ANOVA analysis. We found that overall, mutational pathway impact is fairly low, which may point to the need for further refinement of pathway modeling. In particular, several of the pathway-drug associations that were found to be significant from ANOVA analysis involve pathways that are fairly unconnected (like ‘Other Interleukin Signaling’) and therefore propagation is limited, resulting in a lower overall pathway impact. Secondly, coverage of drugs in the BeatAML screening panel and the Cancer Targetome was limited to 32 drugs in total. This limits the amount of investigation for drug-impacted pathways. However, when we restricted our assessment to these drugs only, we found that

6 out of 11 significant drug-pathway associations are for pathways that are predicted to be drug impacted by the modeling approach. Thus the drug modeling framework also appears to be promising with regards to coverage of impacted pathways in a manner that is meaningful to associations found in the acute myeloid leukemia cohort.

Future work with the modeling framework and acute myeloid leukemia cohort will expand analysis of AML de novo samples to include both paired samples and tumor only sample (no paired normal sample). This will allow us to investigate whether pathway-drug associations found in this analysis hold in the larger cohort. More specific analyses could include subgrouping based on canonical AML mutation types (for instance, FLT3_ITD, NPM1, DNTM3A) and investigation of pathway impact by these subtypes.

Overall, our work sets the stage for more rigorous and quantitative modeling work for understanding drug sensitivity and resistance in patients based on their unique mutational profiles. Our work with the Cancer Targetome in Aim 1 provides a framework for drug-target interaction that is evidence-based and easily harnessed in modeling efforts, as shown in Aim 2, where is it is used in parallel with models for patient mutational impact. This framework captures both strength of impact (for mutation and targets) as well as pathway topology and captures meaningful drug-pathway associations found through other analyses. Future work will focus on the application of this framework to efforts to improve understanding of patient subgroup drug response and be used for treatment development. In particular, future efforts will focus on uniting the

patient mutation impact and drug impact modeling frameworks to further understand how individual patient's mutational profiles affect their response to targeted therapies. In the long term, we aim to be able to predict sensitivity and resistance for patients based on sample mutational information and use this to guide the best course of treatment.

References

1 Al-Lazikani, B. *et al.* (2012) Combinatorial drug therapy for cancer in the post-genomic era. *Nat. Biotechnol.* 30, 679–692

2 Wu, P. *et al.* (2015) FDA-approved small-molecule kinase inhibitors. *Trends Pharmacol. Sci.* 36, 422–439

3 Druker, B.J. *et al.* (2001) Efficacy and safety of a specific inhibitor of the BCR-ABL tyrosine kinase in chronic myeloid leukemia. *N. Engl. J. Med.* 344, 1031–1037

4 Greaves, M. and Maley, C.C. (2012) Clonal evolution in cancer. *Nature* 481, 306–313

5 Jalencas, X. and Mestres, J. (2013) On the origins of drug polypharmacology. *Med Chem Commun* 4, 80–87

6 Mestres, J. *et al.* (2008) Data completeness—the Achilles heel of drug-target networks. *Nat. Biotechnol.* 26, 983–984

7 Santos, R. *et al.* (2016) A comprehensive map of molecular drug targets. *Nat. Rev. Drug Discov.* DOI: 10.1038/nrd.2016.230

8 Hopkins, A.L. (2008) Network pharmacology: the next paradigm in drug discovery. *Nat. Chem. Biol.* 4, 682–690

9 Haupt, V.J. *et al.* (2013) Drug promiscuity in PDB: protein binding site similarity is key. *PloS One* 8, e65894

10 Lavecchia, A. and Cerchia, C. (2015) In silico methods to address polypharmacology: current status, applications and future perspectives. *Drug Discov. Today* DOI: 10.1016/j.drudis.2015.12.007

11 Berger, S.I. and Iyengar, R. (2011) Role of systems pharmacology in understanding drug adverse events. *Wiley Interdiscip. Rev. Syst. Biol. Med.* 3, 129–135

12 Peters, J.-U. (2013) Polypharmacology – Foe or Friend? *J. Med. Chem.* 56, 8955–8971

13 Chi, K.R. (2013) Revolution dawning in cardiotoxicity testing. *Nat. Rev. Drug Discov.* 12, 565–567

14 Xie, L. *et al.* (2009) Drug Discovery Using Chemical Systems Biology: Identification of the Protein-Ligand Binding Network To Explain the Side Effects of CETP Inhibitors. *PLoS Comput. Biol.* 5, e1000387

15 Paolini, G.V. *et al.* (2006) Global mapping of pharmacological space. *Nat. Biotechnol.* 24, 805–815

16 Hu, Y. and Bajorath, J. (2013) Compound promiscuity: what can we learn from current data? *Drug Discov. Today* 18, 644–650

17 Guney, E. *et al.* (2016) Network-based in silico drug efficacy screening. *Nat. Commun.* 7, 10331

18 Lounkine, E. *et al.* (2012) Large-scale prediction and testing of drug activity on side-effect targets. *Nature* DOI: 10.1038/nature11159

19 Li, Q. *et al.* (2010) PubChem as a public resource for drug discovery. *Drug Discov. Today* 15, 1052–1057

20 Wang, Y. *et al.* (2009) An overview of the PubChem BioAssay resource. *Nucleic Acids Res.* 38, D255–D266

21 Seiler, K.P. *et al.* (2007) ChemBank: a small-molecule screening and cheminformatics resource database. *Nucleic Acids Res.* 36, D351–D359
22 Swamidass, S.J. (2011) Mining small-molecule screens to repurpose drugs. *Brief. Bioinform.* 12, 327–335
23 Shun, T.Y. *et al.* (2010) Identifying Actives from HTS Data Sets: Practical Approaches for the Selection of an Appropriate HTS Data-Processing Method and Quality Control Review. *J. Biomol. Screen.* 16, 1–14
24 Fabregat, A. *et al.* (2016) The Reactome pathway Knowledgebase. *Nucleic Acids Res.* 44, D481–D487
25 Fabregat, A. *et al.* (2018) The Reactome Pathway Knowledgebase. *Nucleic Acids Res.* 46, D649–D655
26 De Kouchkovsky, I. and Abdul-Hay, M. (2016) “Acute myeloid leukemia: a comprehensive review and 2016 update.” *Blood Cancer J.* 6, e441–e441
27 Kadia, T.M. *et al.* (2015) Toward Individualized Therapy in Acute Myeloid Leukemia: A Contemporary Review. *JAMA Oncol.* 1, 820–828
28 The Cancer Genome Atlas Research Network (2013) Genomic and Epigenomic Landscapes of Adult De Novo Acute Myeloid Leukemia. *N. Engl. J. Med.* 368, 2059–2074
29 Grunwald, M.R. and Levis, M.J. (2013) FLT3 inhibitors for acute myeloid leukemia: a review of their efficacy and mechanisms of resistance. *Int. J. Hematol.* 97, 683–694
30 OHSU BeatAML Consortium *et al.* (2018) Functional Genomic Landscape of Acute Myeloid Leukemia. *Nature, Under Review*
31 Tyner, J.W. *et al.* (2013) Kinase Pathway Dependence in Primary Human Leukemias Determined by Rapid Inhibitor Screening. *Cancer Res.* 73, 285–296
32 Tyner, J.W. (2014) Functional Genomics for Personalized Cancer Therapy. *Sci. Transl. Med.* 6, 243fs26-243fs26
33 Rotow, J. and Bivona, T.G. (2017) Understanding and targeting resistance mechanisms in NSCLC. *Nat. Rev. Cancer* 17, 637–658
34 Iorio, F. *et al.* (2016) A Landscape of Pharmacogenomic Interactions in Cancer. *Cell* DOI: 10.1016/j.cell.2016.06.017
35 Adzhubei, I. *et al.* (2013) Predicting Functional Effect of Human Missense Mutations Using PolyPhen-2. In *Current Protocols in Human Genetics* (Haines, J. L. et al., eds), pp. 7.20.1-7.20.41, John Wiley & Sons, Inc.
36 Vaser, R. *et al.* (2015) SIFT missense predictions for genomes. *Nat. Protoc.* 11, 1–9
37 Vaske, C.J. *et al.* (2010) Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics* 26, i237–i245
38 Koller, D. and Friedman, N. (2012) *Probabilistic Graphical Models*, MIT Press.
39 Kschischang, F.R. *et al.* (2001) Factor graphs and the sum-product algorithm. *Inf. Theory IEEE Trans. On* 47, 498–519
40 Koller, D. (2009) *Probabilistic graphical models: principles and techniques*, MIT Press.

41 Gat-Viks, I. *et al.* (2005) , The factor graph network model for biological systems. , in *Research in Computational Molecular Biology*, pp. 31–47
42 Dudley, J.T. *et al.* (2011) Exploiting drug-disease relationships for computational drug repositioning. *Brief. Bioinform.* 12, 303–311
43 Ashburn, T.T. and Thor, K.B. (2004) Drug repositioning: identifying and developing new uses for existing drugs. *Nat. Rev. Drug Discov.* 3, 673–683
44 Pijl, H. *et al.* (2000) Bromocriptine: a novel approach to the treatment of type 2 diabetes. *Diabetes Care* 23, 1154–1161
45 Mayr, L.M. and Bojanic, D. (2009) Novel trends in high-throughput screening. *Curr. Opin. Pharmacol.* 9, 580–588
46 Xie, X.-Q.S. (2010) Exploiting PubChem for virtual screening. *Expert Opin. Drug Discov.* 5, 1205–1220
47 Macarron, R. *et al.* (2011) Impact of high-throughput screening in biomedical research. *Nat. Rev. Drug Discov.* 10, 188–195
48 Brideau, C. (2003) Improved Statistical Methods for Hit Selection in High-Throughput Screening. *J. Biomol. Screen.* 8, 634–647
49 Gribbon, P. (2005) Evaluating Real-Life High-Throughput Screening Data. *J. Biomol. Screen.* 10, 99–107
50 Li, Q. *et al.* (2010) PubChem as a public resource for drug discovery. *Drug Discov. Today* 15, 1052–1057
51 Johnston, P.A. *et al.* (2009) Cdc25B Dual-Specificity Phosphatase Inhibitors Identified in a High-Throughput Screen of the NIH Compound Library. *ASSAY Drug Dev. Technol.* 7, 250–265
52 Coffee, E.M. *et al.* (2013) Concomitant BRAF and PI3K/mTOR Blockade Is Required for Effective Treatment of BRAFV600E Colorectal Cancer. *Clin. Cancer Res.* 19, 2688–2698
53 Gaulton, A. *et al.* (2011) ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* 40, D1100–D1107
54 Calhoun, B.T. *et al.* (2012) Automatically Detecting Workflows in PubChem. *J. Biomol. Screen.* 17, 1071–1079
55 Prasad, V. *et al.* (2016) Precision oncology: origins, optimism, and potential. *Lancet Oncol.* 17, e81–e86
56 What Precisely Is Precision Oncology—and Will It Work? - The ASCO Post. . [Online]. Available: http://www.ascopost.com/issues/january-25-2017/what-precisely-is-precision-oncology-and-will-it-work/. [Accessed: 26-May-2017]
57 Saad, E.D. *et al.* (2017) Precision medicine needs randomized clinical trials. *Nat. Rev. Clin. Oncol.* 14, 317–323
58 Andre, F. *et al.* (2014) Prioritizing targets for precision cancer medicine. *Ann. Oncol.* 25, 2295–2303
59 Watson, I.R. *et al.* (2013) Emerging patterns of somatic mutations in cancer. *Nat. Rev. Genet.* 14, 703–718
60 Hudson (Chairperson), T.J. *et al.* (2010) International network of cancer genome projects. *Nature* 464, 993–998
61 McLendon, R. *et al.* (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* 455, 1061–1068

62 Nguyen, D.-T. *et al.* (2017) Pharos: Collating protein information to shed light on the druggable genome. *Nucleic Acids Res.* 45, D995–D1002
63 Biankin, A.V. *et al.* (2015) Patient-centric trials for therapeutic development in precision oncology. *Nature* 526, 361–370
64 Kim, E.S. *et al.* (2011) The BATTLE Trial: Personalizing Therapy for Lung Cancer. *Cancer Discov.* 1, 44–53
65 The I-SPY 1 TRIAL Investigators *et al.* (2012) Chemotherapy response and recurrence-free survival in neoadjuvant breast cancer depends on biomarker profiles: results from the I-SPY 1 TRIAL (CALGB 150007/150012; ACRIN 6657). *Breast Cancer Res. Treat.* 132, 1049–1062
66 Redig, A.J. and Jänne, P.A. (2015) Basket trials and the evolution of clinical trial design in an era of genomic medicine. *J. Clin. Oncol.* 33, 975–977
67 Lopez-Chavez, A. *et al.* (2015) Molecular Profiling and Targeted Therapy for Advanced Thoracic Malignancies: A Biomarker-Derived, Multiarm, Multihistology Phase II Basket Trial. *J. Clin. Oncol.* 33, 1000–1007
68 Pauwels, E. *et al.* (2011) Predicting drug side-effect profiles: a chemical fragment-based approach. *BMC Bioinformatics* 12, 169
69 Johns, D.G. *et al.* (2012) On-and Off-Target Pharmacology of Torcetrapib. *Drugs* 72, 491–507
70 Zhao, S. *et al.* (2013) Systems Pharmacology of Adverse Event Mitigation by Drug Combinations. *Sci. Transl. Med.* 5, 206ra140-206ra140
71 Corsello, S.M. *et al.* (2017) The Drug Repurposing Hub: a next-generation drug library and information resource. *Nat. Med.* 23, 405–408
72 Drews, J. (1996) Genomic sciences and the medicine of tomorrow. *Nat. Biotechnol.* 14, 1516–1518
73 Drews, J. and Ryser, S. (1997) Classic drug targets. *Nat. Biotechnol.* 15, 1350–1350
74 Hopkins, A.L. and Groom, C.R. (2002) The druggable genome. *Nat. Rev. Drug Discov.* 1, 727–730
75 Imming, P. *et al.* (2006) Drugs, their targets and the nature and number of drug targets. *Nat. Rev. Drug Discov.* 5, 821–834
76 Overington, J.P. *et al.* (2006) How many drug targets are there? *Nat. Rev. Drug Discov.* 5, 993–996
77 Rask-Andersen, M. *et al.* (2011) Trends in the exploitation of novel drug targets. *Nat. Rev. Drug Discov.* 10, 579–590
78 Rask-Andersen, M. *et al.* (2014) The Druggable Genome: Evaluation of Drug Targets in Clinical Trials Suggests Major Shifts in Molecular Class and Indication. *Annu. Rev. Pharmacol. Toxicol.* 54, 9–26
79 Wishart, D.S. (2006) DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.* 34, D668–D672
80 Law, V. *et al.* (2014) DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res.* 42, D1091–D1097
81 Chen, X. *et al.* (2002) TTD: Therapeutic Target Database. *Nucleic Acids Res.* 30, 412–415
82 Kanehisa, M. *et al.* (2010) KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res.* 38, D355–D360

83 Kanehisa, M. *et al.* (2017) KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* 45, D353–D361
84 Pawson, A.J. *et al.* (2014) The IUPHAR/BPS Guide to PHARMACOLOGY: an expert-driven knowledgebase of drug targets and their ligands. *Nucleic Acids Res.* 42, D1098–D1106
85 Griffith, M. *et al.* (2013) DGIdb: mining the druggable genome. *Nat. Methods* 10, 1209–1210
86 Ursu, O. *et al.* (2017) DrugCentral: online drug compendium. *Nucleic Acids Res.* 45, D932–D939
87 Hecker, N. *et al.* (2012) SuperTarget goes quantitative: update on drug-target interactions. *Nucleic Acids Res.* 40, D1113–D1117
88 Kuhn, M. *et al.* (2014) STITCH 4: integration of protein-chemical interactions with user data. *Nucleic Acids Res.* 42, D401–D407
89 Koscielny, G. *et al.* (2017) Open Targets: a platform for therapeutic target identification and validation. *Nucleic Acids Res.* 45, D985–D994
90 Bento, A.P. *et al.* (2014) The ChEMBL bioactivity database: an update. *Nucleic Acids Res.* 42, D1083–D1090
91 Gilson, M.K. *et al.* (2016) BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Res.* 44, D1045–D1053
92 Wang, Y. *et al.* (2011) PubChem's BioAssay Database. *Nucleic Acids Res.* 40, D400–D412
93 Wang, Y. *et al.* (2014) PubChem BioAssay: 2014 update. *Nucleic Acids Res.* 42, D1075–D1082
94 Bellis, L.J. *et al.* (2011) Collation and data-mining of literature bioactivity data for drug discovery. *Biochem. Soc. Trans.* 39, 1365–1370
95 Southan, C. *et al.* (2009) Quantitative assessment of the expanding complementarity between public and commercial databases of bioactive compounds. *J. Cheminformatics* 1, 10
96 Southan, C. *et al.* (2013) Comparing the Chemical Structure and Protein Content of ChEMBL, DrugBank, Human Metabolome Database and the Therapeutic Target Database. *Mol. Inform.* 32, 881–897
97 Tiikkainen, P. and Franke, L. (2012) Analysis of Commercial and Public Bioactivity Databases. *J. Chem. Inf. Model.* 52, 319–326
98 Wang, Y. *et al.* (2016) Evidence-Based and Quantitative Prioritization of Tool Compounds in Phenotypic Drug Discovery. *Cell Chem. Biol.* 23, 862–874
99 Koutsoukas, A. *et al.* (2013) In Silico Target Predictions: Defining a Benchmarking Data Set and Comparison of Performance of the Multiclass Naïve Bayes and Parzen-Rosenblatt Window. *J. Chem. Inf. Model.* 53, 1957–1966
100 Mervin, L.H. *et al.* (2015) Target prediction utilising negative bioactivity data covering large chemical space. *J. Cheminformatics* 7,
101 Lim, H. *et al.* (2016) Large-Scale Off-Target Identification Using Fast and Accurate Dual Regularized One-Class Collaborative Filtering and Its Application to Drug Repurposing. *PLoS Comput. Biol.* 12, e1005135

102 Hu, Y. and Bajorath, J. (2012) Many structurally related drugs bind different targets whereas distinct drugs display significant target overlap. *RSC Adv.* 2, 3481
103 Afzal, A.M. *et al.* (2015) A multi-label approach to target prediction taking ligand promiscuity into account. *J. Cheminformatics* 7,
104 Davis, M.I. *et al.* (2011) Comprehensive analysis of kinase inhibitor selectivity. *Nat. Biotechnol.* 29, 1046–1051
105 Finan, C. *et al.* (2017) The druggable genome and support for target identification and validation in drug development. *Sci. Transl. Med.* 9,
106 Qin, C. *et al.* (2014) Therapeutic target database update 2014: a resource for targeted therapeutics. *Nucleic Acids Res.* 42, D1118–D1123
107 Sioutos, N. *et al.* (2007) NCI Thesaurus: A semantic model integrating cancer-related clinical and molecular information. *J. Biomed. Inform.* 40, 30–43
108 Karaman, M.W. *et al.* (2008) A quantitative analysis of kinase inhibitor selectivity. *Nat. Biotechnol.* 26, 127–132
109 Druker, B.J. *et al.* (1996) Effects of a selective inhibitor of the Abl tyrosine kinase on the growth of Bcr–Abl positive cells. *Nat. Med.* 2, 561–566
110 Knowles, P.P. *et al.* (2006) Structure and Chemical Inhibition of the RET Tyrosine Kinase Domain. *J. Biol. Chem.* 281, 33577–33587
111 Wedge, S.R. *et al.* (2002) ZD6474 inhibits vascular endothelial growth factor signaling, angiogenesis, and tumor growth following oral administration. *Cancer Res.* 62, 4645–4655
112 Ciardiello, F. *et al.* (2003) Antitumor effects of ZD6474, a small molecule vascular endothelial growth factor receptor tyrosine kinase inhibitor, with additional activity against epidermal growth factor receptor tyrosine kinase. *Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res.* 9, 1546–1556
113 Jimeno, A. and Hidalgo, M. (2006) Multitargeted therapy: Can promiscuity be praised in an era of political correctness? *Crit. Rev. Oncol. Hematol.* 59, 150–158
114 Espinoza, I. and Miele, L. (2013) Notch inhibitors for cancer treatment. *Pharmacol. Ther.* 139, 95–110
115 Yuan, X. *et al.* (2015) Notch signaling: An emerging therapeutic target for cancer treatment. *Cancer Lett.* 369, 20–27
116 Guerrero-Preston, R. *et al.* (2014) Key tumor suppressor genes inactivated by “greater promoter” methylation and somatic mutations in head and neck cancer. *Epigenetics* 9, 1031–1046
117 Stein, E.M. (2015) Molecularly targeted therapies for acute myeloid leukemia. *Hematology* 2015, 579–583
118 Lai, C. *et al.* (2016) Precision medicine for acute myeloid leukemia. *Expert Rev. Hematol.* 9, 1–3
119 Blucher, A.S. *et al.* (2017) Evidence-Based Precision Oncology with the Cancer Targetome. *Trends Pharmacol. Sci.* DOI: 10.1016/j.tips.2017.08.006
120 Wu, G. *et al.* (2014) ReactomeFIViz: a Cytoscape app for pathway and network-based data analysis. *F1000Research* DOI: 10.12688/f1000research.4431.2

121 Lynch, J.R. and Wang, J.Y. (2016) G Protein-Coupled Receptor Signaling in Stem Cells and Cancer. *Int. J. Mol. Sci.* 17,
122 Benjamini, Y. and Hochberg, Y. (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journral of the Royal Statistical Society, Series B (Methodological)* 57, 289–300
123 Gat-Viks, I. *et al.* (2006) A probabilistic methodology for integrating knowledge and experiments on biological networks. *J. Comput. Biol. J. Comput. Mol. Cell Biol.* 13, 165–181
124 Oprea, T.I. *et al.* (2018) Unexplored therapeutic opportunities in the human genome. *Nat. Rev. Drug Discov.* 17, 317–332
125 Rodgers, G. *et al.* 19-Jan-(2018) , Glimmers in illuminating the druggable genome. , *Nature Reviews Drug Discovery*. [Online]. Available: https://www.nature.com/articles/nrd.2017.252. [Accessed: 08-May-2018]