THE UNIVERSITY OF CHICAGO

COMPUTATIONAL DRUG REPOSITIONING FOR TRIPLE-NEGATIVE BREAST CANCERS

A DISSERTATION SUBMITTED TO

THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES

AND THE PRITZKER SCHOOL OF MEDICINE

IN CADIDACY FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

COMMITTEE ON CANCER BIOLOGY

BY

ROBERT F. GRUENER

CHICAGO, ILLINOIS

JUNE 2021

This work is dedicated to my family.

For the care and attention of my parents and the support of many who helped raise me. For my grandparents, who worked hard to give opportunities to their children who in turn ensured the same for me. Beyond their efforts that made this possible, the love and encouragement I receive from my parents, grandparents, and sisters have made this journey more accessible, more enjoyable, and more gratifying.

I love you all.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

I must first sincerely thank my advisors Stephanie Huang and Geoffrey Greene. They took in a student who didn't know how to culture cell lines or do any coding and gave me space to explore and grow into a computational biologist (who could help with mouse experiments in a pinch). They have been supportive and allowed me to find my own path that aligns most with my interests. Their feedback and support have greatly shaped this work and the scientist I am today. Same goes to my committee: Haky Im, Suzanne Conzen, and Marsha Rosner. Thank you.

I started my graduate career in Molecular Pathogenesis and Molecular Medicine, and I would like to acknowledge the former chair Steve Meredith and my MPMM colleagues, particularly Erin McAuley. I must also thank Kay Macleod who took me into the Cancer Biology program after MPMM was dissolved (even though she didn't have to). Thanks to all my CCB colleagues, particularly Anna Dembo and Sriram Sundaravel. Also, thanks to the Chicago Center for Teaching and my colleagues there including Supriya. All of you have added so much to this experience.

All the members of the Greene and Huang labs have been wonderful to work with. I would like to especially thank Alex Ling for his help and feedback as well as Ya-Fang Chang who taught me everything I know about wet lab experiments. I would not have the skills I have today without these two. Montserrat Tijerina helped tremendously with the work presented in Chapter 4 and was the best mentee anyone could ask for. Lab mates Bradly Green and Muriel Laine made lab more like home. Thank you Justyna, Sean, Dave, Rosie, Paul, Aritro, Yingbo, Amy, Fan, Siddhika, Katie, Ross, Tina, Rose, Colin, Linda, Elizabeth, and Sydney for your support, whether practical or emotional.

Finally, I would like to thank my best friend and partner Emily Webster for her support, care, and love. From coffees and meetings to hikes, games, and much more, it has been wonderful to grow with her. Her enthusiasm and dedication inspire me to become better, strive further, and do good. Achievements are sweeter and the challenges easier with her by my side. I love you Emily.

ABSTRACT

(1) Triple-negative breast cancer (TNBC) is a severe clinical challenge in need of new therapies, but traditional drug development pipelines are time-consuming and expensive. More wholistic methods are needed to efficiently evaluate multiple drug targets in the context of TNBC. Drug response models aim to translate *in vitro* drug response measurements to *in vivo* drug efficacy predictions. While commonly used in retrospective analyses, my goal was to investigate the use of drug response modeling methods for the generation of novel drug discovery hypotheses in TNBC. (2) First, I review the current state of pan-cancer cell line screening datasets as these screening datasets are necessary for building drug response models. (3) Using one of these screening datasets, I generate models of drug response, which are then used to obtain imputed sensitivity scores for hundreds of drugs in over 1000 breast cancer patients. After examining the data for relationships between drugs and patient subtypes, I identified the Wee1 inhibitor AZD-1775 and an XPO1 inhibitor as compounds predicted to have preferential activity in TNBC. For AZD-1775, the imputed drug response formed significant associations with meaningful markers of drug response as well as the compound's mechanism of action. AZD-1775 also efficiently inhibited the growth of preclinical TNBC models. (4) XPO1 *in vitro* inhibition also associated with the TNBC subtype. RNA-Seq analysis implicated two distinct mechanisms for XPO1 inhibition-mediated cell death, with the TNBC-based mechanism being consistent with the pan-cancer gene set associations. (5) Overall, the work here develops a framework to turn any cancer transcriptomic dataset into a dataset for drug discovery and shows the framework's utility to quickly generate meaningful drug discovery hypotheses for a cancer population of interest.

CHAPTER 1: Background & Significance

"Yes, the thing about cancer is to cure it." (Huggins, 1979)

Near where I sit writing this document hangs an old poster from Charles Huggins's former lab. The poster also starts with this quote. It's evocative. It is simple and gets at the heart of translational research. It also evokes the period, a time when cures were discussed, and maybe magic bullets were just around the corner for every cancer. It comes from the context of treating hormone-dependent cancers where a new and clear dependency had been discovered such that "following hormonal intervention, the investigator sees cancer melt away in man and animals" (ibid).

In the 42 years since Huggins published that quote, there has been much progress in understanding the complexities of cancer. We have invented a great number of therapies in the chase to "cure" the disease. But we have also watched and identified resistance mechanisms as tumors evolve from these therapies, even the cancers that were once hormone dependent. For several cancers, we have even begun move away from the notion of "curing" the disease. Some have suggested the goal should not be to "cure" but to manage cancer, to continually adapt therapy so that the patient survives long enough for something else to come along.

With trying to understand the landscape of cancers and tumor evolution, we have generated massive amounts of cancer "-omics" (genomics, transcriptomics, proteomics) data. Patient tumors can be routinely sequenced to provide a list of DNA mutations or a snapshot of RNA expression patterns. There is so much data, but with so much data it is easy to get lost in the forest of cell and molecular biology and not be able see through to an appropriate therapy. Triple-negative breast cancer (TNBC) is a perfect example. There have been copious studies

describing this subtype of breast cancer and thousands of TNBC tumors have been sequenced and audited, but this has translated to relatively few therapeutic options for patients. This disease is, unlike most breast cancers, never hormone dependent and is still waiting for its "cure."

Still, my research goal is not far off from the sentiment expressed by Huggins in this quote. The ultimate goal of my research is to identify therapies that could lead to the more effective treatment of TNBC. By directly integrating the plethora of patient data with preclinical drug response data, I aim to probe drug relationships in the context of patient molecular information. My hypothesis is that drug response modeling can be used to facilitate the translation of preclinical drug data and be used as a novel form of drug discovery.

To this end, I will begin by reviewing the existing literature on TNBC as well as drug response modeling. This will provide the necessary context for my dissertation work that reviews the available datasets for drug response modeling, uses drug response modeling to generate drug repositioning and drug discovery hypotheses in TNBC, and subsequently validates these hypotheses in the search for new therapeutic interventions for these patients.

TRIPLE-NEGATIVE BREAST CANCER

Breast cancer is the most common cancer among women and accounts for the second most deaths among women. Breast cancer has perhaps one of the longest histories of personalized medicine in that molecular subtypes of breast cancer have been long established. Patients are stratified into Hormone Receptor (HR) positive (estrogen receptor (ER) or progestogen receptor (PR) expressing), HER2 positive (Human Epidermal Growth Factor Receptor expressing) or "other" classifications. This other category is labelled Triple-Negative Breast Cancer (TNBC) for lacking expression of any of the three previously mentioned

oncogenic receptors. Each subtype also gets a corresponding therapy. ER and PR-positive tumors get treated with hormonal therapy, most commonly selective ER modulators or degraders. HER2-positive patients receive and anti-HER2 therapy in the form of a kinase inhibitor (such as lapatinib) or a monoclonal antibody against HER2 (e.g. trastuzumab). On the other hand, TNBC cancers are inherently difficult to treat, as unlike the other breast cancer subtypes, they are defined solely by the absence of a distinct molecular target.

TNBC comprises approximately 10-20% of all breast cancer cases. This number can drastically fluctuate among different populations. In African Americans, the number is estimated to be closer to 25-30 percent, with some African communities having upwards of 46% of all cases be TNBC (Siddharth and Sharma, 2018). Latina women (Serrano-Gómez, Fejerman and Zabaleta, 2018) and younger women (Newman et al., 2015) also have an increased proportion of TNBC patients. Although a smaller segment of the overall breast cancer population, TNBC accounts for a disproportionate number of breast cancer deaths. TNBC has the lowest 5-year overall survival among the breast cancer subtypes regardless of race or ethnicity (Howlader et al., 2018; DeSantis et al., 2019). Beyond this, TNBC is known for its aggressive behavior and has been associated with high mean tumor size, higher grade of tumors at diagnosis as well as increased recurrence rate and metastasis after diagnosis. Metastasis is also more likely to occur in the lungs and brain when compared to ER+ disease (Aysola et al., 2013).

Part of the explanation for the aggressive behavior is biological and caused by a diverse molecular landscape. TP53 is the most commonly mutated gene in TNBC, and occurs at a rate of approximately 70-88% of all TNBC tumors compared to an overall prevalence of only around 30% in all breast cancer patients (Abubakar et al., 2019). However, aside from TP53

mutations, only PIK3CA, PTEN, KMT2C and RB1 have been identified as mutated in greater than 5% of TNBC patients. MYC amplification is the most common copy number alteration in TNBC, and in one study occurred in 81% of patient samples. Other common copy number alterations (occurring in 40-58% of patients) were amplifications in in E2F3, IRS2, CCNE1, EGFR, NFIB, CCND1 and MYB and losses in CHD1, PTEN, RB1, and CDKN2A. (Jiang et al., 2019; Zhao et al., 2020)

Despite some recent progress, TNBC patients continue to have the worst 5-year overall survival among breast cancer patients, and most TNBC patients are still treated with cytotoxic chemotherapies (Li et al., 2017). There is a clear and present need to identify new and effective therapies for TNBC to help reduce morbidity and mortality in these patients.

IMPUTING DRUG RESPONSE IN PATIENTS

Traditional drug development pipelines are time-consuming and take years for target identification, validation, and subsequent design optimization of the lead candidate compounds (Ashburn and Thor, 2004). While these approaches are indispensable for generating new therapeutic compounds, methods are needed to holistically explore and expand the potential use of existing drugs to different cancer contexts. The high costs, low success rates, and protracted development time for establishing new clinically-viable compounds has generated interest in expanding the use of (utility extension) and finding new uses for (repurposing/repositioning) existing drugs (Pushpakom et al., 2019; Wong, Siah and Lo, 2019). The challenge that remains is to identify appropriate contexts for drug repositioning and utility extension. Pathway mapping and signature-based approaches are both examples that utilize gene/protein expression patterns to identify such opportunities (Hurle et al., 2013). In vitro screening is another common approach to test existing drugs for phenotypic changes in cancer cell lines (Corsello et al., 2020).

In the area of precision medicine, these in vitro screening data are used as the inputs in machine learning models that aim to obtain accurate predictions of patient drug response. Researchers have developed many ways to build models depending on the type of input data and desired algorithm (reviewed in (Azuaje, 2017)). In the lab of Stephanie Huang, we previously established a general approach to impute/predict drug response in patients that was shown to be accurate in retrospective analyses of clinical studies (Geeleher, Cox and Huang, 2014). Our method involves building predictive models between baseline gene expression values in cell lines and their respective drug efficacy metrics (e.g., EC50 or AUC). In our original publication, this modeling approach was shown to be equally good or better at predicting patient response as the gene signatures derived directly from the clinical datasets. This retrospective study and others (Geeleher et al., 2015; Li et al., 2015) have shown that our methodology is accurate and useful for identifying meaningful relationships between drug response and patient populations.

Most advancement in drug imputation has focused on improving modeling methods so that researchers can better stratify "responder" from "non-responder." Recently we have begun to investigate extensions beyond obtaining accurate predictions of patient response. For example, we previously linked patient imputed drug response with genomic features and, in doing so, recapitulated known and discovered new biomarkers of drug response (Geeleher et al., 2017). Here I propose a novel use case for patient drug modeling: drug repositioning and utility extension. I hypothesize that I can flip the traditional paradigm of patient drug response modeling in order to identify drugs targeted towards a particular patient population. That is, instead of stratifying patients into responder/non-responder populations, I could begin with the patient population I would like to respond and test for compounds predicted to target this patient subset. Overall, I contend that drug sensitivity prediction methods can fill in the often-

missing pharmacological data from clinical patient datasets, providing a virtual drug screen of patients to hundreds of compounds and allowing for the identification of trends among imputed drug response, clinical features, and patient subtypes.

OVERVIEW

Drug response is complex. There are many pathway dependent factors that may lead to drug sensitivity or resistance as well as more general multidrug resistance mechanisms. Still, drug imputation can be quite accurate. This accuracy indicates that the modeling is able to condense large amounts of -omics and drug sensitivity data into drug response heuristics, something much more interpretable and more easily translated than multidimensional -omics data. Additionally, the molecular target(s) of a compound are not always known, but the mechanism of action is not needed for translational models of drug response. The models are built with whole-genome flexibility, which allows us to assess compounds regardless of how well they are annotated unlike in other more traditional forms of drug discovery hypothesis generation.

Drug response models can directly link biological differences to differences in drug sensitivity. Drug response modeling, then, has the potential to be highly translational. Through imputing drug response in patients via drug response modeling, we can transform any clinical cancer dataset to one primed for drug discovery. Questions, such as what patient clinical or genomic features associate with a compound's efficacy, can now be readily answered. Searching this data for drug repositioning and drug discovery opportunities could lead to the identification of compounds more tailored to particular patient populations and ultimately impact patient care. While this is a great theoretical potential for drug response modeling, testing is needed to determine whether modeling-based drug discovery hypotheses are meaningful. TNBC makes an excellent context for which to study drug imputation as TNBC is well-studied, biologically

distinct from other breast cancer, and lack effective targeted therapeutics. The significance of this work is high and two-fold: drug repositioning analyses in TNBC could lead to the identification of novel treatments and impact patient care for thousands of breast cancer patients as well as showcase a novel way to perform drug discovery in cancer.

My goal is to evaluate drug response modeling as a tool for drug discovery and use it in order to identify and accelerate therapies for TNBC. To understand the strengths, weaknesses, and biases of the cell-line based drug response modeling, it is important to first review the available datasets which serve to train the models (Chapter 2). Using these datasets, I built models of drug response and looked for compounds that could lead to the more effective treatment of TNBC in Chapters 3 and 4. Chapter 3 covers my work identifying AZD-1775, a Wee1 kinase inhibitor, for TNBC as well as validation of both the drug response model and the efficacy of AZD-1775 to inhibit growth of TNBC cells. In Chapter 4, I discuss the XPO1 inhibitor KPT-330 and the possible mechanisms of action that allows the for the efficient killing of breast cancer cells. Methods, results, and discussions are included for each individual chapter. Finally, Chapter 5 serves as a summary of this work, a discussion of additional uses for the imputation methodology, and a look-ahead at other future directions.

CHAPTER 2: Cell Line Datasets Review and Impact on Drug Response Modeling

2.1 INTRODUCTION

Cancer cell lines (CCLs) have been used in phenotypic screens for potential oncology compounds for decades. CCLs are perfect tools for drug screening as they are easy to grow in multi-well plates and easy to assay for phenotypic changes such as viability or apoptosis induction. Beyond this, CCLs are easy to manipulate genetically. shRNA and more recent advances in CRISPR technology have allowed us to knockdown or knockout every gene in gene in the genome in CCLs and measure the effects on the cells. Advances in liquid handling systems, such as acoustic dispensing of extremely small volumes, has made screening large number of CCLs against large compound libraries even easier, faster, and as a result more common place.

The history of publicly available CCL screens goes back several decades to the NCI-60. The NCI-60 began in the 1980s with the goal to evaluate and facilitate the translation of drugs in oncology. As a public platform, any drug could be submitted to screen against 60 CCLs (the CCLs have changed over the years, but most have remained the same). To date, the NCI-60 initiative has screened over 100,000 compounds with much of that screening data available publicly. This screen was the first to perform an integrated analysis of gene expression and it pharmacological sensitivity data in a large panel of cell lines (Scherf *et al.*, 2000). The screen and this analysis is credited for making a number of important drug discoveries in cancer (Shoemaker, 2006).

Since 2010, many other institutes have provided their own phenotypic screening data. These screens have typically focused on expanding the number of cell lines screened against

each compound and providing genomic and transcriptomic data for better integration that allows for biomarker discover. Examples include the Broad Institutes Cancer Cell Line Encyclopedia (CCLE), which has been subsequently updated into the Cell Therapeutics Response Portal (CTRP) and the Sanger Institute's Genomics of Drug Sensitivity in Cancer (GDSC) as part of their Cancer Genome Project. These cell line screens have also provided genomic and transcriptomic information for a majority of the cell lines screened. This allowed the original publications to systematically identity markers of drug sensitivity in CCLs (Garnett et al., 2012) as well as identify mechanism of action for certain drugs by correlation analysis (Rees et al., 2016). Beyond compound screening, many efforts have been made to use siRNA and CRISPR, the most prominent publicly available dataset for these screens (called Achilles) is another effort from the Broad Institute.

For my dissertation work, it became necessary to investigate and become familiar with these resources as they serve as the training data for the imputation models. It is only with the plethora of genomics and drug efficacy data provided by these large publicly available screening datasets that we are able to build models that utilize the entire transcriptome of cells to model drug response metrics. Thus, it is important to know the limitations regarding how the screens were performed as well as potential biases that might exist within this data to know how the potential biases and limitations of the models.

In order to better understand these resources, I performed a review of the publicly available drug screening data with Jessica Fessler and Alexander Ling of the Huang lab. This led to the review Alex and I co-first authored entitled "More than Fishing for a Cure: The Promises and Pitfalls of High Throughput Cancer Cell Line Screens" (Ling *et al.*, 2018). Jessica started this

venture in a lab rotation where she helped to identify potential screening datasets to include. Alex and I took over the project where we identified additional datasets and summarized the data. In general, the summary of the shRNA and CRISPR datasets was performed by Alex, the summary of the compound data was performed by me, and the summary of the CCL data was a joint effort (see this chapter's methods for more detailed breakdown). While the backbone of our publication is similar to this chapter, there are a number of significant changes that make the review presented here new and unique. I performed and wrote up this updated analysis exclusively.

This chapter differs from the original review in a number of important ways. First, only the data from compound screens that use cell lines from multiple tissue of origins (pan-cancer) are reviewed since these screens are more relevant for the model building done in later chapters. For specific information on the genomic screens or on single-cancer screens (such as (Daemen *et al.*, 2013; Teicher *et al.*, 2015; van de Wetering *et al.*, 2015), etc.) or the genomic screens (such as (Cheung *et al.*, 2011; Aguirre *et al.*, 2016), etc.), the reader is referred to that the original review. Additionally, for the purposes of the original review, we included screening data for historic reasons which was removed for this dissertation (e.g. CTRPv1). Finally, and most importantly, since the publication of the review in 2018, there has been a significant change and update to the GDSC and PRISM datasets. Chapters 3 and 4 rely in particular on data from the GDSC and CTRP datasets, so additional details are given for these datasets in the text and figures. The data reviewed in this chapter reflects all of these changes as well as some additional analyses and as such is markedly different from the original publication, even though most of the overarching conclusions remain similar. A summary of all the publicly available datasets reviewed in this dissertation can be found in Table 2.1.

Institution	Data Set Title	# Cell Lines	# of Tested Compouds	Citation
Wellcome Trust Sanger Institute and Massachusetts General Hospital	GDSC1	987	367	Garnett et al., 2012; Iorio et al., 2016
	GDSC2	809	198	Garnett et al., 2012; Iorio et al., 2016
NCI	NCI60	74	49,278	https://wiki.nci.nih.gov/display/NCIDTPdata/
	NCI-ALMANAC	60	104 (5,334 combinations)	Holbeck et al., 2017
Broad Institute	CTRP v2	887	496	Seashore-Ludlow et al., 2015
	PRISM Repurposing	578	4,518*	Corsello et al, 2020
GlaxoSmithKline	GlaxoSmithKline	310	19	Greshock et al., 2010
Genentech	gCSI	429	16	Haverty et al., 2016
Institute for Molecular Medicine Finland	FIMM	50	52	Mpindi et al., 2016

Table 2.1. Available in vitro CCL Screen Datasets.
This table provides summary information for the CCL screens I review in this chapter. Cell line and compound numbers reflect the latest releases of each dataset (removing any duplicated cell lines or compounds). Further details for each study (except GDSC2) can be found in our published paper Ling *et al.*, 2018.

2.2 METHODS

Identifying High-Throughput Cancer Cell Line Screens

Using online search engines, PubMed, and previous reviews (such as as (Smirnov *et al.*, 2018)), we compiled a list of CCL screens with publicly available data for the review. For this chapter, I have removed screens that are simply older screens that have evolved into newer screens (e.g. CTRPv1 which has evolved into CTRPv2 or earlier versions of a dataset). A few other inclusion criteria were used: greater than 10 compounds screened, greater than 50 pan-

cancer cell lines used, and data that was publicly available and easily accessible. See Table 2.1 for the screens included.

Investigating Trends in Cell Line Data In Cancer Cell Line Screens

Alex Ling worked primarily on harmonizing and annotating the cell line information using Cellosaurus *(https://web.expasy.org/cellosaurus/),* BioSample (Barrett *et al.*, 2012) and COSMIC (Forbes *et al.*, 2017). The harmonized data from the original publication is also used here, though updated with the data from the new drug screening datasets. I also updated the ethnicity analysis using by integrating inferred genetic ancestry from (Dutil *et al.*, 2019). For ancestry, the predominant ancestry (highest percentage in any ancestry category) was used. European ancestry (North and South) were combined similar to the paper.

We both contributed to the published analysis of the trends in cell line data including the proportion of cell lines screened to proportion of cancer incidence and mortality from the (Siegel, Miller and Jemal, 2017) data, as well as basic distribution regarding the patient age when the cell line was collected, cell line gender, and the cell line ethnicity. The data presented here is an update on the original analysis given the inclusion criteria from above and was solely performed by me. Calculations and graphing were all performed in R, graphs were made using the ggplot2 R package (Wickham, 2009).

Annotating and Summarizing Compound Information

Compound names were obtained from the original cell line screening publications or relevant online data repositories. I then harmonized the compound identifiers to correct for inconsistent formatting and name usage. Using PubChem's Identifier Exchange Service (https://pubchem.ncbi.nlm.nih.gov/idexchange/idexchange.cgi), I identified synonyms for all named compounds in the original datasets and to convert these synonyms to PubChem IDs.

After checking the results manually, some compounds still didn't match/harmonize between the datasets. To ensure the matches, I then took the list of PubChem ID and matched it back to compound synonyms and then once again matched that list to PubChem IDs. This created the highest degree of overlap such that even compounds that were highly related, for example irinotecan and its active metabolite SN-38, would then be identified as the same compound, which was deemed appropriate for our purposes. The results were further checked and curated as needed to ensure correctness. For information on the molecular targets and clinical phase of the compounds, The Drug Repurposing Hub (Corsello et al., 2017) was used to add information for any drug that used in this dataset. For any drugs that weren't contained in The Drug Repurposing Hub, I used the annotated information from the original publication or the updated web portals if the datasets contained similar annotations. This resulted in a list of drugs with their original identifiers from each CCL screen, a harmonized identifier, their clinical phase, their general mechanism of action, and their specific molecular targets. This information is provided in our Ling et al 2018 publication as Table S3.

I used R and the msigdbr package (*MSigDB gene sets R package*, no date) to obtain genes and pathway information to map compound molecular targets onto pathways. I utilized the Canonical Pathways gene set (Milacic et al., 2012; Liberzon et al., 2015; Kanehisa et al., 2017; Fabregat et al., 2018) (http://www.biocarta.com/) from the Broad Institute's MSigDB database (Subramanian et al., 2005a; Liberzon et al., 2015) for the pathway analysis.

2.3 RESULTS

Trends and Details for Screened Cancer Cell Line

Cancers are heterogeneous, and so should the models of cancer. To recapitulate the diversity seen in cancer, CCL screens need to utilize a wide variety of cancer cell line types. The

choice of which cell lines to include in a screen is an important one. NCI-60 for example only uses 60 cell lines, so the number of cancers that can be represented is limited. Also, limitations on the number of models available for each tissue of origin vary. To continue on the example of NCI-60, this dataset includes 9 different tissue types, with tissues being represented by as few as 2 cell lines (prostate) or up to 9 (lung). Prostate cancer researchers have very few prostate cancer cell lines available compared to other similarly common cancers like lung and breast cancer. Other characteristics of the cell line could be important to consider besides just tissue of origin, but could include the age, gender, and ethnicity of the patient from which the cell line was obtained. All of these are clinically relevant features that should ideally match the characteristics of patients.

To assess the diversity in the CCLs screened in these pan-cancer datasets, I obtained annotated cell line data from all the datasets listed in Table 2.1. Among these datasets were 1,494 unique cell lines covering over 200 different disease classifications and 30 tissue of origins. These disease characteristics span from very common cancer types (e.g. lung) to rare cancers (e.g. leiomyosarcomas) and highly specific subtypes (e.g. Acute biphenotypic leukemia). To give better context to the distribution of these cancer types, I correlated the cancer tissue of origin with the American Cancer Society's (ASC) estimated number of cases and deaths for the year 2021 (Siegel *et al.*, 2021). Interestingly, the number of cancer cells screened for a tissue of origin was highly correlated with the estimated incidence and death rate (Figure 2.1 A-B, spearman correlation equal to 0.69 and 0.68 respectively). The pan-cancer drug screens have, it seems, captured part of the heterogeneity seen in cancer type distribution and prioritizing cancers for which the incidence and death rates are highest.

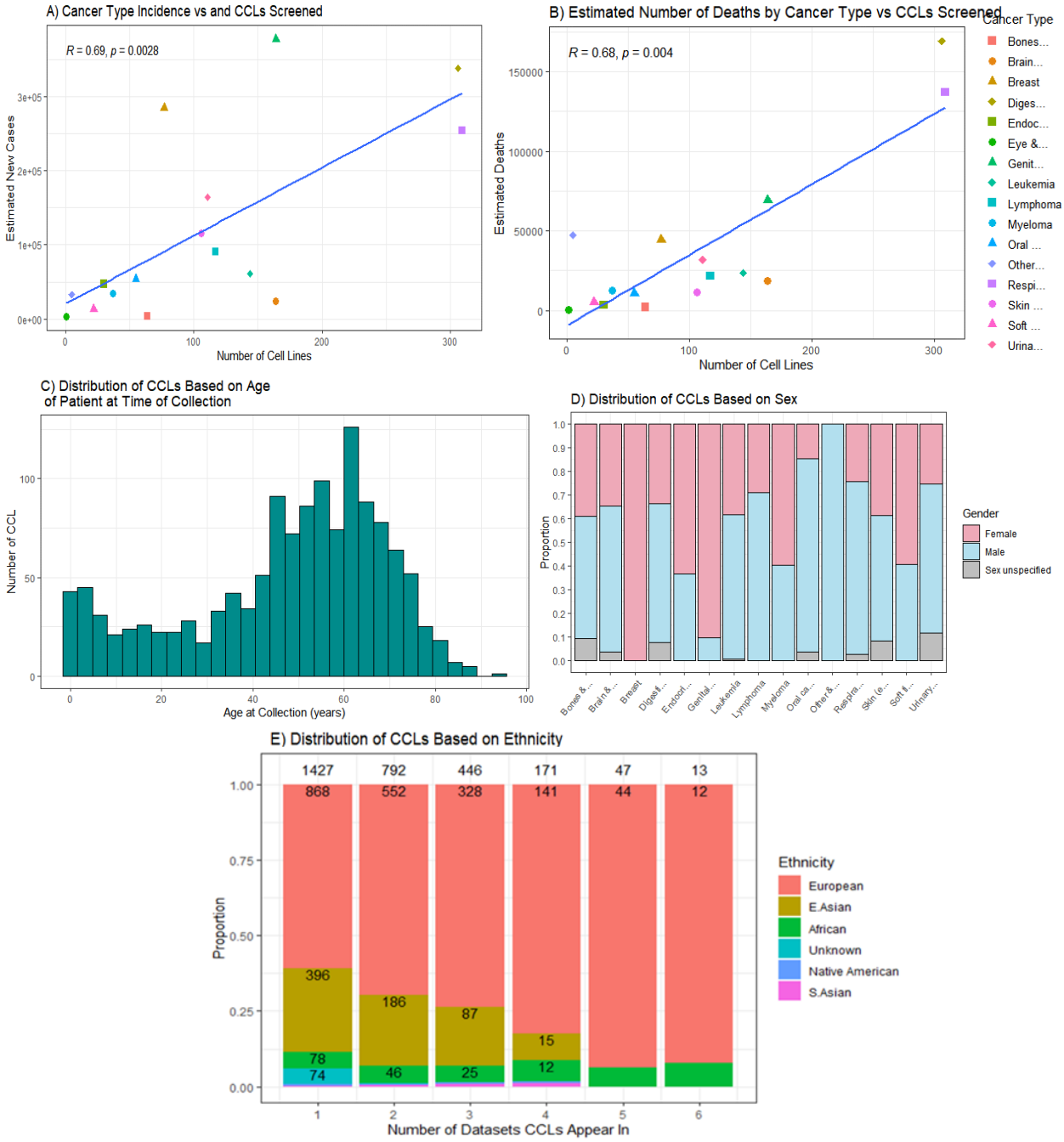


Figure 2.1. Screened Cell Lines Capture A Degree of Heterogeneity and Clinical Relevance Seen in Patients.

A-B Correlation between the estimated number of cancer cases (A) or number of estimated deaths (B) in 2021 for the indicated cancer types (obtained from the American Cancer Society and Siegel et al., 2021) and the number of unique cell lines screened from each cancer type. (**C**) Histogram of age of the patient at the time of CCL collection for all unique CCLs. (**D-E**) Proportional bar plots for the relative distribution of cell line sex (D) and ethnicity (E). For E, the number printed on top is the total number of CCLs with ethnicity information that appear in at least that number of datasets. For example, the first column represents the ancestry proportions used in all 1427 cell lines with ancestry information. The second column is the same for the 792 CCLs that appear in two or more institutional screens, etc. This is to show that European ancestry is overrepresented both overall and more so in the most commonly used CCLs.

Cancer heterogeneity doesn't just come from tissue of origin, but also other clinical patient features such as age, sex, and ethnicity. A histogram for the age of the patient from which a cell line was generated is shown in Figure 2.1C and does roughly correspond with expected patient ages with the mode centered around 60. Pediatric cancers are also fairly proportionally represented as an additional increase in the bars at less than 10 years of age. Broken down by cancer type, the two most common cancers types for the under 10 years of age CCL group are also from brain or leukemic origins.

However, for sex (Figure 2.1D) and ethnicity (Figure 2.1E) the breakdown isn't quite proportional to expectations. Sex is not always distributed evenly for all cancer types. For example, for oral cavity and pharynx tumors as well as the urinary system cancer categories, men do have an incidence rate that is 2.55 times higher and 2.40 times higher than females respectively and the CCL breakdown is aligned (more-or-less) with this breakdown. On the opposite side, the CCL gender breakdown for the genital systems category is more a showcase for how few prostate cancer cell lines are available so it is unsurprising that most CCLs in this category are female. However, for two categories there is some unevenness that can't be accounted for in this way: digestive and respiratory cancers. Looking that the cancer subcategory data, the second largest imbalance is that there are 2.88 times as many male lung cancer CCLs to female CCLs (216 male CCLs to 75 female CCLs) when the ACS data shows these to be nearly even in regards to both estimated new cases (119,000:117,000) and deaths (69,000:64,000). The largest subcategory imbalance is for liver cancer, which has 7.333 times as many male to female CCLs (22:3). Although liver cancers are more common in men, the amount is only 2.4 times for incidence (30,000:12,000) and 2 times for death (20,000:10,000). Since sex differences can affect the biology of cancer and therefore the response to treatment (Rubin *et al.*, 2020), it is important

to increase the diversity of these CCLs, especially if CCLs are going to be used to study the effect of sex-specific cancer phenomenon and drug sensitivity.

Ethnicity is generally poorly annotated during the creation of CCLs, so efforts have been made to impute ethnicity for these CCLs (Dutil *et al.*, 2019). It is readily apparent that most CCLs are from white European or Asian descent (Figure 2.1E). Not only this, but if we look at the CCLs that are most commonly used across the institutions (those that appear in 3 or more different screening institutions), White/European ancestry dominates the CCLs with 141/171 (82%) or 41/43 (95%) of CCLs being of European ancestry for CCLs that appear in 4 or 5 institutions respectively. It is certainly true that European ancestry is highly over-represented in CCLs overall and in particularly in the most commonly used CCLs. Additionally, for the Asian CCLs, over 80% of Asian CCLs are East Asian, specifically from Japanese origin. Looking at ethnicity by tissue of origin, 3 cancer subcategories, there were no CCLs representing African ethnicity. Hispanic and Native American categories are not consistently given in CCL ethnicity annotations, but based on the available data, only one category had a CCL of Hispanic ethnicity and five categories had a CCL of Native American ethnicity.

Compound Screens Cover a Diverse Set of Cancer-Relevant Targets

Apart from having a diverse set of CCLs that represent the heterogeneity seen in cancers, impacting a diverse set of molecular targets is an equally important aspect of CCL screens. All together there are over 50,000 unique drugs across all the screening datasets. However, this impressive number is almost solely due to the NCI-60, which has screened a large number of probes or chemicals without annotation data. All the other screens combined have only screened a total of 2,029 unique drugs. To assess the diversity of the gene targets of the compounds included in the 9 screens, I identified the molecular targets of screened compounds using the

Broad Drug Repurposing Hub (Corsello et al., 2017) and any target data provided by the screening datasets themselves. I was able to match clinical trial status, mechanism of action, and gene target information for over 2,000 compounds.

Regarding clinical stage, of the 2,029 compounds with annotated FDA clinical trial information, 843 compounds were already FDA approved and 623 more had been in some phase of clinical trial (Figure 2a). Surprisingly, the distribution shown in that figure doesn't change after filtering out the PRISM repurposing screen, which almost exclusively screened approved non-oncology compounds. A total of 1,538 unique gene products were impacted by the 2,363 compounds with annotated molecular targets. 910 genes were targeted by more than one compound, meaning 628 genes were only targeted by one drug in any of these datasets. It should be noted that compounds often had more than one molecular target, with 1237compounds annotated as hitting at least 2 protein targets. Figure 2.2B shows the ten most frequently targeted genes, which were each targeted by at least 38 unique compounds in the CCL screens I reviewed. Encouragingly, many of these top gene targets are recognizable as important in cancer. However, when overlaying these 1,234 genes to known cancer genes (either those frequently mutated or implicated in cancer), the resulting overlap was less than anticipated. I queried known cancer genes through two different resources: the Cancer Genome Atlas (TCGA) and the Cancer Gene Census (CGC) (Futreal *et al.*, 2004). Of the 127 most frequently mutated genes identified by the TCGA, only 45 were targeted by these compounds. Additionally, the Sanger Institute (CGC) has catalogued genes that have been causally implicated in cancer. Only 152 of the 723 CGC genes were targeted by at least one compound screened in CCLs. Both limitations on the number of compounds screened as well as general limitations regarding protein "druggability" likely play a role in explaining these proportions. Indeed, of the

targeted genes in CGC, close to 60% are classified as oncogenes while 15% were classified tumor suppressor genes.

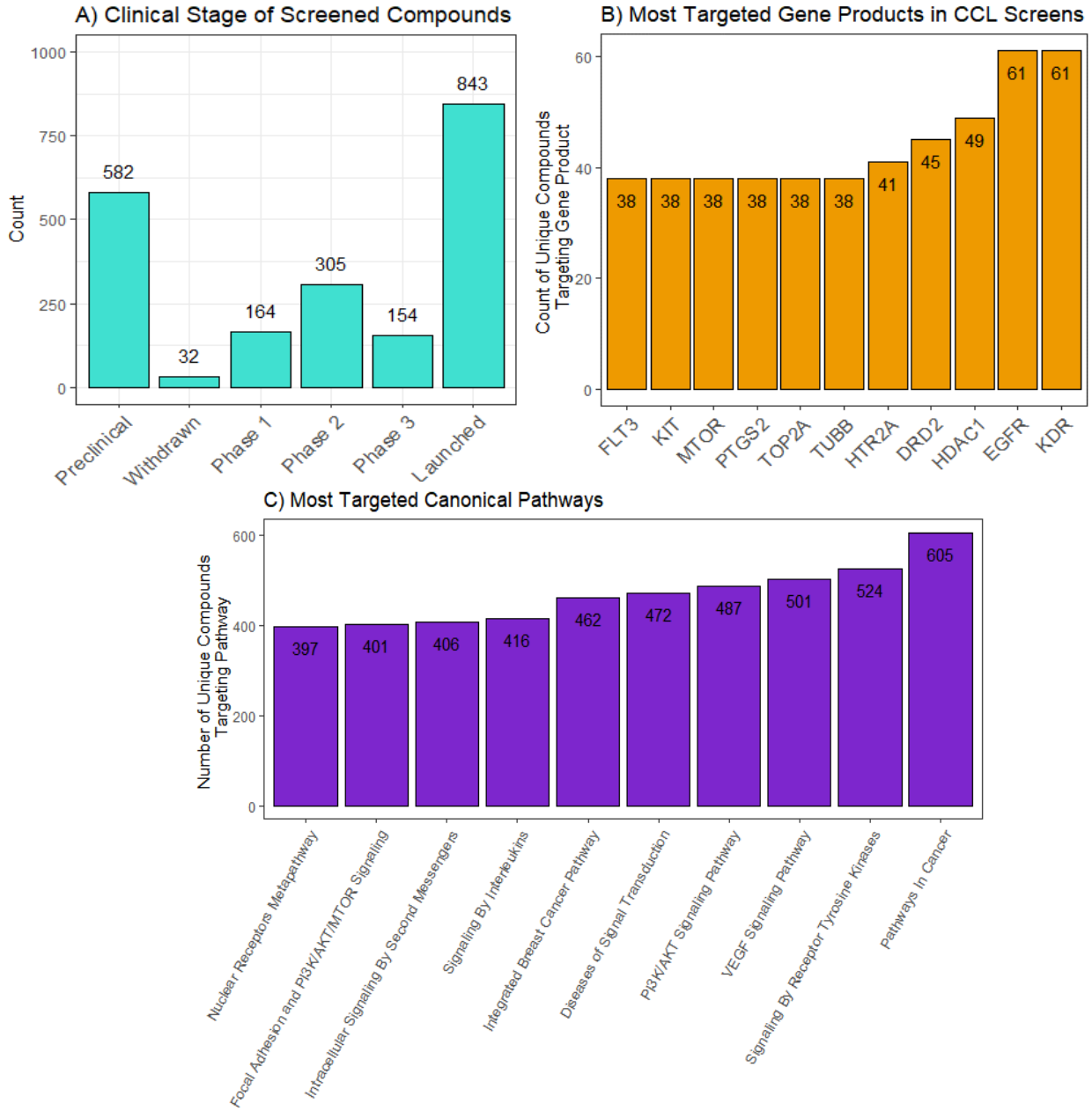


Figure 2.2 Targets and Clinical stage of compounds in CCL screens.
The compounds used in this figure are from the 9 CCL screens listed in Table 2.1. Only compounds that were able to be annotated with the relevant information are shown in each graph. **(A)** Bar chart showing the clinical stage distribution for the drugs 2029 drugs with annotated information. **(B)** Bar chart showing the ten most commonly targeted molecular feature and the number of unique compounds that target this protein. Molecular target data from original database or Broad DRH. **(C)** Shows the 10 most commonly targeted pathways in MSigDB's Canonical Pathway Gene Set (C2:CP) based on the number of unique compounds which target at least one gene target in that pathway.

To further investigate the role these druggable genes played in general cell biology and cancer pathways, I utilized the Broad Institute's MSigDB database, and, in particular, the database's Canonical Pathways gene set to represent general cell biology pathways. While the most commonly targeted canonical biology pathway was unsurprisingly "pathways in cancer," many other biologically significant pathways are also impacted (Figure 2.2C). Indeed, 2,480 of the 2,871 canonical biology pathways are impacted by at least one compound, with a median of 26 unique drugs impacting a given pathway. Regarding cancer specific pathways, 592 of the 620 pathways were impacted by at least one compound in the Cancer Modules (C4:CM, (Segal et al., 2004)) and Oncogenic Signature gene sets (C6). Overall, the coverage of the majority of the general biology pathways and cancer relevant pathways along with the proportion of drugs approved or in clinical trials suggests that CCL screens have, in general, selected a relevant yet broad array of compounds for screening. In these past two sections, I outlined the total coverage presented in these 9 datasets. Next, I will discuss the overlap among the screens for both the CCLs and the compounds to investigate the similarities among the datasets.

Cancer Cell Lines Have Considerable Overlap Among the Screening Datasets

In Figure 2.1, it was shown that together the different screening datasets cover a wide variety of cancer types. Figure 2.3 shows the coverage of cancer types for each individual cell line screen. Generally, the screens show similar proportion for all the cancer type covered with the most prominent deviations occurring the datasets with the fewest cell lines (NCI and FIMM screens). NCI-60 and the NCI-ALMANAC screening datasets comprise only 60 cell lines, which omit some more common cancer types such as sarcomas or pancreatic cancer cell lines. The FIMM dataset has the fewest cancer types available with only 5 cancer types available when accounting for the "other" category.

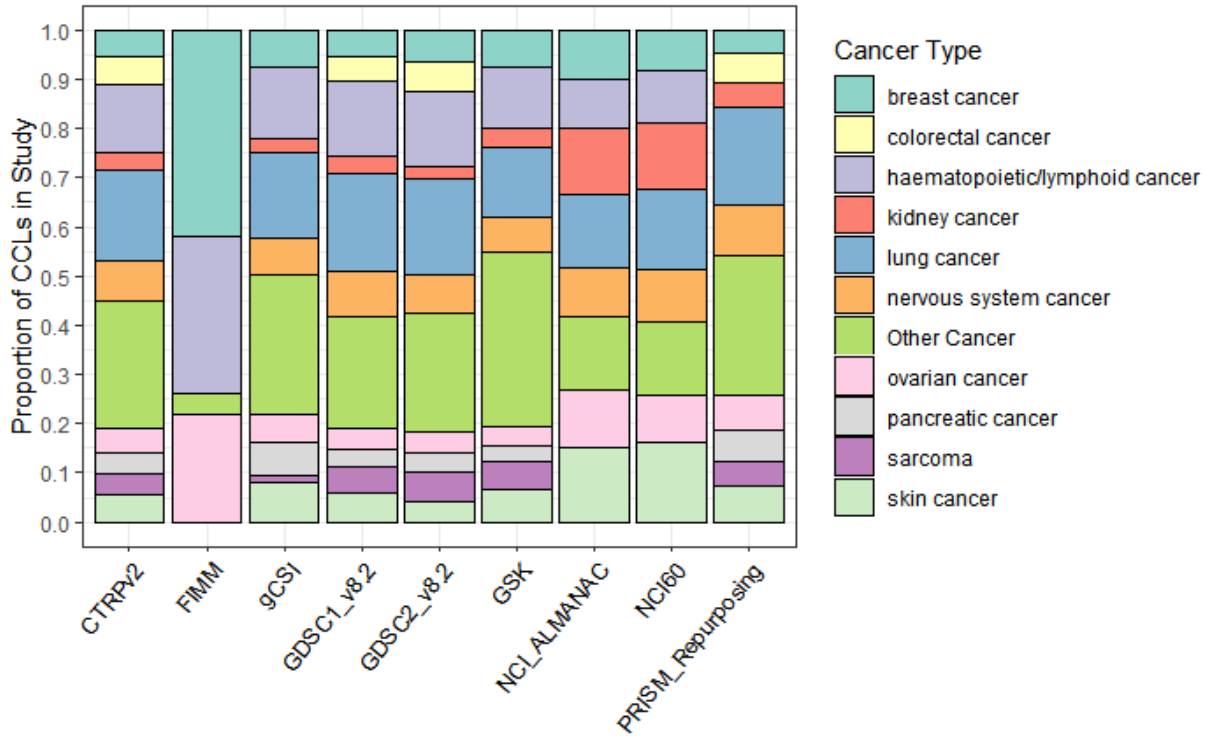


Figure 2.3 Tissue Representation Across CCL Screens
Cell line tissue type representation in each dataset. Tissue type was determined by bioinformatic and manual curation using Cellosaurus, the BioSample database, COSMIC, or annotations provided by the datasets themselves. Individual tissue of origin are shown except for grouping hematopoietic and lymphoid tumors as well as combining any cancer not represented into the "other" category.

Regarding individual cell lines, the datasets often containing a great deal of overlap among them (Figure 2.4A). Some of this is unsurprising given some of the datasets are from the same institution. For example, PRISM most highly overlaps with CTRPv2 which are both Broad Institute Screens and NCI-Almanac is of course comprised of the current set of 60 CCLs used in the NCI-60. Nevertheless, the overlap is quite high in general as can be seen in the heatmap by the coloring of cells below the diagonal. While the overlap between any two studies is generally over 50%, only 13 CCLs appear in all the datasets (Table 2.2). It is interesting to note that of these 13 CCLs, 12 of them are of Caucasian background. Most of these CCLs are female, but this is expected given the cancer types represented. Investigating further at the 252 CCLs that appear in at least 3 different institutional screens, male and female cell lines are roughly

equally represented. However, as noted before and shown in Figure 2.1E, 141 of these CCLs are of Caucasian background with 15 E. Asian, 12 African, 2 S. Asian, and 1 Native American.

CellosaurusID	CCL Name	Cancer Type	Age	Gender	Ethnicity
CVCL_0031	MCF7	Breast	69	Female	Caucasian
CVCL_0062	MDAMB231	Breast	51	Female	Caucasian
CVCL_0419	MDAMB468	Breast	51	Female	African
CVCL_0553	T47D	Breast	54	Female	Caucasian
CVCL_0004	K562	CML	53	Female	Caucasian
CVCL_1711	SR786	NHL	11	Male	Caucasian
CVCL_0465	NIHOVCAR3	Ovary	60	Female	Caucasian
CVCL_0532	SKOV3	Ovary	64	Female	Caucasian
CVCL_1304	IGROV1	Ovary	47	Female	Caucasian
CVCL_1627	OVCAR4	Ovary	42	Female	Caucasian
CVCL_1628	OVCAR5	Ovary	67	Female	Caucasian
CVCL_1629	OVCAR8	Ovary	64	Female	Caucasian
CVCL_0035	PC3	Prostate	62	Male	Caucasian

Table 2.2 Most Common CCLs across the Screening Datasets
CCLs were harmonized using the Cellosaurus with additional annotation provided by COSMIC or BioSample. The data was then filtered by the number of datasets a given CCL appears in. The CCLs here have been screened in all the institutional screening datasets.

Of the 1494 total CCLs across these datasets, 387 are unique to a single study (Figure 2.4B). This number is potentially skewed since several of these screens come from the same institution. When grouping the screens by institution from Table 2.1, 697 of the 1494 cell lines are unique to a single institution (Figure 2.4C). For the 387 CCLs unique to a single study, most of these come from CTRP as shown in Figure 2.4D, which is no surprise as it is the largest CCL screen. However, it is clear though that the overlap between GDSC1 and GDSC2 likely limit the number of unique CCLs in these datasets. When grouped by institution, both the Sanger (GDSC) or Broad (CTRP and PRSIM) screens have the most unique CCLs (Figure 2.4E).

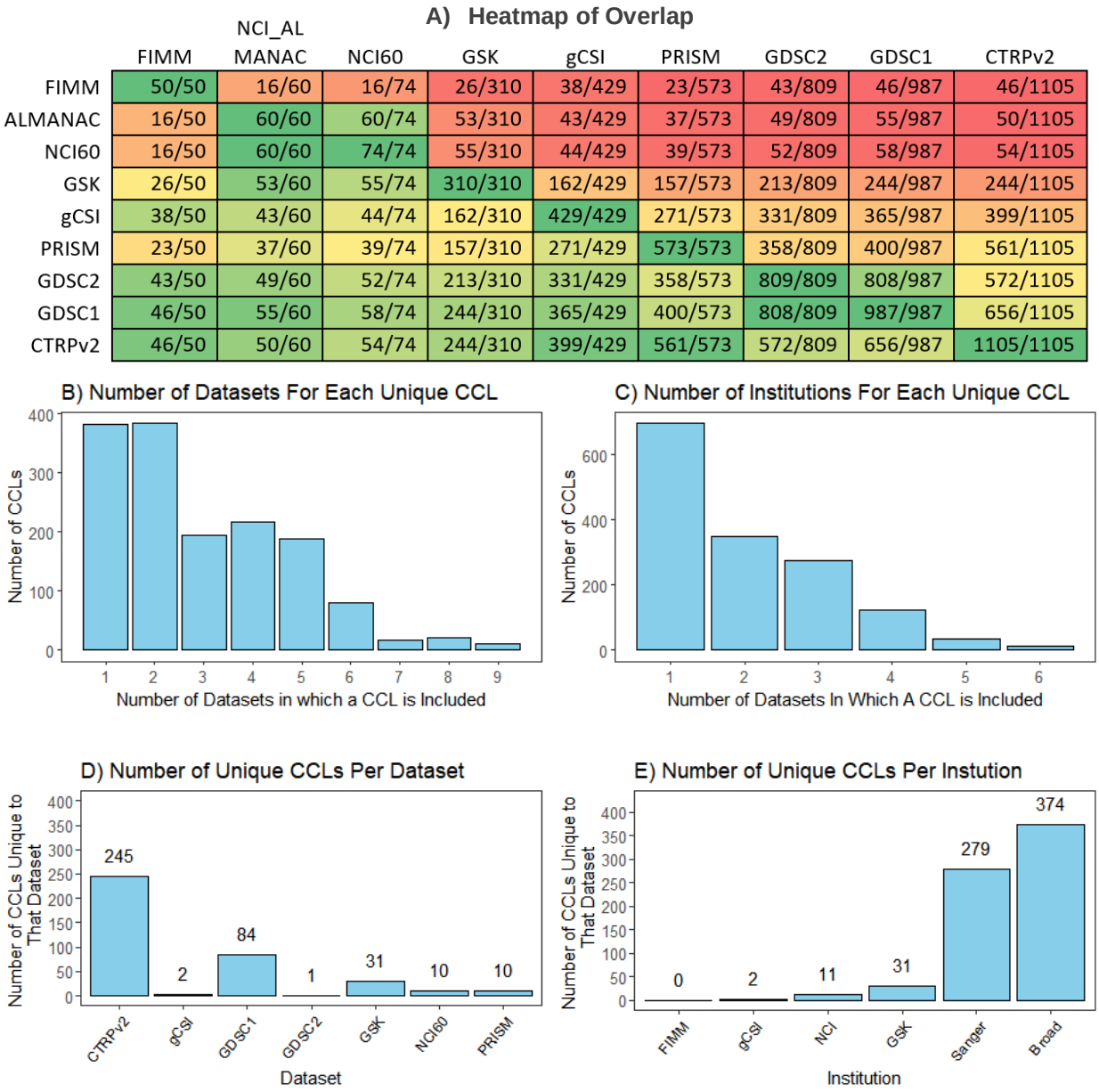


Figure 2.4. Cell Line Overlap/Frequency Among CCL Screens

(A) Heatmap of cell line overlap between reviewed studies. The heatmap columns and rows are organized from fewest to most CCLs, so anything below the diagonal indicates a true proportional overlap. The color values above the diagonal are both a size and overlap comparison. **(B-C)** Bar plots showing the distribution for the number of datasets (B) and number of institutions that any given cell line appears in. **(D-E)** Bar plot showing the number of cell lines that are unique to a single dataset (D) or institution (E).

Compound Overlap Among the CCL Screens

In total, there are over 50,000 unique screening agents with publicly available data in these datasets, most of which can be attributed to NCI-60. NCI-60 has data for close to 49,300 compounds with almost 49,000 of these agents being unique to the NCI-60 screen. However, it should be noted that the majority of these compounds failed to meet NCI-60's screening standards by either missing the minimum range requirements, not passing a minimum consistency among replicates, or by having results for fewer than 35 cell lines. Taking this into consideration, only ~21,000 compounds are both publicly available from the NCI-60 and passed their standards. Comparatively, the other CCL screens I reviewed tested a combined total of approximately 2,746 agents, of which about 2,029 are unique.

There is an appreciable amount of overlap among the compounds tested in the drug screens as shown in Figure 5A below the diagonal. It should be noted that the data from the FIMM and gCSI screens were released specifically to examine the reproducibility of cell line screens, with particular relationship to the Broad and Sanger Screens. As can be seen in the heatmap, these screens have the highest overlap with the GDSC and CTRP screens. GSK on the other hand was an early screen brought on by GlaxoSmithKline and only screened GSK molecules, so it is not unexpected that these 19 molecules don't appear with too much regularity in the other screening datasets. PRISM, which being a dataset focused on repositioning, actually has a fair amount of overlap with the other screening datasets. All in all, the overlap is not quite as high as that of the CCLs, but this is unsurprising given there are more potential molecular compounds than CCLs to choose from.

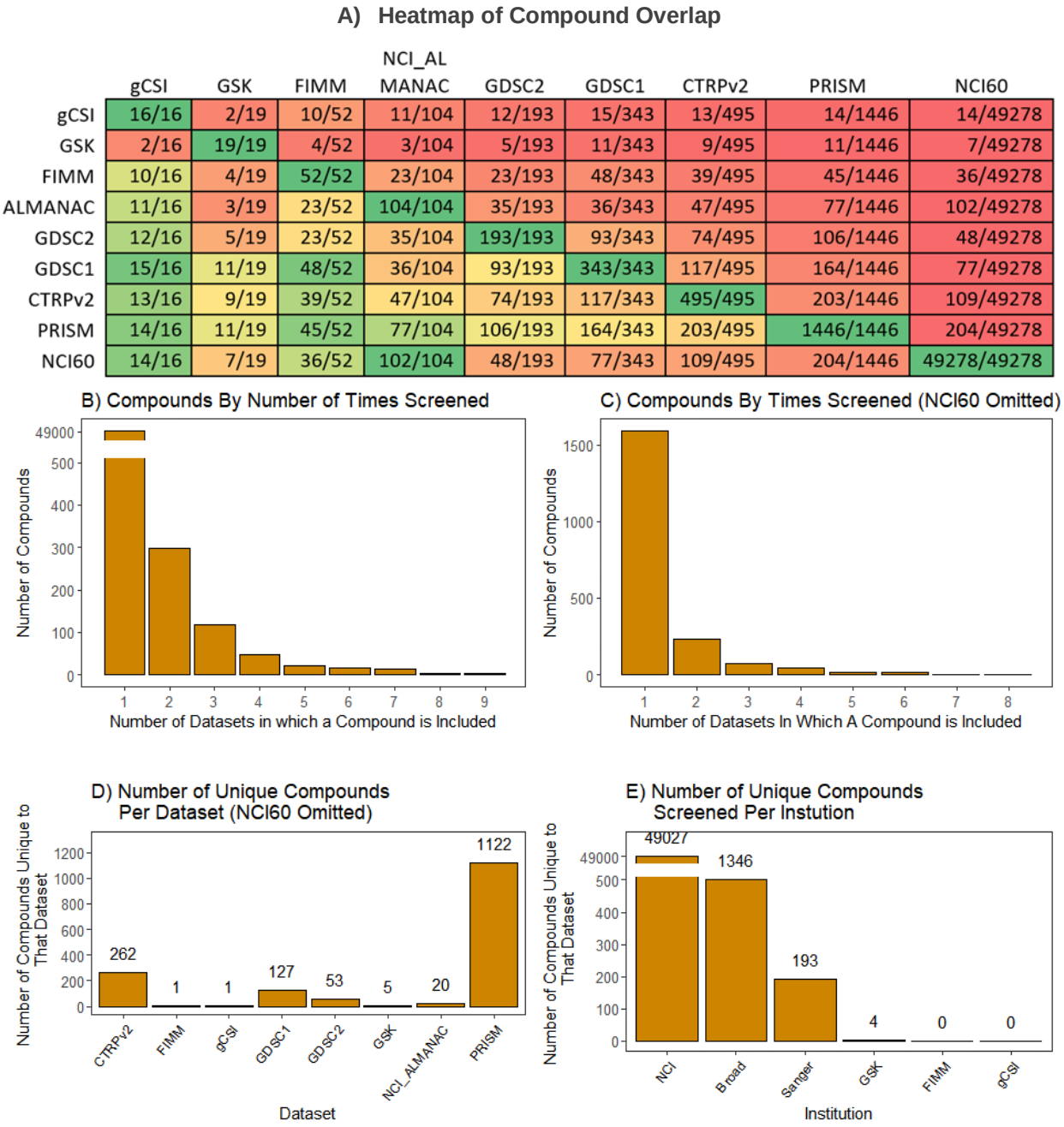


Figure 2.5 Overlap of Compounds and Frequency Among CCL Screens
(A) Heatmap of compound overlap between reviewed studies. The datasets (heatmap columns and rows) are organized from fewest to most compounds screened, so anything below the diagonal indicates a true proportional overlap. The color values above the diagonal are both a size and overlap comparison. **(B-C)** Bar plot showing the distribution of the compounds by the number of times they were screened across the datasets. (C) Same as B except the analysis was performed completely omitting NCI-60. (**D**) Number of compounds unique to a given screen, once again with the NCI60 data omitted for visualization purposes. (**E**) Number of unique compounds screened by every institution.

However, Figure 5B-C shows that there are a number of compounds that have been included in more than one screen. 518 drugs appear in at least 2 screens. However, if the NCI-60 is removed as an outlier for screening more than 49,000 compounds, then this number drops to 386 (Figure 5C). If I take into account the institution by grouping the screens as before, the number of drugs that have been screened by different institutions (still ignoring NCI60) is only 281. Surprisingly, 2 compounds were screened by every single screening dataset here: Lapatinib and Paclitaxel. Looking at the screens with the most unique compounds, this is clearly NCI-60 with over 45,000 unique compounds. For illustrative purposes, the NCI-60 was omitted for Figures 5D and it is clear that the PRISM and CTRP datasets have the next highest number of unique compounds. The NCI has by far screened the most unique compounds thanks to the NCI60 in particular, with the Broad Institute having the next most data on 1,300 unique compounds (Figure 5E). This is not to overlook the Sanger screens since the proportion of unique compounds in the GDSC screens is still fairly high (193 compounds unique of 443 total compounds between GDSC1 and GDSC2). While each of the major CCL screens contain unique information, enough overlap exists among the datasets to test the consistency of these screens and indeed models of drug response derived from this data (see discussion for section on Consistency Among Screens).

Only a Few Screens Are Appropriate for Model Building

Of the screens reviewed here, only three have screened a large enough number of CCLs for every compound to enable robust model building: GDSC1, GDSC2, and CTRP. These screens have a median CCL/drug of 831, 742, and 916 respectively. Each screen has a few compounds that have only been screened in a few hundred cell lines, even down to just 44 CCLs for one compound in GDSC2. However, the interquartile range for each of these screens are

similar: 880-926, 728-752, 700-851 CCLs/compound for GDSC1, GDSC2 and CTRPv2 respectively. In this way, the number of CCLs screened per compound doesn't differentiate among these or give preference to one for model building.

The choice of drug concentrations used in CCL studies is important. On one hand, the concentration should be relevant for the amount of drug a tumor would see in a patient to ensure the dosing is biologically relevant. However, there is an opposing and competing need to fit accurate four-parameter dose-response curves, which is often aided by having a larger dose range. GDSC in general has a smaller and typically more tailored dose range compared to CTRP. For example, the toxic chemotherapeutic paclitaxel has a maximum tested dose of 102.4 nM in GDSC but is tested up to 66,000 nM in CTRPv2. While likely more appropriate in this case, I looked at potential issues this may cause by comparing the reported EC50 value to the maximum tested dose range for each of these datasets (Figure 2.6). GDSC is highly skewed toward the right in these graphs indicating that most of the CCLs tested by these compounds are not responding at the maximum tested concentration. For example, 106/343 compounds in GDSC1 and 61/193 compounds in GDSC2 are screened such that 90% or more of the CCLs screened have a reported EC50 value above the maximum tested concentration. In comparison, only one drug in CTRPv2 is screened such that a majority (50%) of the CCLs have a reported EC50 value higher than the maximum screening concentration.

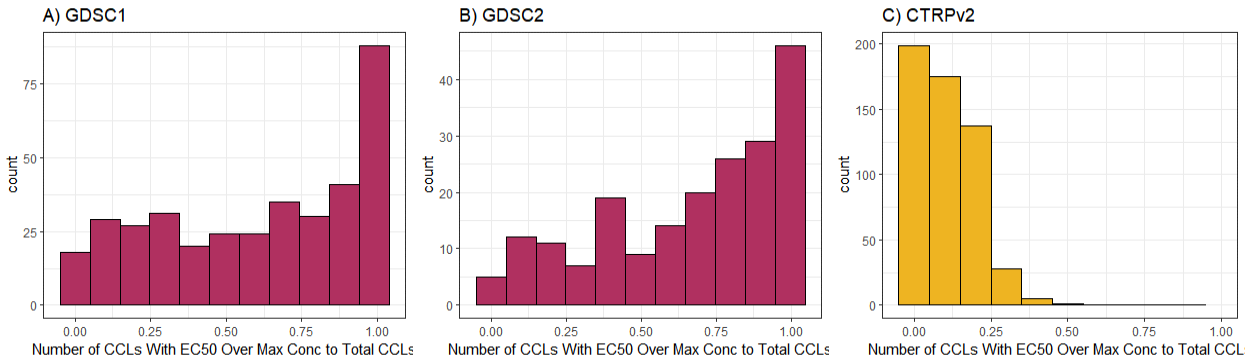


Figure 2.6 Dose Range vs EC50 values comparison among GDSC and CTRP screens
Data for these plots came from the most current release at the time of writing (April 2021). **(A-C)** Maximum tested dose and EC50 values were obtained for the dataset indicated. For every drug, the number of cell lines with a reported EC50 value over the maximum tested dose range was counted and then divided by the total number of cell lines screened against giving the proportion of cell lines that were resistant to that therapy. The plots show a histogram of that proportional value for all the drugs

2.4 DISCUSSION

There are many obvious benefits for having the data from these large CCL compound screens publicly available. As reported earlier, this data can be helpful to identify novel drugs for the treatment of cancer or a particular indication of cancer as well as biomarkers of response and even the biological mechanisms behind growth inhibition. These resources can be used for anything from drug discovery, to simply check the effective dose range of a compound for study design, to complex drug response modeling and identification of gene-drug regulatory networks. The utility of having screening data available for over 50,000 compounds (2,000 if NCI is not included) and drug response information on close to 1500 CCLs is obvious and shouldn't be understated. However, the limitations are more nuanced. In this discussion section, I will focus on the major limitations of these screens, the consistency of the data among the screens, as well as the implications for model building.

Limitations of Screens

There are always improvements that can be made to the number and diversity of the CCLs used in the CCL screens. In general, the CCLs reviewed covered a diverse array of cancer types. Many cancer types though are most often represented by only a few CCLs. While some of this is unavoidable in some cases like prostate cancer, this limits the direct study of biological specific phenomenon and drug interactions in CCLs (such as studying the AR-V7 splice variant's effect on drug response). The same goes for the ethnicity background of these cell lines, which is of increasing importance as more understanding is made on the effect ethnicity has on cancer progression and treatment response (Sekine et al., 2008; Keenan et al., 2015; Costa and Gradishar, 2017). Having a more varied set of CCL ethnicities would make it possible to determine if differences in ethnic background causes phenotypic changes (differences in drug response or genomic data) at the cancer cell line level. Even more concerning would be if the discoveries found in these CCL screens wouldn't translate to patients of African, Asian, or other underrepresented backgrounds. As cancer biologists, we often think about how well CCLs represent patients as a model of cancer, yet we rarely think about cell line sex and ethnic background. Still population-specific genetic variation contributes to health disparities in cancer, cancer risk, and outcomes and so these factors should be considered more closely in our models.

Molecularly, increasing the number and diversity of CCLs used would also be highly beneficial to the identification of biomarkers relevant to targeted therapies. Targeted therapies target particular vulnerabilities in cancer and as such are expected to only work for the subset of cell lines with that vulnerability. Indeed, previous reports have suggested that up to 85% of the cell lines tested in some screens are insensitive to the majority of tested compounds (Bouhaddou

et al., 2016), placing serious limits on the power to identify biomarkers associated with response to those treatments.

Fortunately, new efforts have begun to generate *in vitro* cancer models which capture this diversity (Williams and McDermott, 2017). The Cancer Cell Line Factory at the Broad Institute aims to generate more than 10,000 CCLs for research use (Boehm and Golub, 2015). The Human Cancer Model Initiative (HCMI) is another effort to increase the number of CCLs available. The HCMI is a collaboration between the NCI, Cancer Research UK, the Sanger Institute, and the foundation Hubrecht Organoid Technology, which aims to create as many as 1000 new in vitro cancer models with detailed clinical information, carefully controlled culture condition, and modern culture techniques such as conditionally reprogrammed cells and organoids (https://ocg.cancer.gov/programs/HCMI). Second, new types of model systems have recently become available that may increase the biological relevance and diversity of large screening dataset. Patient-derived tumor xenografts (PDX) have also been explored as a means of expanding the genetic diversity of pre-clinical drug screens (Gao et al., 2015). Similarly the establishment of organoid models and organoids based on PDX (also called PDOs) have been shown to model be promising models of in vivo response (Huang et al., 2020). Hopefully, these efforts will greatly increase the diversity and clinical relevance of available pre-clinical cancer models for future screens.

Consistency Among Screens

Concerns regarding the consistency of these CCL screen datasets heightened in 2013 when a study reported a large degree of inconsistency between the GDSC and Cancer Cell Line Encyclopedia (CCLE, which later spun off CTRP dataset). (Haibe-Kains et al., 2013). However, the statistical methods and approaches employed in this study were subsequently called into

question and multiple follow-up studies that reanalyzed the results concluded that both the pharmacological and genomic data are largely consistent and reproducible between these datasets (The Cancer Cell Line Encyclopedia Consortium and The Genomics of Drug Sensitivity in Cancer Consortium, 2015; Bouhaddou et al., 2016; Geeleher et al., 2016; Haverty et al., 2016; Mpindi et al., 2016; Pozdeyev et al., 2016).

Beyond reanalyzing the data, these follow-up papers also proposed potential reasons for any remaining inconsistencies. Aside from technical considerations, there was a consensus that a major source of variation was due to the abundance of insensitive cell lines for a majority of compounds tested. That is, often when compounds target specific cancer dependencies, their pharmacological metrics are dominated by cells insensitive to the compound's effects. When comparing the IC50 or area under the drug dosage and response curve (AUC) metrics, the few sensitive cell lines appear as outliers while the technical variability (noise) can then dominate the correlation of these statistics.

Additionally, problems with consistency are compounded when the datasets use different dose ranges or when IC50s are used for comparison. Mpindi et al., for example, investigated the effect of dose ranges by harmonizing the data to the same dose range, and found improved agreement between the datasets when these differences were accounted for. Several studies called into question the utility of the IC50 metric in comparing such large-scale datasets for two major reasons (Bouhaddou et al., 2016; Haverty et al., 2016; Pozdeyev et al., 2016). First, it was noted that extrapolating the IC50 when it lies beyond the maximum dose tested often leads to increased variability, which would further decrease consistency among the studies. Second, several studies argued that IC50s do not capture the diversity of pharmacological profiles, and

thus reanalyzed the data using variations on the AUC metric, which better combines information on the potency (IC/EC50) and efficacy (i.e. maximal activity value or Emax) of the drug (Bouhaddou et al., 2016; Haverty et al., 2016; Mpindi et al., 2016; Pozdeyev et al., 2016). Both harmonizing the dose range and analyzing the data with AUCs instead of IC50 helped to account for some of the variability expected from insensitive cell lines and thus allowed for a more accurate assessment of consistency.

Technical variability exists in all biological experiments, and CCL drug screens are no exception. A benefit of having overlap among these screens is that it allows for cross-validation when identifying novel drugs or cancer/molecular settings for existing drugs. However, in conducting cross-validation, one needs to keep in mind all potential sources of variability and take these into consideration when determining if diverging results from one study to another represents something truly biological. That said, all in all, the results from CCLE and GDSC (along with FIMM and gCSI) have been found to be largely consistent.

Choice of Screen for Model Building

To build the types of regression models needed for drug response modeling, both transcriptomic and compound efficacy metrics are needed. The only studies that provide transcriptomic information have been the NCI, Broad Institute, and Sanger Institute. Strictly speaking, transcriptomic information from the Cancer Cell Line Encyclopedia—Broad's attempt to categorize CCLs at the genomic, transcriptomic, and methylomic level—could be matched for any of the studies provided here. However, since it was shown that cell lines don't have complete overlap among the studies and since genomic drift in a cell line is possible between the datasets, using matched data is preferred. With this consideration and the need for large

amounts CCLs screened per drug, the best datasets for model building are the GDSC and CTRP datasets.

Regarding genomic information between these screens, there are some differences. The transcriptomic data from CTRP is RNA-Seq based and housed by the CCLE which does update, with the most recent release of the expression data being in 2019. Unfortunately, the screening data in CTRPv2 itself hasn't received many updates since its 2015 release data. For GDSC the opposite is true. The expression data is microarray based and was last updated in 2015 while GDSC compound screening data had its last major update in 2019 with the creation of GDSC1 and GDSC2, but has received additional screening data as recently as February 2020 (*News - Cancerrxgene - Genomics of Drug Sensitivity in Cancer*, 2021). It is unfortunate that the transcriptomic data isn't collected at exactly the same time the compound efficacy data is, though obviously this is wholly impractical for such large datasets. CCLs have been shown to drift in labs which can lead to differences in gene expression, cell morphology, and proliferation (Ben-David *et al.*, 2018). However, a recent study has also shown that while drift exists across datasets, in general there is only a small association between total genetic drift and differences between drug response (Quevedo *et al.*, 2020). Since these publications, efforts have been made to minimize drift in these large CCL screens and the evidence suggests that the major transcriptomic pathways in a CCL aren't altered by drift.

For model building, there is no evidence for better performance of models trained with microarray compared to RNA-Seq data, though RNA-Seq does have over twice as many features compared to the microarray data. This may lead to more relevant features being

included for a model but could also exacerbate problems of overfitting if not considered properly.

The largest difference between these two screens may come from the approach used for preferred screening concentration and efficacy metric. As seen in Figure 2.5, the screening concentration can make a substantial impact on the ability to call sensitive or resistant cell lines. Anytime the EC50 is over the maximum concentration, some interpolation of the graph needs to be made in order to estimate the EC50 value. This can be done to some accuracy after the maximum concentration; however, for many of these drugs, the estimated EC50 value is much higher than the maximum screening concentration. Of the 106 GDSC1 and 61 GDSC2 drugs with 90% CCL resistance, 75 and 60 of those drugs have an *average* screening concentration over 8 *times* larger than the maximum screening concentration. For many of these EC50 estimates, there is likely no way for them to be accurate and as such the values derived from any attempt to estimate them is going to be driven simply by noise.

Having even a majority of CCLs resistant to a particular drug is expected for some compounds, especially targeted therapy that may target a vulnerability present in only a few CCLs. However, these noisy values are likely detrimental to the modeling process. If the variability in 90% of the observations are driven by stochastic noise, the models I build using our current framework are simply not going to be biologically meaningful. Additionally, the preferred sensitivity metric is different for each dataset. GDSC typically reports EC50 values while CTRP reports normalized area under the curve (AUC) values. AUC values are bound between 0 and 1and represent the proportional area under the fitted dose response curve. AUC values are likely better in this situation since any value measured above the dose response curve

will have a value near or equal to 1. This helps with outliers as well, in the reported EC50 values in the CTRP data some of the measured EC50 values are over 10^{300} times larger than the maximum treated concentration. While such outliers typically don't exist in the GDSC data, using AUC values instead helps to ensure that outliers and the fit of a curve are less of an issue.

CHAPTER 3: Virtual Screening Breast Cancer Patients and The Identification Of AZD-1775 For TNBC

3.1 INTRODUCTION

In Chapter 1, I outlined the potential drug response modeling has to be used to identify potentially effective compounds for specific cancer types of interest. There is evidence that drug response models can capture some of the biology of the disease to allow for accurate predictions of drug response as well as biomarker identification. However, it had not yet been determined if I could flip the traditional paradigm of patient drug response modeling in order to identify drugs targeted towards a particular patient population. That is, instead of stratifying patients into responder/non-responder populations, can I begin with the patient population I would like to respond and test for compounds predicted to target this patient subset?

In Chapter 2, I outlined the landscape of drugs and cell line screening datasets that could be used for drug response modeling. With this information, it became clear that the data from CTRP was larger, provided more detailed RNA-Seq expression data, and had a larger screening range (i.e. potentially less noise) than the only comparable dataset, GDSC. CTRP then became the obvious choice for model building. The detailed patient information provided from the TCGA dataset made this 1,000+ breast cancer patient data suitable for imputing drug sensitivity. Even with a prediction accuracy of 75%, that would mean over 750 patients would have their drug response predicted accurately enough that trends between patient clinical features and drug response should become apparent. With over 1,000 patients and the potential to impute response for 496 compounds, this would lead to the potential for the interrogation of over 496,000 patient drug response scores. Effectively, this allows us to create a virtual screen of breast cancer patients and ask questions such as what drugs are predicted to have preferentially

activity in a particular breast cancer subtype or what breast cancer mutations are associated with this drug's predicted activity.

For this study, I aimed to identify compounds that could lead to the more effective treatment of TNBC. Using a candidate drug as an example, I demonstrate the process of identifying a lead candidate drug and perform biomarker discovery for the drug of interest. Then, I validate the method using an independent cell line drug screening dataset and use *in vitro* and *in vivo* experiments to explore the utility of the candidate drug with existing standard of care treatment. These results are presented in two phases: the discovery phase and the validation phase. Overall, I contend that drug sensitivity prediction methods can fill in the often-missing pharmacological data from clinical patient datasets, providing a virtual drug screen of patients to hundreds of compounds and allowing for the identification of trends among imputed drug response, clinical features, and patient subtypes. This analysis that makes up this chapter is similarly presented in my paper: *Facilitating Drug Discovery in Breast Cancer by Virtually Screening Patients Using in Vitro Drug Response Modeling* (Gruener *et al.*, 2021).

3.2 MATERIALS AND METHODS

Data Acquisition and Code Availability

The Broad Institute's Cell Therapeutics Response Portal v2 (CTRP) (Seashore-Ludlow *et al.*, 2015) AUC data were obtained from the Cancer Target Discovery and Development Network established by the National Cancer Institute's Office of Cancer Genomics (*https://ocg.cancer.gov/programs/ctd2/data-portal*, no date). The corresponding gene expression values for these cell lines were obtained directly from the Broad Institute's Cancer Cell Line Encyclopedia (CCLE) data portal (*Broad Institute Cancer Cell Line Encyclopedia (CCLE)*, no date). The Cancer Genome Atlas (TCGA) (Cancer Genome Atlas Research Network *et al.*, 2013) gene

expression data was downloaded from firebrowse.org and the clinical data (IHC status, PAM50 subtype, etc.) was downloaded using the TCGAbiolinks R package (Colaprico *et al.*, 2016). The Genomics of Drug Sensitivity in Cancer (GDSC) (Yang *et al.*, 2013) data was downloaded the GDSC website (*Drug Download Page - Cancerrxgene - Genomics of Drug Sensitivity in Cancer*, no date).

R Code to reproduce the entire computational analysis is available from the Github repository at (https://github.com/RFGruener/Gruener-et-al_2021). There, R scripts and additional documentation allows for the download of the CCLE, CTRP, GDSC, and TCGA data, CTRP/CCLE model generation and imputation in TCGA, statistical analyses to identify compounds of interest, and biomarker analysis.

Generating Models for Imputing Drug Response and Statistical Analysis

The methods for imputing drug response in TCGA patients using the CTRP/CCLE cell line data are based on those previously described (Geeleher, Cox and Huang, 2014). To summarize the methodology here, TCGA and CCLE expression data were filtered for common genes between the two dataset and then integrated using ComBat (Johnson, Li and Rabinovic, 2007). Feature selection was performed by removing 20% of genes with the lowest variation in gene expression across the samples. After a power transformation of the AUC values, a linear ridge regression model was fit between the CCLE gene expression and corresponding cell line AUC values from CTRP for every drug independently. Once the models were fit, I input the homogenized TCGA patient gene expression data into the models to obtain a drug sensitivity estimate (imputed sensitivity score) for each patient to every drug in CTRP.

Criteria for Lead Compound Identifcation and Statistical Analysis

Patients were grouped into clinical or PAM50 subtypes and the imputed sensitivity scores for each patient were compared using a two-sided Welch Two Sample t-test. For the

proof-of-concept comparisons for drugs effective in the hormone receptor-positive (HR+) setting, patients were separated into HR+ and TNBC, and t-tests were performed on the respective imputed sensitivity scores. For comparisons looking for drugs effective in the TNBC setting, patients were stratified into TNBC and non-TNBC groups and t-tests were performed on the respective imputed sensitivity scores. Given the large sample size ($n > 1000$) for these t-tests, the number of significant associations and degree of the significance could be quite high even after multiple-test corrections. This enabled us to be stricter in our criteria for compound-of-interest identification. For the HR+ analysis, only the 10% most significant compounds predicted to be more effective in the HR+ subset were investigated further. For the TNBC analysis, similar criteria were employed, selecting discoveries based on both a top 10% significance and a top 10% effect size thresholds. This second criterium was added because the effect size values were in general skewed towards TNBC for biologically unspecific reasons, as mentioned in the discussion. T-tests, *p*-value adjustments, and Spearman's correlation tests were performed using the base functions in R. Data was graphed using the package ggplot2 (Wickham, 2009). Mechanism of action, target information, and clinical phase were obtained from a recent review (Ling *et al.*, 2018).

Gene-Set Enrichment Analysis

Gene-set enrichment analysis (Subramanian *et al.*, 2005b) as performed using the software package GSEA v4.0.2 for Windows downloaded from gsea-msigdb.org. TCGA BRCA RNA-Seq data was used as the expression dataset, MSigDB's hallmark gene sets (Liberzon *et al.*, 2015) were used for the gene sets database, and patient imputed sensitivity scores to AZD-1775 were used as a continuous phenotype label. Default software parameters were used except Pearson correlations were used for ranking genes to reflect the use of a continuous phenotype label.

Obtaining Biomarker Associations Between Imputed Drug Response and Nonsynonymous Somatic Mutations and GDSC ANOVA Biomarker Associations

The associations between imputed drug response in TCGA and somatic mutations were calculated using linear models in R as previously described (Geeleher *et al.*, 2017). Briefly, gene mutation information was obtained from firebrowse.org (2016/01/28 release), which were summarized at a gene level with mutations called if a gene contained a nucleotide change that would affect the protein's amino acid sequence. I controlled for cancer type when the analysis was applied to all TCGA or PAM50 subtype for the TCGA BRCA cohort when specified in the text by including cancer type/subtype as a covariate (encoded as a factor) in the linear models.

ANOVA associations between drug response and TP53 for all 185 drugs in the GDSC2 dataset was downloaded directly from the GDSC data portal (*Cancer feature: TP53_mut - Cancerrxgene - Genomics of Drug Sensitivity in Cancer*, no date). The *p*-values were FDR corrected for the 185 associations tested.

In Vitro Cell Line Experiments

BT549, HS578T, and MDA-MB-231 cell lines (ATCC) were maintained in RPMI (ThermoFisher Scientific, Waltham, USA), DMEM, and DMEM (GE Healthcare Life Sciences, Hyclone, Logan, USA) media respectively. All media was supplemented with L-glutamine and 10% FBS (ThermoFisher Scientific, Gibco, Waltham, USA). For viability assay, cells were seeded at 5000 cells/well in 96-well plates. After 24 hours, the media was removed and replaced with media containing AZD-1775 at various concentrations between 0 and 3.2 μM, DMSO was used as vehicle and given in control wells. Growth was monitored every 4 hours to ensure control wells reached but did not exceed 95% confluence. After approximately 72 hours of treatment for each cell line, Cell Titer Glo® (Promega, Madison, USA) viability assay was performed as suggested by manufacturer. Luminescence values were obtained from VICTOR

Multilabel plate reader (PerkinElmer, Waltham, USA) and normalized to control well before plotting. Graphing and IC50 determinations were done using Prism 8 software (GraphPad, San Diego, USA).

Xenograft Experiments

All mouse studies were performed under the approved Institutional Animal Care & Use Committee (IACUC) protocol number 72037. C.B17 SCID mice were purchased from Taconic Biosciences. At 8 weeks of age, 3×10^6 MDA-MB-231 cells in PBS and Matrigel™ (Corning, Corning, USA) were injected into the mammary fat pads of the mice. When average tumor size reached 150 mm^3, mice were randomized into 4 treatment groups including vehicle and combination. AZD-1775 was received from AstraZeneca through the NIH's CTEP program, prepared in 0.5% methylcellulose solution, and delivered via oral gavage at 75 mg/kg on the first three days of the week for 4 consecutive weeks. Doses and schedule of AZD-1775 were suggested by AstraZeneca in order to best mirror use in patients. Paclitaxel from Teva Pharmaceutical (NDC 1703-4768-01) was purchased from the University of Chicago Pharmacy, prepared in PBS, and delivered by IP injection at 12 mg/kg on the first day of the week for 4 consecutive weeks. Tumor volume was monitored twice weekly by caliper and measured using the formula $\pi/6 \times L \times W^2$. Survival analyses are based on when tumors reached a study endpoint of 2000 mm^3. Graphing and statistical analyses performed using Prism 8 software (GraphPad, San Diego, USA).

3.3 RESULTS

Discovery Phase: Imputing Patient Response to Medications Enables the Discovery of Candidate Drugs for TNBC.

CTRP is the largest publicly available cancer cell line screening dataset with 496 unique compounds screened in 887 cancer cell lines. I used CTRP's publicly available in vitro drug response data and the corresponding RNA-Seq gene expression data from CCLE (Barretina *et al.*, 2012) for model building as described in the methods. Because each model is generated independently, the accuracy of these models can vary. Therefore, I performed a 20-fold cross-validation, and, of the original 496, only the 427 drug response models that had both a significant and positive Spearman correlation between measured and predicted response were further examined for the rest of this paper (Figure 3.1). I then applied these drug response models to the breast tumor RNA-Seq data from TCGA (Cancer Genome Atlas Network, 2012) to obtain a drug sensitivity estimate for each drug against each patient. The complete file of imputed drug response is available on the github repository or as supplementary table 1 of my associated publication (Gruener et al, 2021).

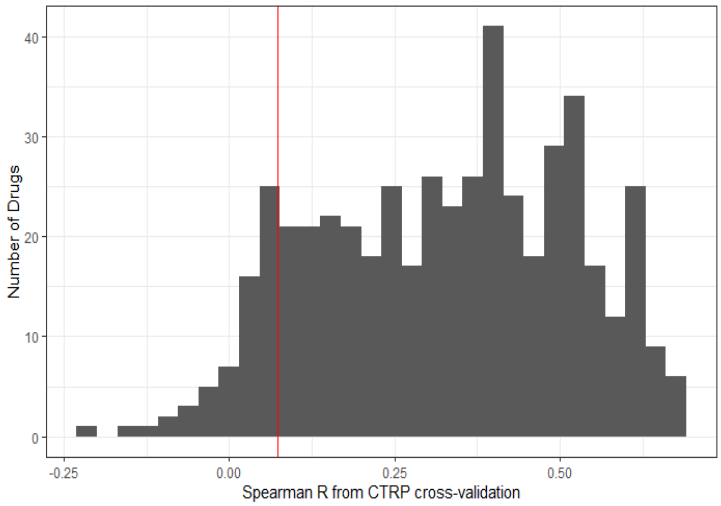


Figure 3.1 Distribution of Cross-Validation Results
The spearman rank correlation coefficient was determined for each of the 496 drug-response models based on a 20-Fold cross validation. A histogram of the correlation coefficients is shown. The red line indicates the model with the minimum spearman correlation that maintained significance. Roughly, every model to the right of this line was included in further analysis.

In order to discover drugs that are targeted towards TNBC, patients were stratified based on tumor IHC status for ER, PR and HER2 and patterns of imputed drug sensitivity in each subtype were compared. As a proof-of-concept, I first sought out drugs that were predicted to be targeted towards hormone receptor-positive (HR+, i.e., ER+ and/or PR+) breast cancers. By stratifying patients by their HR-positivity and comparing patient imputed drug sensitivities, I identified 11 compounds predicted to be preferentially effective in HR+ cancers (Figure 3.2). The two most significant results were a BCL2 inhibitor and tamoxifen, the standard-of-care ER antagonist. BCL-2 is overexpressed in 80% of ER+ cancers and inhibitors have already been investigated for HR+ cancers in clinical trials (Lok *et al.*, 2019). These results were encouraging and suggested that this approach could indeed identify relevant compounds-of-interest for a patient population.

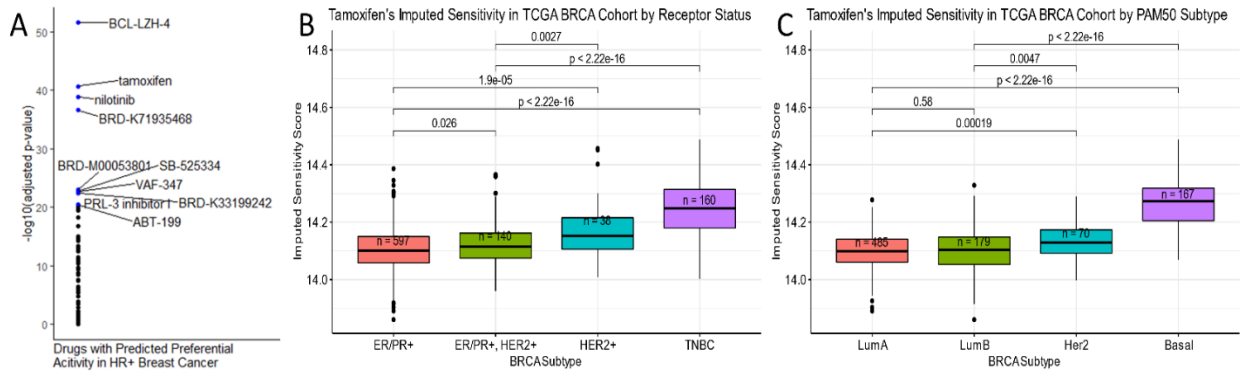


Figure 3.2. Proof of Concept: Drugs associated with ER+ Breast Cancers
Drug response models were built for 427 CTRP drugs and drug response predictions were made for every TCGA breast cancer patient. **A.** Patients were stratified into HR+ and HR- groups and t-tests were performed for every drug between the respective imputed drug response values. The results for all the drugs predicted more effective in the HR+ setting are plotted in the strip chart by the FDR adjusted p-value. **B-C.** Boxplots of tamoxifen imputed drug response with patients subtyped by either their IHC molecular status (B) or PAM50 subtyping (C).

I employed the same approach to identify compounds effective for TNBC patients. I dichotomized patients into TNBC and receptor positive (ER+, PR+, or HER2+; abbreviated RPBC) categories and looked for compounds that showed greater predicted efficacy in TNBC

compared to RPBCs by t-test. Figure 3.3A shows a volcano plot of the t-test results for all 427 drugs investigated. Due to the large sample size (n_{total} = 1095), 251 drugs showed higher sensitivity in TNBC when compared to RPBC at a Bonferroni adjusted *p*-value of less than 0.01 (see discussion). Thus, I chose to enforce a stricter threshold and focused on the top 10% of compounds showing higher predicted efficacy in TNBC based on both effect size and statistical significance, the results of which can be seen in Table 3.1. Of note, the most frequently represented pathway targeted by these compounds was cell cycle related (12 out of 17 drugs of interests; mechanism of action starred in Table 3.1). This agrees with previous studies that have identified the cell cycle as a vulnerability in TNBC (Hwang, Park and Kwon, 2019). Furthermore, all ten pathways identified by this analysis have been implicated as dysregulated in TNBC. Several of the candidate compound nominated by our approach have already been investigated in preclinical or clinical settings. References to the preclinical and clinical investigation of these compounds in TNBC can be found in Table 2. Overall, our approach appears consonant with and in support of the more traditional approaches that led to the identification of these drugs for TNBC in the literature; thus, substantiating the accuracy of our results and the potential of our computational approach to help speed up the drug selection pipeline.

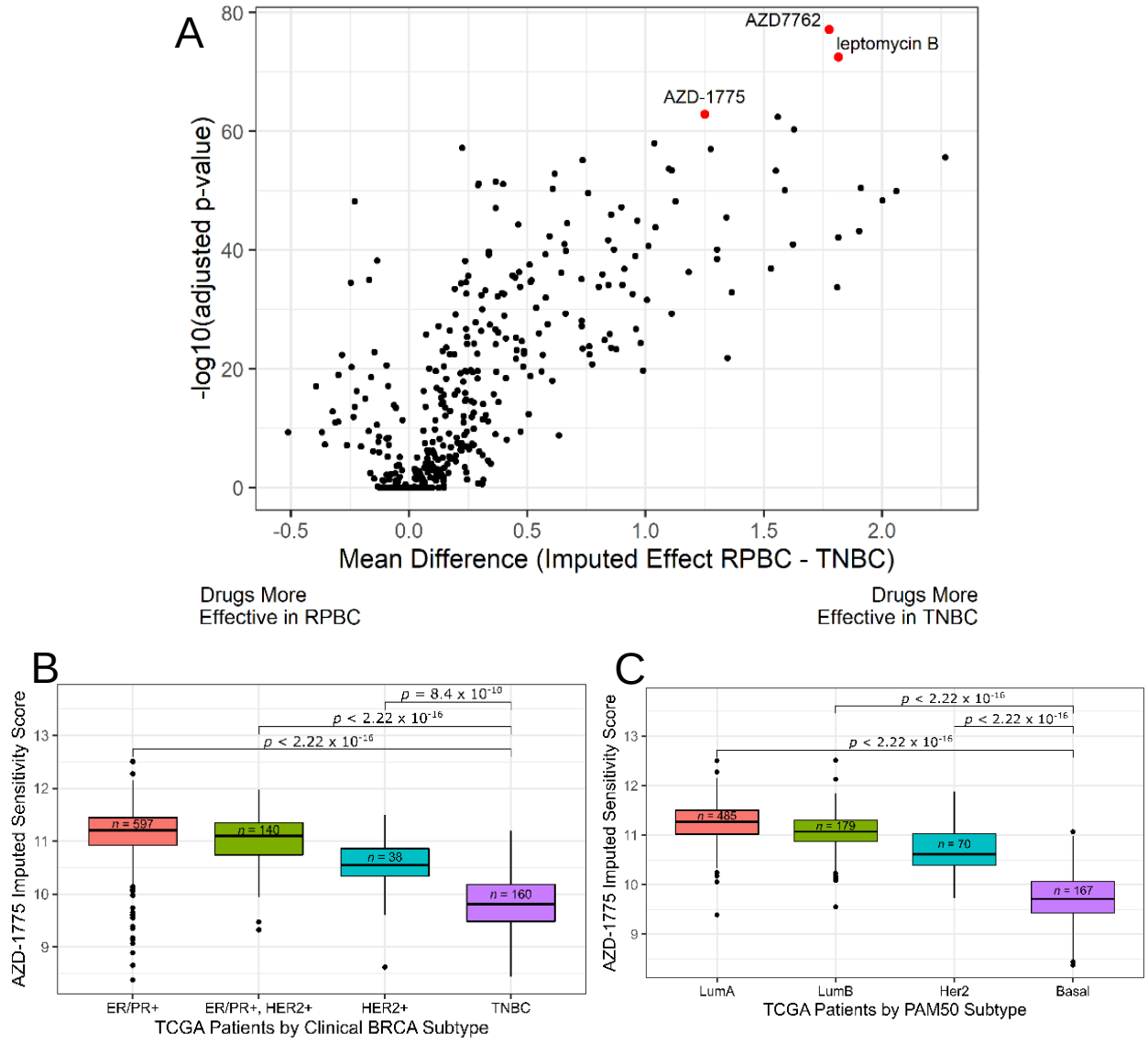


Figure 3.3 AZD-1775 is predicted to be most effective in TCGA triple-negative breast cancer (TNBC) patients.

(A) Volcano plot of TNBC vs. receptor positive breast cancer (RPBC, i.e., non-TNBC) imputed sensitivity t-test results for all drugs in CTRP. 427 drug response models were applied to the TCGA breast cancer RNA-Seq data resulting in an imputed sensitivity score for each patient. A t-test was then performed for every compound between the compound's imputed response in TNBC and non-TNBC (RPBC) patients. The p-values were Bonferroni-adjusted to correct for multiple testing. Highlighted in red are the top 3 most significant results (AZD7762, leptomycin B, and AZD-1775). (B,C) AZD-1775 imputed sensitivity in TCGA breast cancer tumors by receptor status (B) and PAM50 subtyping (C). Boxplots summarize results of each tumor sample's imputed sensitivity score to AZD-1775 in the TCGA breast cancer cohort by subtype. The n values indicate the number of patients in each group and p-values shown are adjusted for multiple testing. Lower values on the y-axis indicate increased predicted sensitivity. Dataset Abbr: TCGA, The Cancer Genome Atlas; CTRP, Cell Therapeutics Response Portal.

Mechanism of Action	# of Drugs in Top 10%	Total # of Drugs in Database	Drug(s) in Top 10%
CHK inhibitor *	1	1	AZD7762
exportin antagonist	1	1	leptomycin B
WEE1 kinase inhibitor *	1	1	AZD-1775
CDK inhibitor *	5	6	dinaciclib, alvocidib, SNS-032, PHA-793887, BRD-K30748066
translation (eIF4F complex) inhibitor	2	2	CR-1-31B, SR-II-138A
PLK inhibitor *	3	4	GSK461364, BI-2536, rigosertib
proteasome inhibitor	1	2	MLN2238
tubulin polymerization inhibitor *	1	4	docetaxel
phosphodiesterase inhibitor	1	2	ML030
kinesin-like spindle protein inhibitor *	1	1	SB-743921

Table 3.1 Drug pathways predicted to be effective for TNBC based on the differential imputed response analysis.

Table 3.1 contains the mechanisms of action (MOA) of the compounds that were in the top 10% most effective for TNBC based on both effect size and *p*-value from the imputed sensitivity analysis in the TCGA breast cancers (see also Figure 3.3). The rows are in ordered from most to least significant based on t-test *p*-value of the first drug listed in the drug column. A count column for the total number of drugs with the same MOA are also included. Asterisk (*) in Mechanism of Action column indicate drugs that target cell cycle/DNA repair pathways.

Drug	MOA Investigated in TNBC (Pubmed ID)	Clinical Phase	Preclinical TNBC Evidence for Drug's Use in TNBC (Pubmed ID)	Clinical Trial of Drug in TNBC
AZD7762	Reviewed in 30825473; 25104095, 22446188	Phase 1	25104095	
leptomycin B	24431073, 28810913, 30996012	Phase 1		
AZD-1775	Reviewed in 30825473; 29088738, 30181387, 29605721	Phase 2	29088738, 30181387, 29605721	NCT03012477
Dinaciclib	Reviewed in 28108739; 29144137, 27486754, 31704972, 29137393, 28678584, 25485498	Phase 3	27486754	NCT01676753, NCT01624441
alvocidib		Phase 2		
SNS-032		Phase 1	31704972	
PHA-793887		Phase 1		
BRD-K30748066		NA		
CR-1-31B	18644990, 19628077, 31106142	NA		
SR-II-138A		NA		
GSK461364	Reviewed in 30825473; 31751384, 30996295	Phase 1	31751384	
BI-2536		Phase 2	30996295	
rigosertib		Phase 3		
MLN2238	30400780, 23948298, 30601533, 25575864	Launched		NCT02993094
docetaxel	Reviewed in 26273192; 27966988	Launched	27966988	Many (e.g. NCT02413320)
ML030	27901486, 23536305	NA		
SB-743921	20068098, 29190901, 29535384, 24928852	Phase 2	29190901	

Table 3.2. Literature support for drugs identified by imputed drug response modeling
Table 3.2 contains references for the same drugs as Table 3.1. The references are grouped by investigation of the MOA to which the compound belongs with additional references for that specific compound in either preclinical or clinical setting. References are given as PubMed IDs for the MOA (Mechanism of Action) and preclinical columns, whereas clinical trials are given by their the clinicaltrials.gov identifier numbers (NCT). The furthest clinical phase the compound has been tested in is also given in the indicated column.

Of the most significant hits, the top two compounds—AZD7762 and leptomycin B—have been studied in clinical trials in cancer. However, the development of these two compounds was halted due to toxicities. Leptomycin B is an XPO1 inhibition and new inhibitors of XPO1 have been generated in recent years. In Chapter 4, I examine the use of XPO1 inhibitors in the context of breast cancer further. For now, the third most-significant hit, AZD-1775 (aka MK-1775 and Advosertib), was well tolerated in patients in a phase I clinical

trial in advanced solid tumors (Leijen, van Geel, Pavlick, *et al.*, 2016). AZD-1775 also had a high in vitro cross-validation score (r_s: 0.55, *p*-value 3.3×10^{-60}) and targets the cell cycle like many of the other most significant results. Finally, AZD-1775 was also consistently one of the most significant results based on both clinical subtypes based on receptor status (Figure 3.3B) and PAM50 subtype classifications (Figure 3.3C). For these reasons, I chose to focus on AZD-1775 for subsequent validation and to demonstrate the potential/feasibility of our drug selection/validation pipeline.

Discovery Phase: Identify Biomarkers for AZD-1775

Proof-of-concept: Tumors Predicted to be Sensitive to AZD1775 are Enriched with Cell Cycle Gene Sets

The primary target of AZD-1775 is the Wee1 kinase, which is known to play a critical role in inhibiting the cell cycle at the G2/M checkpoint. I hypothesized that, if our model is picking up on biological meaningful patterns, the RNA expression profiles of patient tumors predicted to be more sensitive to the AZD-1775 should be enriched for cell cycle gene sets. To test this hypothesis, I performed gene-set enrichment analysis (GSEA) on the TCGA breast cancer RNA-seq data using patient imputed response to AZD-1775 as the continuous phenotype label. Using the hallmark gene set, I found that tumors predicted to be more sensitive to AZD-1775 were enriched for the G2/M checkpoint signature (Figure 3.4A, FDR = 0.04). G2/M is the only significantly enriched pathway that associated with AZD-1775 sensitivity at an FDR of less than 0.05, indicating a specific and significant concordance between the imputed results and the biological action of AZD-1775.

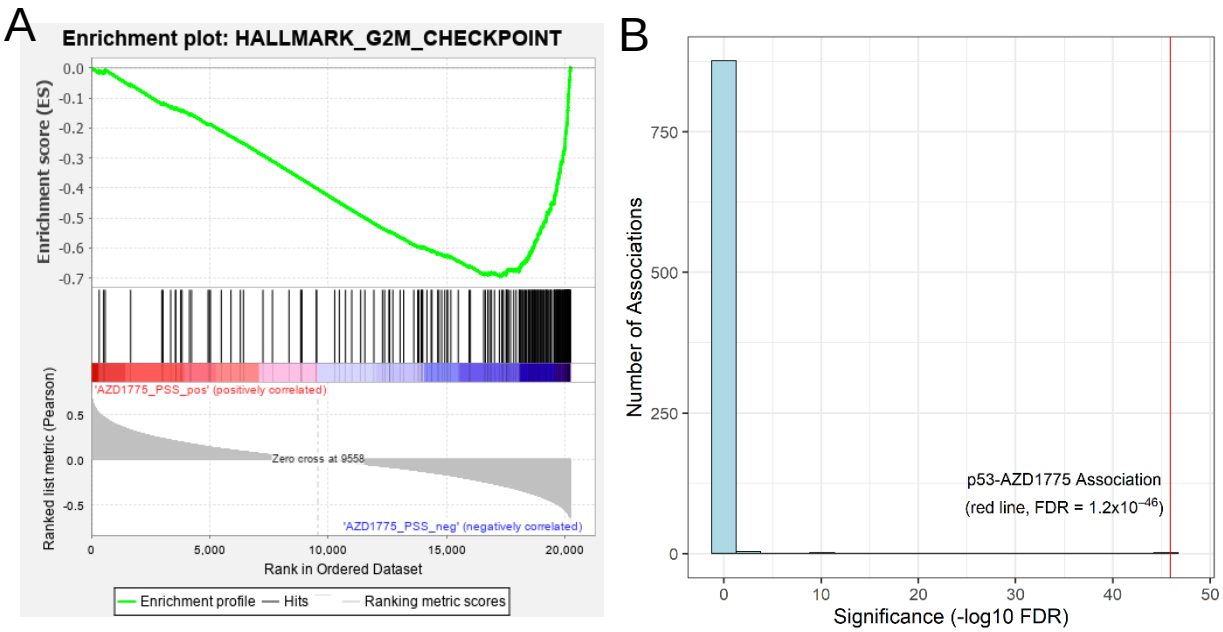


Figure 3.4 Biologically meaningful associations with AZD-1775 imputed sensitivity in TCGA breast cancers.
(**A**) Gene set enrichment analysis was performed in MSigDB's "Hallmark" gene sets using TCGA's breast cancer expression data with AZD-1775 imputed sensitivity score as the continuous phenotype variable for enrichment. The G2M gene set was significantly up-regulated in breast cancer patients predicted to be sensitive to AZD-1775 with an enrichment score of -0.695, normalized enrichment score of −1.86 and an FDR q-value of 0.04. A negative enrichment score associates with sensitivity to the drug as smaller imputed sensitivity values indicate more sensitive. This was the most significantly enriched for pathway. (**B**) A histogram of p-values achieved for all the associations between AZD-1775 imputed response and any gene with a somatic protein-coding change in at least 20 samples (n = 882 genes) in TCGA breast cancer cohort. TP53 mutation and AZD-1775 achieves the strongest association at an FDR = 1.2 x 10^{-46}, with the next most significant association at an FDR of 1.2×10^{-13}.

Imputation-Based Drug-Wide Association Analysis Reveals Potential Biomarkers for AZD-1775

I previously published a method to form associations between imputed drug responses and genomic features in a manner analogous to genome-wide association studies (GWAS) (Geeleher *et al.*, 2017). I employed this methodology (termed IDWAS for imputed-drug wide association study) to link genomic features with our imputed drug response to search for potential biomarkers of response to AZD-1775. Results from this analysis identified mutation status of 13 genes in the TCGA breast cancer cohort that significantly associated with AZD-1775 sensitivity at an FDR < 0.05. In particular, AZD-1775 response formed a highly significant

(FDR = 5.6×10^{-46}) association with *TP53* mutational status in breast cancer patients (Figure 3.4B). This p53-association holds when IDWAS was performed on the pan-cancer TCGA with tissue type correction (n = 9968, FDR = 2.1×10^{-8}) as well as in TCGA breast cancer with correction for PAM50 subtype (n = 1098, FDR = 1.6×10^{-9}). The effect size of this association indicates that AZD-1775 would be more effective in the p53 mutated/null setting, which is consistent with other literature that shows Wee1 inhibition in p53 mutated cell lines increases genomic instability and in turn forces apoptosis or mitotic catastrophe (Hirai *et al.*, 2009). I also investigated the genomic association data in the independent GDSC cell line screening dataset. Of the 185 drugs tested in this data, AZD-1775 was the only drug that significantly associated with increased efficacy in the TP53-mutated setting after multiple test correction (FDR = 0.035), giving further confidence in the association between TP53 mutation and AZD-1775 activity.

To evaluate whether the biomarkers were unique to AZD-1775, I performed the same IDWAS analysis with all 17 drugs from Table 3.1. I identified 38 genes that associated with these 17 compounds. Six gene mutations (e.g., *TP53*, *PI3KCA*, *MAP3K1*, and *NOTCH3*) associated with over half of all the drugs. This overlap is expected, as many of these drugs have similar mechanisms of action and are imputed to generally work better in the TNBC population. For example, the 8 genes that significantly associated with the PLK inhibitor GSK461364 also significantly associated with the other two PLK inhibitors.

Of the 13 significant gene-drug associations with AZD-1775, six genes were found to associate with four or fewer other drugs and one gene associated specifically with AZD-1775 (Figure 3.5). Of interest were the associations with *ANKRD30A*, *SPTBN1*, and *GALNTL5*, as these genes have known and well-defined cancer associations. *ANKRD30A* encodes for a breast

differentiation antigen, NY-BR-1, which is often dysregulated in breast cancer and has been investigated as a general breast cancer biomarker previously (Balafoutas *et al.*, 2013). *SPTBN1* and *GALNTL5* have also been implicated in breast cancer. Mutations in these genes have been associated with increase in metastatic phenotypes, which is of interest because both IDWAS associations indicate AZD-1775 would be more effective in the mutated (i.e. more metastatic) setting (Hussain, Hoessli and Fang, 2016; Chen *et al.*, 2020).

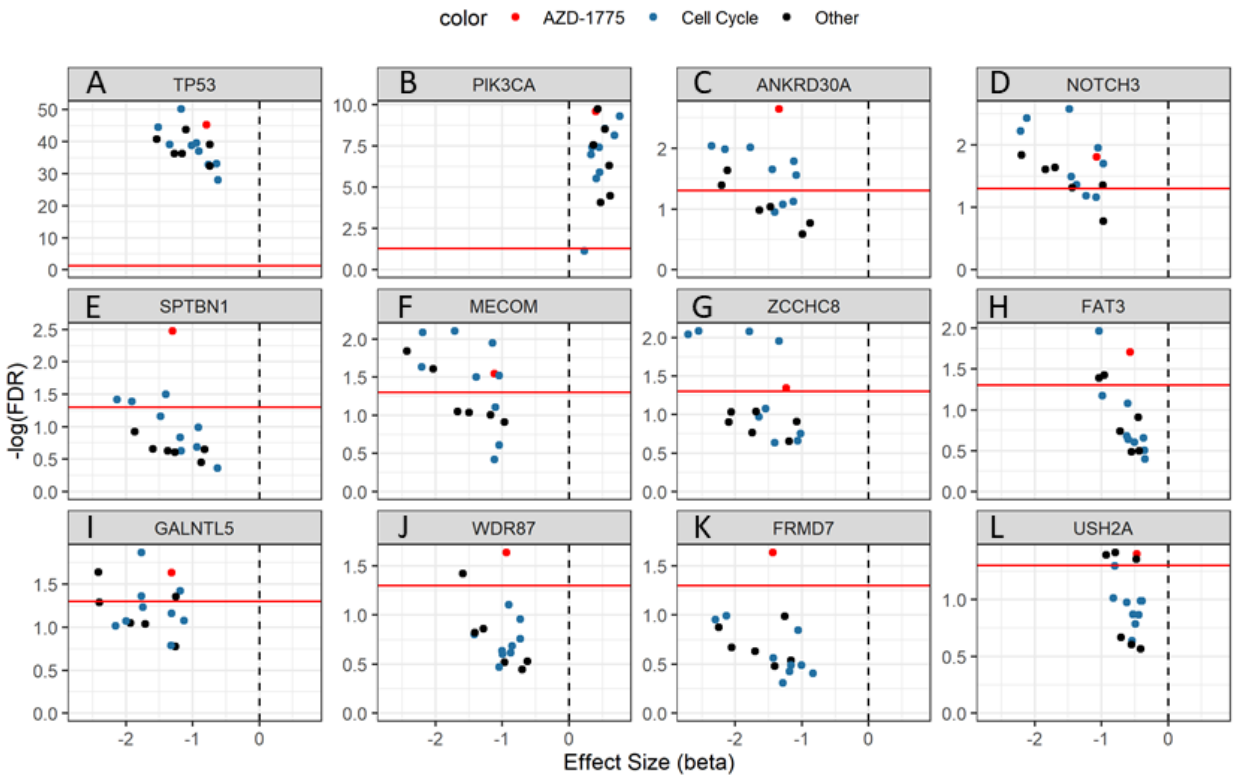


Figure 3.5. Genomic Associations of Imputed Response for All Table 3.1 Drugs.
(**A-L**) For each gene indicated, a volcano plot shows the significance and effect size (beta from a linear model fitting gene mutation by imputed drug sensitivity) of the gene-drug associations. The drugs are colored by cell cycle modulators (blue, see drugs starred in Table 3.1) or other compounds (black) with AZD-1775 highlighted in red. These genes represent the 12 genes that form the highest gene-drug associations with AZD-1775.

Overall, I identified a number of candidate drugs and potential corresponding biomarkers in the discovery phase. Next, I shifted to validation of both our ability to generate accurate predictions with AZD-1775 and AZD-1775's efficacy in TNBC preclinical models.

Validation Phase: Measured Cell Line Response to AZD-1775 in an Independent in vitro Dataset Validate Our Predictions

Even though the cross-validation within the CTRP dataset showed a good reproducibility between imputed and actual measured AZD-1775 sensitivity, I was interested in further exploring prediction accuracy with another independent dataset. Therefore, I imputed drug response scores in GDSC and compared GDSC measured AZD-1775 IC50 to our imputed results. I find a significant positive Spearman-rank correlation between our imputed sensitivity scores and GDSC's available IC50 data ($n = 749$, $r_s = 0.63$, $p < 2.2 \times 10^{-16}$). This correlation between imputed and measured sensitivity remains significant when I only analyze cell lines unique to GDSC (Figure 3.6A, $n = 297$, $r_s = 0.62$, $p < 2.2 \times 10^{-16}$). For comparison, the Spearman-rank correlation between AZD-1775 sensitivity in the cell lines that were measured in both datasets was 0.50 (overlapping cell lines: $n = 452$, $r_s = 0.50$, $p < 2.2 \times 10^{-16}$, Figure 3.6B). That the correlation of the imputed and measured results in GDSC was comparable to (if not greater than) the correlation between measured GDSC and CTRP results indicates that our imputed values capture the majority of AZD-1775's in vitro sensitivity variation. A breast cancer cell line only analysis reveals a similar correlation between imputed AZD-1775 response and GDSC measured sensitivity ($r_s = 0.64$); as did an analysis using only breast cancer cell lines unique to GDSC ($r_s = 0.61$).

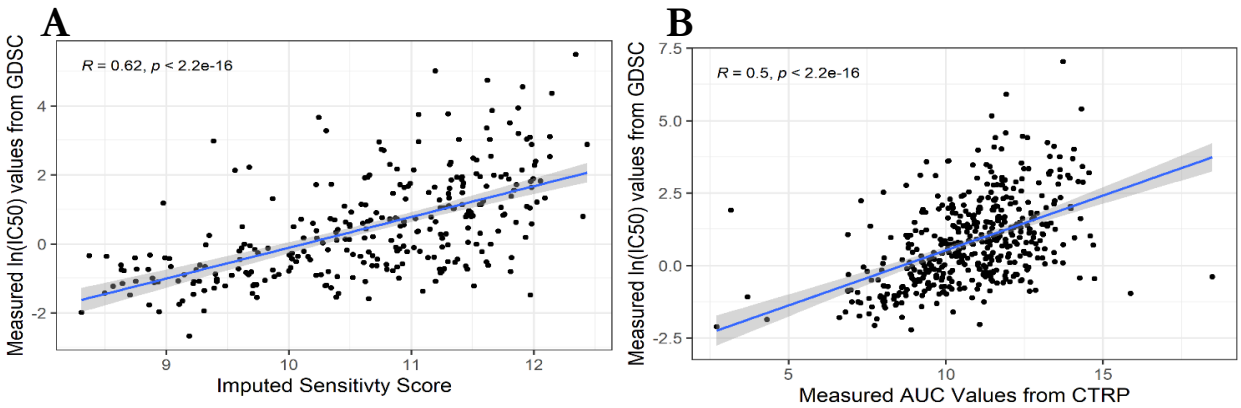


Figure 3.6 Correlation of Wee1 inhibitor predicted and actual cell line response to AZD-1775 in an independent in vitro dataset.
(**A**) For each cell line in GDSC not present in CTRP (i.e., cell lines unique to GDSC), imputed drug response to AZD-1775 using the CTRP imputation model (x-axis) is graphed against the reported GDSC IC50 sensitivities (y-axis). Spearman-rank correlation (R) and p-value as reported. (**B**) For each cell line in common between GDSC and CTRP, the measured drug response values are graphed against each other: IC50 values from GDSC (y-axis) against the measured AUC values from CTRP. Spearman-rank correlation and p-value as indicated. (B) serves as a point of comparison for how consistent two completely independent measured datasets are compared to the consistency of our imputed values built in one dataset to the measured values of the other (A). Abbr: GDSC, Genomics of Drug Sensitivity in Cancer; IC50, Half maximal inhibitory concentration; AUC, Area Under the dose response Curve.

Validation Phase: In Vitro and in Vivo Assessment of Cellular Sensitivity to AZD-1775 in Combination with Standard-of-Care Paclitaxel

My overall goal is to identify compounds that will be effective in the treatment of breast cancer patients. While the differences in imputed response indicate that a druggable vulnerability exists between the two populations, this data does not prove that the drug will be potent/effective at the appropriate dose levels. Thus, it is of paramount importance to follow-up the lead-compound identification with tests to verify a drug's potency in the population of interest. This section addresses this by looking at AZD-1775's inhibitory ability in in vitro and in vivo models of TNBC.

Single Agent use of AZD-1775 is Able to Inhibit Growth of TNBC Cell Lines

Thus far, I have focused mainly on CTRP data, which, as a high-throughput screen, is not optimized for particular compounds or cell lines. For example, AZD-1775 was screened at a

concentration range between 0–33 μM while the reported clinical C_{max} is only 1.4 to 2.6 μM depending on the dosing schedule (Do *et al.*, 2015). I therefore performed cell viability assays (Cell Titer Glo®, Promega, Madison, USA) tailored to three TNBC cell lines at a more pharmacologically meaningful dose range of 0 to 3.2 μM. I found EC50 to be 392.8 nM, 531.2 nM and 602.5 nM in the BT549, HS578T and MDA-MB-231 cell lines respectively after 72 hours of AZD-1775 treatment (Figure 3.7A–C). These results are in line with CTRP's reported EC50 values, where 18/19 TNBC cell lines were found sensitive to AZD-1775 at an EC50 1.4 μM or less (Figure 3.7D). These findings support that AZD-1775 is efficacious in inhibiting the proliferation of TNBC cell lines at pharmacologically achievable concentrations.

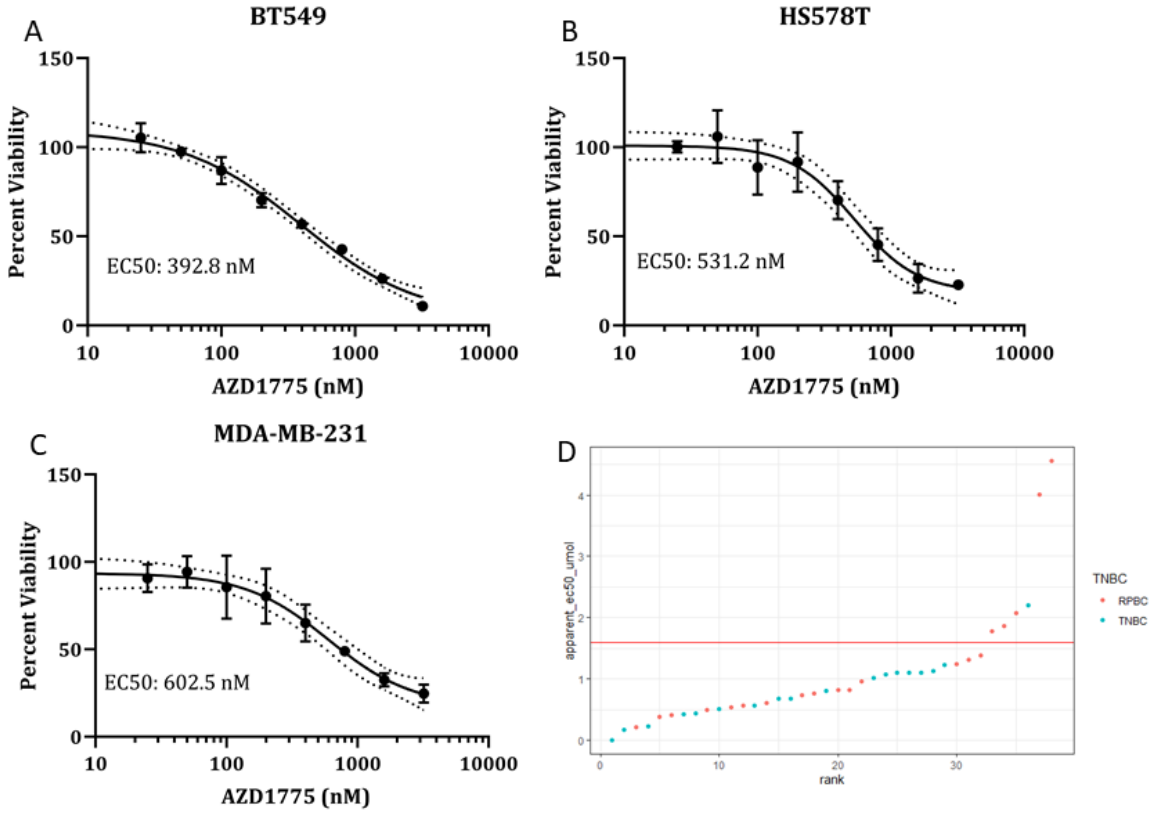


Figure 3.7. In Vitro Efficacy of AZD-1775.
(**A-C**) For the cell line indicated, an 8-point dose response curve was fit to 72-hour post-treatment viability measurements with Cell Titer Glo®. Percent viability was calculated normalizing to the vehicle only (DMSO) control wells. The 4-parameter logistic fit (full line) with 95% confidence intervals (dashed lines) are shown. (**D**) The reported EC50 values (y-axis) for the all the breast cancer cell lines in the CTRP dataset are plotted in order of sensitivity. Color is added to show the TNBC status of the cell lines.

AZD-1775 Alone and in Combination with Paclitaxel Inhibits MDA-MB-231 Xenograft Growth

To further validate AZD-1775 for use in TNBC, I performed tumor growth experiments in a mouse xenograft model. Using orthotopic MDA-MB-231 xenografts, I tested single agent AZD-1775 compared to vehicle control and/or paclitaxel, a standard of care agent in treating TNBC, as a positive control. Regarding tumor outgrowth, single-agent paclitaxel and AZD-1775 had similar efficacy at most time points, but neither showed consistent statistically different results when compared to control after adjusting for multiple comparisons (Figure 3.8A, AZD-1775 alone only reached significance at day 10, adj. $p = 0.0491$). However, time-to-endpoint analysis showed that AZD-1775 significantly improved time to endpoint while paclitaxel did not (Figure 3.8B, AZD-1775: $p = 0.021$; paclitaxel: $p = 0.18$ by Log-rank test). Additionally, I tested AZD-1775 in combination with paclitaxel; the rationale being that novel drug candidates are often first evaluated in clinical trials in combination with a standard-of-care agent. Combination treatment with AZD-1775 and paclitaxel showed statistically significant impairment of tumor outgrowth by day 10, which remained significant over the treatment period (Figure 3.8A, multiplicity adjusted p-value <0.05). Mice receiving a combination of AZD-1775 and paclitaxel did significantly better in time-to-endpoint than both the control ($p = 0.0003$) and paclitaxel alone ($p = 0.017$). These results support the use of AZD-1775 as a monotherapy, but they more strongly support the future testing of the combination of paclitaxel and AZD-1775 in treating TNBC.

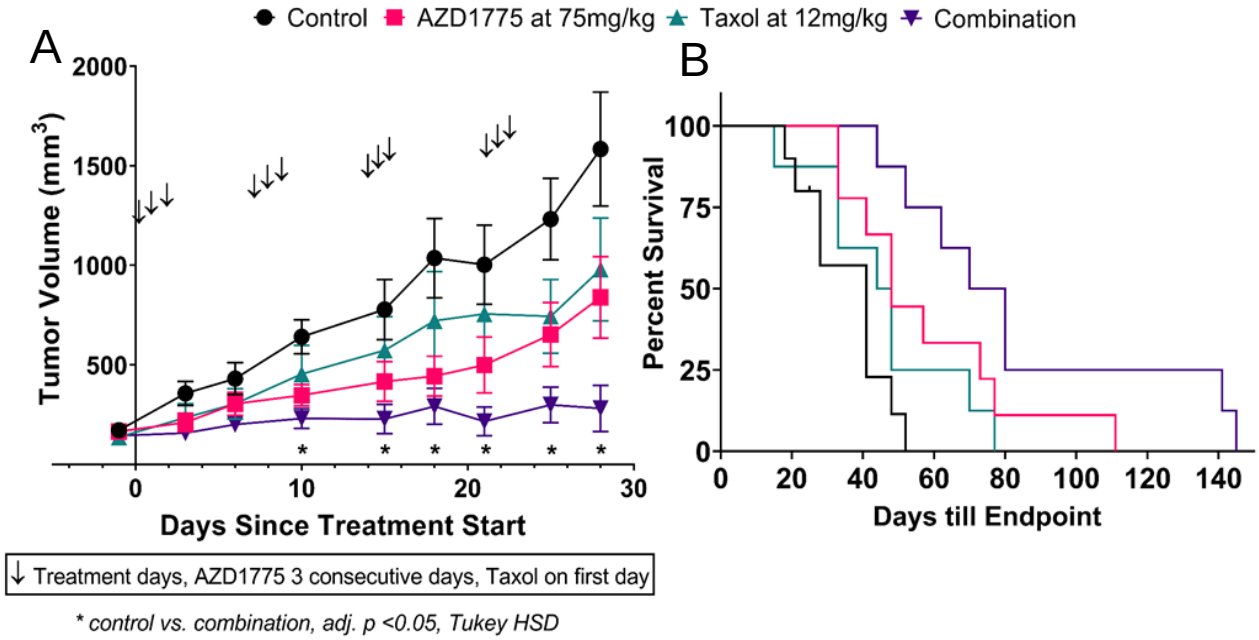


Figure 3.8. AZD1775 and paclitaxel combination inhibits TNBC tumor growth in vivo. (**A**) Tumor growth assay of MDA-MB-231 xenografts treated with AZD1775 and paclitaxel. Mice were orthotopically injected with MDA-MB-231 cells and allowed to grow to an average tumor size of 150 mm^3 before being randomized into 4 treatment groups: vehicle control, AZD-1775 given PO at 75 mg/kg on consecutive days 1–3 of each week for 4 weeks, Paclitaxel bolus given IP every week on day 1 of each week for 4 weeks, and the combination of the paclitaxel and AZD-1775 treatments. Tumor size was measured by caliper every 3–4 days after treatment start and tumor volumes and standard errors are plotted by treatment group over the first four treatment weeks. * indicate significant values between combination and control by Tukey's honestly significant difference (HSD) for each timepoint. The only other value that was significantly different during this time period by HSD was the AZD-1775 compared to control treatment at day 10 (not indicated). (**B**) Days till endpoint analysis of MDA-MB-231 xenografts. For the same mouse experiment shown in (A), days until tumor growth endpoint (2000 mm3) were recorded and plotted as a Kaplan–Meier survival curve. Log-rank (Mantel–Cox) test performed for all 4 curves reached a significance of $p = 0.0014$. Individual analysis was performed between each curve, the log-rank test was only significant ($p < 0.05$) when comparing Control vs combination, $p = 0.0003$, Taxol vs combination, $p = 0.0086$, and Control vs. AZD1775, $p = 0.0207$.

3.4 DISCUSSION

While patient datasets such as The Cancer Genome Atlas (TCGA) often contain a plethora of "-omics" data, the corresponding drug response information are limited and not suited for novel drug discovery. By integrating in vitro high throughput drug screening data and patient tumor molecular information, I created a virtual drug screening pipeline that enables drug discovery with simultaneous biomarker identification for a patient population. I demonstrated the pipeline from lead identification, to biomarker discovery, to in vitro and in vivo validation of the compound AZD-1775 for TNBC.

TNBC is a very aggressive disease with a high recurrence rate and limited therapeutic options (Foulkes, Smith and Reis-Filho, 2010). Our main goal was to take advantage of existing large high-throughput in vitro drug screening datasets and apply novel computational methods to systematically identify effective drugs for the treatment of TNBC. To this end, I imputed drug responses by making 427 individual drug-transcriptome models and identified candidate drugs that are likely to be more effective in TNBC tumors. Among them, I focused on AZD-1775, a Wee1 kinase inhibitor, testing the model's accuracy in an independent screening dataset (GDSC) and the compound's efficacy in inhibiting tumor growth using an in vivo xenograft model. I also demonstrated the pipeline of employing computational methods to identify biomarkers for compounds of interest that could help further tailor a targeted therapy to the appropriate patient population.

Overall, this method has several unique advantages in comparison with other approaches for drug discovery. First, I am able to investigate a large number of compounds in a relatively short amount of time by taking advantage of existing and publicly available datasets (available databases reviewed in Chapter 2). This analysis was able to quickly query 427 compounds

targeting over 400 known cancer targets in a clinical population of interest. Additionally, utilizing machine learning models trained on large panels of *in vitro* cell line data allows me to make unbiased models with no assumptions about the compounds' modes of actions, which can often be partially or completely unknown. For cancer indications with limited number of cell lines or useful preclinical models, this method could be particularly useful as the patterns of drug response are identified using large panels of pan-cancer cell lines and then applied to the disease type of interest. Furthermore, the choice of patient population is flexible and controlled by the investigator such that this approach could be used to identify compounds targeted towards patient groups defined clinically or, for example, patients defined by a genetic alteration of interest to the investigator. Finally, imputing in patient datasets directly allows us to translate the in vitro patterns of drug response into the most clinically relevant population and leverage other types of information contained in these datasets. This is something that up until this point has not been done for drug discovery, and, as shown in this work, is quite capable of identifying relevant and useful compounds. Use of this methodology in other cancer contexts is discussed further in the future directions in Chapter 5.

This approach has potential limitations and considerations. This and other drug modeling methods are unable to make absolute potency or efficacy predictions. As such, I can only compare relative imputed sensitivity metrics within a drug and not the metrics between drugs. Issues related to differences in relevant dosing ranges, toxicities, and other pharmacokinetic parameters make comparing drug sensitivity metrics among compounds impractical. In addition, the transformation of sensitivity metrics for modeling and independence of the models add additional barriers that prevent the direct comparison of model outputs between drugs. However, I would argue that even if predicting absolute potency of a

compound were possible, it would necessarily enrich for toxic chemotherapies since these are definitionally the most potent compounds.

My approach of comparing patient subtypes is therefore more useful for identifying agents that target particular genomic vulnerabilities in the patient population of interest. This does preclude finding compounds that may target vulnerabilities present in both populations at similar frequencies. Under the assumption that a targeted therapy will work for a certain subtype and not others, this limitation is not likely to hinder identifying the most interesting compounds. Indeed, finding targeted therapies and therapies that work specifically for a particular subtype is of critical interest because there is no panacea or magic bullet in cancer. TNBC is a good example use-case then, as TNBC is generally quite distinct from other breast cancer subtypes and indeed in pan-cancer analyses often cluster away from other breast cancers (Prat *et al.*, 2013; Hoadley *et al.*, 2014). The choice of patient populations to compare should be carefully considered when using this approach.

Of course, both because the method is computationally driven and because predictions cannot check drug potency or toxicities, validation is necessary to check the effectiveness of the compound. As such this method is more complimentary to the traditional drug development pipeline than a substitute for it, allowing for the speedy identification of lead compounds for further study and validation.

The results of the differential imputed drug sensitivity between TCGA's TNBC and RPBC cohorts indicate that there is a skew towards identifying more compounds predicted to be effective in TNBC rather than RPBC. There are likely several contributing factors to explain this skew. First, regarding the translational models, it has been shown that well performing

computational models are not necessarily specific to a drug of interest (Schätzle, Esfahani and Schuppert, 2020). That is, general mechanisms of drug response, such as factors associated with multi-drug resistance, are likely encoded in many of the independent models and as such might skew the imputed distribution of all the drugs in a similar direction. Additionally, many of the agents tested in CTRP target DNA-damage or cell cycle, which is a known vulnerability of TNBC, whereas very few compounds target hormone pathways present in RPBC. Finally, RPBC captures a mixture of hormone dependent and growth factor (HER2) driven diseases, thus it is not surprising that not many drugs are specifically enriched in this mixed population. However, for our list of candidate compounds, I only focused on the top 10% most significant hits to account for this skew towards TNBC.

Of the 17 compounds I focused on, I observed multiple cell cycle inhibitors and DNA-damaging agents. Of note were the CHK, PLK, and CDK inhibitors that have been explored in depth in the TNBC literature. CHK inhibitors have already shown promise in TNBC preclinical models to induce cell death by promoting DNA damage and subsequent apoptosis (Albiges et al., 2014; Bryant, Rawlinson and Massey, 2014), with particular efficacy in both Rb (Witkiewicz et al., 2018) and p53 deficient (Ma et al., 2012) settings. While one of our top results here, the use of CHK inhibitors have been limited due to poor pharmacokinetic properties, with several earlier formulations being discontinued due to toxicities (Thompson and Eastman, 2013). PLK inhibitors have similarly been explored as an overexpressed factor in TNBC whose inhibition can induce G2/M arrest in TNBC in vitro models (Ueda et al., 2019). Recapitulating these targets in TNBC supports the ability of our methodology to systematically identify effective agents for a cancer population of interest.

There has been increasing interest in the use of AZD-1775 in various cancer settings, with several early phase clinical trials showing a favorable toxicity profile in a number of different cancer settings (Do et al., 2015; Leijen, van Geel, Pavlick, et al., 2016; Leijen, van Geel, Sonke, et al., 2016; Méndez et al., 2018; Sanai et al., 2018; Cole et al., 2020). Wee1 is a protein kinase that acts as an inhibitor of the cell cycle through phosphorylation of the CDK1/cyclin B complex resulting in G2 cell cycle arrest (Parker and Piwnica-Worms, 1992). Thus, inhibition of Wee1 by AZD-1775 blocks DNA damage repair through loss of the cell cycle checkpoint and promotes early entry into mitosis. This is often fatal for cells, especially in p53 mutated cells that mainly rely on the G2/M checkpoint for DNA repair (Geenen and Schellens, 2017), which is often the case in TNBC (Hwang, Park and Kwon, 2019). Wee1 inhibition has been shown to sensitize p53-deficient tumor cells to DNA-damaging agents in various cancer contexts (Hirai et al., 2009; Mizuarai et al., 2009; Pappano et al., 2014; Clausse et al., 2016; Cuneo et al., 2016; Webster et al., 2017; Yin et al., 2018; Diab et al., 2019) and to prolong mitosis in breast cancer cells (Lewis et al., 2017). Since I began investigating AZD-1775 for TNBC, several studies have also come out in TNBC: one showing that cyclin E overexpression sensitizes TNBC to AZD-1775 (Chen et al., 2018) and one reporting synergy between AZD-1775 and ATR inhibition in TNBC cells (Jin et al., 2018). Our work here serves to complement these studies as the first computational approach to use existing pharmacologic data to identify AZD-1775 as a compound of interest for TNBCs as well as the first to explore the combination of AZD-1775 with standard-of-care paclitaxel in TNBC treatment. Of the over 50 clinical trials listed on clinicaltrials.gov with AZD-1775, only one is in (metastatic) TNBC (NCT03012477) and four focus on tumors harboring p53 mutations (NCT02272790, NCT01357161, NCT01164995, NCT02448329). These clinical trials thus far have not provided genomic mutational information

to assess the potential of the associated biomarkers I identified. I believe our findings, along with others mentioned here, support further investigation of AZD-1775 in TNBC as well as the investigation of the potential biomarkers identified to tailor therapy to the appropriate patients.

In summary, I provide a novel approach for the identification and prioritization of candidate compounds for a particular patient group of interest and showcase the use of this method in TNBC. In doing so, I identified and validated AZD-1775 for use in TNBC both in vitro and in vivo. However, this was not the only top target that showed predicted efficacy in TNBC. Among the compounds in this list, the use of XPO1 inhibitors (the second most significant compound associated with TNBC from Table 4.1). Chapter 4 extends the drug discovery work started here by examining the efficacy and mechanism of response of XPO1 inhibition in breast cancers.

CHAPTER 4: Investigating inhibition of nuclear export in breast cancer cells

4.1 INTRODUCTION

Chapter 3 showcased the potential utility of imputed drug response-based drug discovery. Table 3.1 generated a number of potential hits, many of them with other preclinical evidence for their use in TNBC. In particular, the second most significant compound in the list, leptomycin B stands out. Unlike most of the other compounds, Leptomycin B is not directly related to the cell cycle. Leptomycin B is an XPO1 inhibitor that represents a fairly unique mechanism of action of inhibiting nuclear-to-cytoplasmic protein translocation. However, Leptomycin B was very toxic to patients in a clinical trial setting. Second generation XPO1 inhibitors (SINEs or selective inhibitors of nuclear export) have been recently developed and have been approved for multiple myeloma patients in 2019 (*Drug Approval Package: XPOVIO*, 2019), demonstrating these new generation of XPO1 inhibitors are a safe and efficacious treatment. With promising safety and efficacy of SINEs and the indications from my drug response imputation that XPO1 may be a preferentially effective treatment for TNBC results, investigating XPO1 as a therapy for TNBC is a novel and promising research avenue. Additionally, investigation of XPO1 could further showcase the potential for the imputed drug response to be used for drug discovery.

To begin I overview the role of XPO1 and exportins play in the cell and in cancer, and then discuss the previous studies on XPO1 in breast cancer in particular. While some other studies have already identified XPO1 inhibition as a potentially effective treatment in breast cancer, the mechanism by which XPO1 inhibition is leading to decreases in cell viability hadn't been sufficiently demonstrated. Thus, this section of my dissertation is focused on the initial exploration of XPO1 inhibition-mediated cell death through RNA-Seq as well as leveraging the

imputed drug sensitivity results to find mechanistic associations between imputed response and patient features.

Exportins in the Cell

The nucleus of a cell is a controlled subcellular location. Unlike other subcellular locations like the endoplasmic reticulum or the mitochondria which are much more closed off to the cytoplasm, the nucleus contains pores that allow molecules and proteins of less than about 30-60 kDa (depending on shape and charge) to freely diffuse in and out. Proteins of larger size can still pass through the nuclear pore complex (NPC); however, the rate of diffusion drastically diminishes with increasing size. For the efficient movement of large proteins and RNA structures, the karyopherin and NTF-like proteins facilitate the transportation of across the NPC. This family is made up of 3 subcategories: importins (facilitate nuclear import), exportins (facilitate nuclear export), and transportins (facilitate both nuclear import and export).

For proteins, the movement across the NPC is generally controlled by an amino acid sequence that tags a protein for either nuclear import (an nuclear localization signal, NLS) or export (a nuclear export signal, NES), with energy provided by a Ran-GTPase. For the case of import, an importin binds to a protein with an NLS, the complex interacts with and goes through the NPC. In the nucleus the complex binds to Ran-GTP which destabilizes the complex allowing for release of the cargo protein. Ran-GTP is hydrolyzed which releases the importin, both proteins are then recycled to start the process anew. Export is the exact opposite, with exportin binding a Ran-GTPase first which induces a conformational shift in the exportin and allows for the binding of the NES containing cargo protein. Transport through the NPC follows, and hydrolyzation of Ran-GTP in the cytoplasm releases the complex.

XPO1 in Cancer

The NES is a peptide sequence of the form LXXLXXXL where L is a hydrophobic residue (often a leucine) and X can be any other amino acid. There are over 300 proteins with an NES sequence that have been confirmed functional in the literature with many more putative examples (Xu, Grishin and Chook, 2012). Of the 368 proteins on NESdb (ibid, http://prodata.swmed.edu/LRNes/IndexFiles/namesGood.php, date accessed 04/01/2021), 89 of these proteins are either match the "oncogene" or "tumor suppressor genes" flags on Uniprot (The UniProt Consortium, 2021). After further manual curation for the proteins with the most commonly altered in cancer, I created Table 4.1. In this table, we can see many different cancer relevant genes and pathways are potentially impacted by XPO1 inhibition. To highlight a few areas, many proteins involved in the G1/S transition are present in this table: E2F1, E2F7, p21, and cyclin D1. Additionally, DNA repair relevant proteins BRCA1, BRCA2, p53, PARP10 are present, with p53 exclusion from the nucleus being a well-established phenomenon in cancers (Moll, Riou and Levine, 1992). NK-$\varkappa$B pathway proteins and WNT signaling proteins are also well represented in this table.

Gene Symbol	Full name	NESDB_ID	oncogene	TSGs
AhR	Aryl hydrocarbon receptor	55	onco	tsg
Akt1	RAC-alpha serine/threonine-protein kinase	261	onco	tsg
APC Protein	Adenomatous polyposis coli protein	47	onco	tsg
ATF-2	Cyclic AMP-dependent transcription factor ATF-2	208	NA	tsg
Bach1	Transcription regulator protein BACH1	171	onco	NA
Beclin 1	Beclin-1	132	onco	tsg
BRCA1	Breast cancer type 1 susceptibility protein	10	onco	tsg
BRCA2	Breast cancer type 2 susceptibility protein	169	onco	tsg
CASC3	Cancer susceptibility candidate gene 3 protein	379	onco	NA
cIAP1	Baculoviral IAP repeat-containing protein 2	103	NA	tsg
Cyclin B1	G2/mitotic-specific cyclin-B1	9	onco	NA
Cyclin D1	G1/S-specific cyclin-D1	54	onco	tsg
E2F1	Transcription factor E2F1	59	onco	tsg
E2F7	Transcription factor E2F7	297	onco	NA
FAK	Focal adhesion kinase 1	83	onco	NA
FOXO3	Forkhead box protein O3	64	onco	NA
HDAC1	Histone deacetylase 1	252	onco	tsg
hTERT	Telomerase reverse transcriptase	53	onco	NA
KLF5	Krueppel-like factor 5	115	onco	NA
MAPKK1/MEK1	MAP kinase kinase 1	11	onco	NA
MK5	MAP kinase-activated protein kinase 5	165	onco	tsg
mTOR	Mammalian target of rapamycin	384	NA	tsg
NANOG	Homeobox protein NANOG	224	onco	NA
p100	Nuclear factor NF-kappa-B p100 subunit	237	onco	tsg
p120ctn	Catenin delta-1	19	onco	NA
p21Cip1	Cyclin-dependent kinase inhibitor 1	210	onco	tsg
P53	Cellular tumor antigen p53	6	onco	tsg
p73	Tumor protein p73	39	onco	tsg
PARP-10	Poly [ADP-ribose] polymerase 10	248	onco	NA
Pdcd4	Programmed cell death protein 4	349	onco	tsg
RelA/p65	Transcription factor p65	25	onco	tsg
Smad4	Mothers against decapentaplegic homolog 4	18	NA	tsg
Snail	Zinc finger protein SNAI1	211	onco	NA
STAT3	Signal transducer and activator of transcription 3	50	onco	tsg
YAP1	Transcriptional coactivator YAP1	320	onco	tsg

Table 4.1 Cancer Proteins Containing verified NES motifs
A list of proteins containing NES motifs was obtained from NESdb. This list was subsequently filtered to proteins that also appeared in Uniprot using the search terms "tumor suppressor genes" or "oncogene" (as indicated in the final columns). Of the original 378 proteins, 88 cancer relevant proteins remained. These were further manually curated to the list of 35 genes presented here to showcase the breadth of potential XPO1 targets that could be affected by XPO1 inhibition.

As can be seen in Table 4.1, many of the proteins XPO1 interacts are regulatory proteins that moderate cellular growth and apoptosis. Unsurprisingly then, XPO1 overexpression is common in various cancer types, including AML, lymphoma, myeloma, and ovarian cancers (Gandhi et al., 2018). XPO1 inhibition has been investigated as a potential therapeutic target for many of these cancer types, with the XPO1 inhibitor Selinexor being recently approved to treat refractory or relapsed multiple myeloma ('XPO1 Inhibitor Approved for Multiple Myeloma', 2019). While XPO1 inhibition has shown promise in inhibiting cancer growth in some of these settings, the mechanism by which XPO1 inhibition induces cell death has been difficult to define. Since there are many proteins XPO1 interacts with, there are many potential downstream signaling pathways that might lead to XPO1-inhibition-mediated cell death. In general, the hypothesized oncogenic role XPO1 plays in cancer is that XPO1 interacts with tumor suppressor genes and expels them outside the nucleus. For tumor suppressor genes that interact with DNA (either through DNA damage signaling or transcriptional activation or inhibition), this renders them functionless allowing the cancer cell to continue to proliferate. Changes in the localization of p53 serves as a good example for this since outside the nucleus, p53 can no longer act as a transcription factor. Implicated downstream effects of XPO1 inhibitors have included: the increased nuclear retention of p53 in AML (Ranganathan et al., 2012), NF-κB in CLL (Lapalombella et al., 2012), and FOXO proteins in ovarian cancer (Corno et al., 2018), pancreatic cancer (Azmi, Aboukameel, et al., 2013), and NHL (Azmi, Al-Katib, et al., 2013).

XPO1 Inhibition in Breast Cancer

In breast cancer, there have been several studies on the efficacy of XPO1 inhibitors in preventing cell growth. A 2014 study showed the inhibitor was effective in inhibiting cell growth in breast cancer cell lines as well as a triple-negative breast cancer (TNBC) xenograft (Cheng et

al., 2014) and a 2017 study showed Selinexor was able to inhibit the growth of 17/26 breast cancer cell lines (this paper also tested combination treatments with Selinexor) (Arango et al., 2017). Of particular interest from this paper is that XPO1 inhibition was able to inhibit the growth of all 14 TNBC cell lines and 4/5 TNBC PDX models. TNBC is a very heterogeneous disease, so it is particularly surprising that this inhibitor is able to show efficacy in such a wide panel of TNBC samples. The observation that XPO1 is able to inhibit TNBC tumor growth have led to testing Selinexor in a small Phase II Clinical Trial of Advanced (Metastatic) TNBC tumors (Shafique et al., 2019). This clinical study showed Selinexor was generally well tolerated in patients, but no objective responses were observed in the first 10 patients so the study was terminated early. 30% of the patients (3/10) did reach a Stable Disease Recist Criteria. The patients enrolled in this study were advanced stage patients and the study was clearly underpowered to be definitive. Still, understanding the mechanism by which XPO1 inhibition is working in these TNBC models could help determine biomarkers or potential combinations that might synergize with XPO1 inhibition and improve patient response. Furthermore, understanding the mechanism by which XPO1 inhibition leads to cell death may shed light on biological vulnerabilities in TNBC that may increase our understanding of this disease.

While the efficacy of XPO1 inhibition in TNBC has been investigated, little is known about the mechanism of XPO1 in breast cancer. In the 2014 paper mentioned above, XPO1 inhibition was linked to STAT3 inactivation which in turn blocked the expression of the survivin oncogene (Cheng et al., 2014). While this paper showed XPO1 inhibition inhibited TNBC growth (figures 1-2 of Cheng et al), they only showed that XPO1 inhibition blocks survivin export in one ER+ cell line and one HER2+ cell line (figure 3) and that XPO1-mediated apoptosis was dependent on survivin in only the HER2+ cell line (figure 4). Furthermore, the

study only showed that XPO1 inhibition activated caspase-dependent cleavage of survivin happened in the HER2+ cell line (with less evidence of activation in MCF7) (figure 5) and that inhibition of XPO1 represses STAT3 transactivation in HEK293 cells (figure 6). Besides the lack of consistent model system used in this paper, the paper chose survivin as the important mediator because the authors had previously shown survivin interacts directly with XPO1. Finally, the 2017 Arango et al paper largely rejected the survivin-mediated explanation of XPO1 response since XPO1 inhibition was shown to decrease survivin expression in both sensitive and resistant cell lines and called on researchers to identify new mechanisms and pharmacodynamic markers of response. Additionally, the general mechanisms by which TNBC attains more general sensitivity to XPO1 inhibitors than other breast cancer subtypes are still undetermined. An unbiased and wholistic approach is needed to examine the mechanism that leads to XPO1 inhibition mediated cell death in breast cancer.

Summary and Overview

XPO1 is an important regulator in the cell that controls the localization of hundreds of proteins, many of them important for cancer. XPO1 inhibition is a promising clinical therapy with success in a number of different cancer types. Of particular interest is the noted ability for XPO1 inhibition to have a clear sensitive/resistant response profile and to be effective in a wide variety of molecular backgrounds. While XPO1 can hit multiple targets, it still remains undetermined which ones are the most relevant for cancer and whether its ability to target a heterogeneous population of cells is through the simultaneous targeting of multiple pathways. Understanding the pathway by which breast cancer cells are responding to XPO1 inhibition may help explain the heterogeneous response and also lead to important biomarkers of response.

My goal is to evaluate and understand the means by which XPO1 inhibitors drive an anti-tumor effect in breast cancer. To begin, I analyzed cell line screening data for leptomycin B (an XPO1 inhibitor) from the Broad Institute's Cell Therapeutics Response Portal (CTRP) and perform my own set of experiments to validate the efficacy *in vitro*. Next, I further evaluated the imputed drug response and patient data to obtain gene sets and mechanisms that are associated with imputed drug response. Finally, I performed and analyzed an RNA-Seq experiment in two cancer cell lines to determine what genes are being differentially expressed after treatment with the SINE KPT-330 as a way to begin the mechanistic exploration of XPO1 inhiibiton-induced cell death.

4.2 METHODS

Data Availability

The Leptomycin B data was obtained from the Broad Institute's Cell Therapeutics Response Portal v2 (CTRP) (Seashore-Ludlow *et al.*, 2015), which is hosted by the Cancer Target Discovery and Development Network established by the National Cancer Institute's Office of Cancer Genomics (*https://ocg.cancer.gov/programs/ctd2/data-portal*, no date). The corresponding mutation data for TP53, PI3KCA, and PTEN were obtained directly from the Broad Institute's Cancer Cell Line Encyclopedia (CCLE) data portal (*Broad Institute Cancer Cell Line Encyclopedia (CCLE)*, no date).

Data presented from the cMAP (Subramanian *et al.*, 2017), was obtained direction from the website (https://clue.io/, accessed 04/01/2021). From this website, searching for XPO1 brings up the XPO1 inhibitor referred to as "BRD-K61829047" and selecting "/CONN" allows for viewing the top connections/internal connectivities. From here, the cell lines HCC515, HT29, MCF7 and PC3 were selected as these shared the most consistent connections.

Viability and Apoptosis Experiments

A panel of breast cancer cell lines used in Figure 4.2 were grown in their ATCC suggested media. All media was supplemented with L-glutamine and 10% FBS (ThermoFisher Scientific, Gibco, Waltham, USA). For viability assay, cells were seeded at 5000 cells/well in 96-well plates. After 24 hours, the media was removed and replaced with media containing KPT-330 (Selleck Chem, Houston, USA) ranging between 25 and 3200 nM or with DMSO as vehicle for control wells. Growth was monitored every 4 hours to ensure control wells reached but did not exceed 95% confluence. After approximately 72 hours of treatment for each cell line, Cell Titer Glo® (Promega, Madison, USA) viability assay was performed as suggested by

manufacturer. Luminescence values were obtained from VICTOR Multilabel plate reader (PerkinElmer, Waltham, USA) and normalized to control well before plotting. Graphing and EC50 determinations were done using Prism 8 software (GraphPad, San Diego, USA). EC50 values were then plotted using ggplot functions in R (Wickham, 2009).

Apoptosis Experiments

MDA-MB-231 and MCF7 cells were cultured in DMEM (GE Healthcare Life Sciences, Hyclone, Logan, USA) supplemented with L-glutamine and 10% FBS (ThermoFisher Scientific, Gibco, Waltham, USA) and seeded onto 96 well plates at 5000 and 7000 cells/well respectively. After an initial 24 hours, media was replaced with media containing drug, Incucyte® Caspase-3/7 Green Dye for Apoptosis(Essenbio Sciences, Satorius, Ann Arbor, USA) and Incucyte® Cytotox Red Dye (Essenbio Sciences, Satorius, Ann Arbor, USA). Growth was monitored via Incucyte® S3 (Essenbio Sciences, Satorius, Ann Arbor, USA) with images taken every 4 hours. Images were analyzed with Incucyte software for red/green counts per image plotted over time.

RT-qPCR Experiments

Survivin gene expression was determined using RT-qPCR. MDA-MB-231 and MCF7 cells were cultured as described previously and treated with KPT-330 for 2, 8, or 24 hours. Cells were then lysed and RNA was extracted with the RNeasy mini kit (Qiagen, Hilden, Germany). RNA was converted to cDNA using qScript (QuantaBio). Transcript levels were analyzed by real time PCR using Taqman reagents (Life technologies) in the StepOnePlus Real-Time PCR System (Applied Biosystems), with probes for Survivin and B2M were purchased from IDT

Imputation based analysis

Gene-set enrichment analysis (Subramanian *et al.*, 2005b) as performed using the software package GSEA v4.0.2 for Windows downloaded from gsea-msigdb.org. TCGA BRCA RNA-

Seq data was used as the expression dataset, MSigDB's hallmark gene sets (Liberzon *et al.*, 2015) were used for the gene sets database, and patient imputed sensitivity scores to leptomycin B were used as a continuous phenotype label. Default software parameters were used except Pearson correlations were used for ranking genes to reflect the use of a continuous phenotype label.

RNA-Seq Experiments

Cell lines were grown in 6cm dishes and allowed to establish for 48 hours. The media was then replaced with either DMSO or 200 nM of KPT-330 media. At 2, 8, and 24 hours cells were scraped, spun down, and snap frozen. RNA was isolated as previously described. Library preparation and sequencing was performed by the UChicago Genomics Facility. RNA QC (quality and quantity) was performed using an Agilent bio-analyzer. RNA-SEQ libraries were generated using Illumina mRNA stranded kits (using Illumina protocols) to obtain approximately 60 million pair-end reads per sample and the libraries were sequenced on an Illumina NovaSEQ 6000 using Illumina reagents and protocols. Adapters used dual unique molecular indices from IDT (10bp index).

The University of Minnesota's pipeline for bulk RNA-Seq processing was used to process the RNA-Seq data (Baller *et al.*, 2019). To summarize the pipeline briefly, raw reads were trimmed using Trimmomatic using the options to trim for: all_illumina_adapters, LEADING:3 TRAILING:3, SLIDINGWINDOW:4:15, MINLEN:18. Alignment was performed via HISAT2 using the latest release of GRCh38.100 as the reference genome. Finally, aligned reads were counted using featureCounts.

Differential gene expression analysis was performed using the *DESeq2* R package (Love, Huber and Anders, 2014). Genes were called differentially expressed if the FDR corrected p-

value was less than 0.1 and absolute log2 fold-change was above 1. Differentially expressed genes were then used gene enrichment analysis using KEGG pathways (Kanehisa *et al.*, 2017) via the R package *clusterProfiler* (Yu *et al.*, 2012). KEGG pathway visualizations were created by the *pathview* R package (Luo and Brouwer, 2013). GSEA enrichment analysis was performed using *clusterProfiler* and the *msigdbr* package (*MSigDB gene sets R package*, no date).

4.3 RESULTS

The results can be divided into three separate areas: evaluation of XPO1 inhibitor in cell lines, computational-driven analysis of potential mechanism of actions, and RNA-Seq based analysis of mechanisms of action in two distinct breast cancer cell lines. To begin, it is first important to explore the efficacy of XPO1 inhibition and patterns of drug response in breast cancer cell lines. This analysis will highlight some of the unique characteristics of XPO1 sensitivity: heterogeneous response and dichotomous response. With that context, the drug imputation results are then analyzed to determine whether the imputed mechanistic results are in line with known biology. Finally, I explore RNA-Seq data to gain experimental insight into the mechanism of XPO1 inhibitor-mediated cell death.

Efficacy of XPO1 Inhibitors in Breast Cancer

Leptomycin B was one of the top hits from my imputed differential analysis based on CTRP data (Chapter 3). Thus, I investigated the observed effects on cell line viability this compound had in the CTRP dataset. Leptomycin B drug response values from CTRP are shown in Figure 4.1A. As can be seen, most (30/36) breast cancer cell lines and all but one TNBC (16/17) cell lines respond to the leptomycin B at an EC50 value of less than 20 nM. However, the 5 most resistant cell lines (cell lines that don't respond below 20 nM) have an over 400-fold decrease in sensitivity (not shown on graph). Additionally, the cell lines responding to treatment are heterogeneous. When classifying the cancer cell lines into HER2+, HR+, or TNBC, it can be seen the response rate is 85%, 66.6%, and 94% respectively. This extends further when investigating the genetic mutations of the cells. As can be seen in Figure 4.1B, response to leptomycin B appears independent of three of the most common breast cancer mutations: p53, PTEN, and PI3KCA.

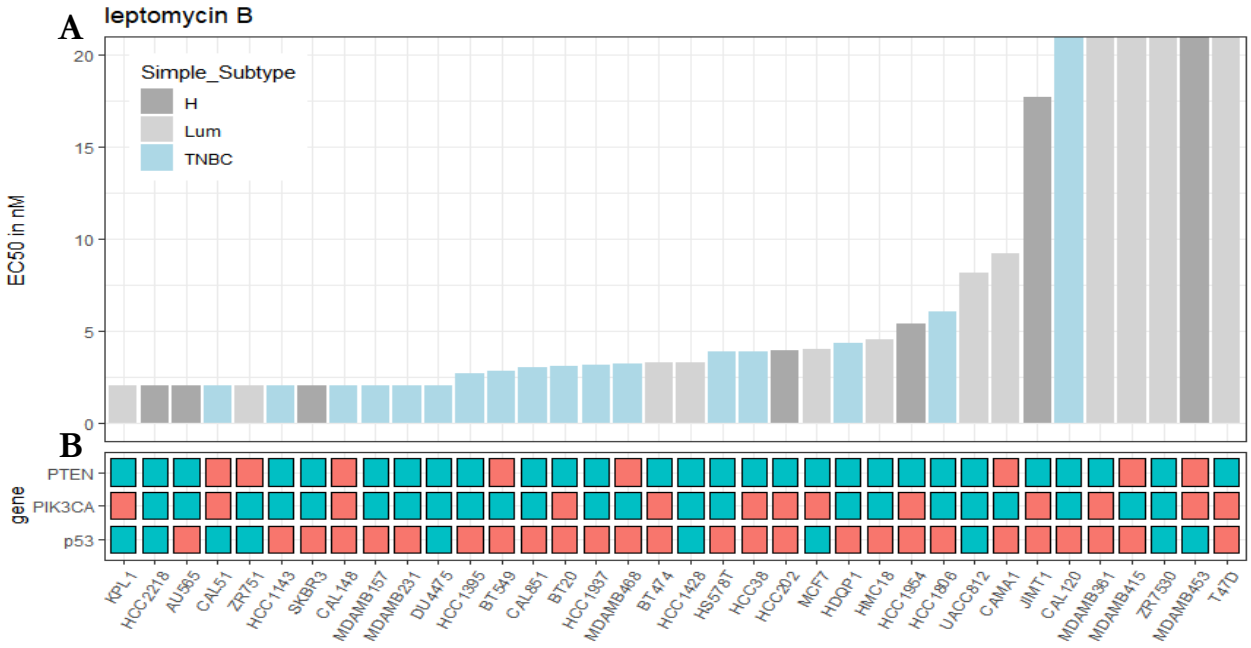


Figure 4.1. Cell Line Response to XPO1 Inhibitor Leptomycin B from CTRP
(**A**) Plot showing CTRP cell line EC50 values for each cell line in order of sensitivity. Cell line names are in line with the lower graph. Columns are colored by breast cancer subtype as determined from literature review. The 11 most sensitive cell lines had a reported EC50 less than the minimum tested concentration, so for simplicity these were given the minimum tested concentration (2.2 nM). The scale is linear and zoomed in to increase interpretability. The six right-most bars continue off the edge of the graph. CAL-120 has an EC50 of 403nM and the five remaining cell lines have an estimated EC50 of over 8,000 nM (over the max tested value). (**B**) Heatmap of gene mutations. Red indicates the cell line has a mutation in the specified gene, blue-green indicates the gene is WT in that cell line. Mutation information was obtained from CCLE.

This analysis of publicly available data is in line with the reported literature and demonstrate the heterogeneous and dichotomous response nature to XPO1 inhibition. The response is heterogeneous in that many cell lines respond even though these cell lines can represent very different biology (breast cancer subtype and driver mutations). I use the term dichotomous to represent the clear sensitive/resistant phenotype present in this data where the rule could be if a cell line is going to respond, they are likely to respond at less than 20 nM or won't respond until 400-fold higher concentrations (with only one exception, CAL-120, which responds at 400 nM).

Leptomycin B was tested in clinical trials but quickly abandoned because of toxicity issues and SINEs with better toxicity profiles were developed. To test whether SINEs have similar profiles, I tested KPT-330 (aka Selinexor or brand name: Xpovio) in a panel of breast cancer cell lines (Figure 4.2A). Similar heterogeneous and dichotomous response can be observed in this panel of cell lines where every TNBC cell line responded to therapy at or below 200 nM whereas T47D had an EC50 of close to 1400 nM. These results are in line with a similar literature report that tested KPT-330 in a panel of 26 breast cancer cell lines(Arango *et al.*, 2017), but are around 2 to 10-fold less effective in my tests. While Arango et al reported EC50 values for many of the same cell lines as between 11-50 nM, the overall results are consistent since the same cell lines respond in either case. It should be noted that Arango et. al shows some inconsistency compared to Figure 4.1. For the 9 ER+ cell lines tested in Arango et. al., only 1 cell line (MCF7) was shown to be sensitive to KPT-330. Of the three other ER+ cell lines that overlap between Arango et al and the CTRP leptomycin B data shown in Figure 4.1, BT474 is inconsistent with it being sensitive above but not in Arango et al. T47D and MDA-MB-361 are shown resistant in both. Thus, it is not clear if the claim in Arango et al that KPT-330 is not effective in ER+ models is entirely substantiated or if testing in additional ER+ models (such as the ones used in CTRP) would uncover more responders. However, another source of this variation could be due to the use of different XPO1 inhibitors (leptomycin B compared to KPT-330). Either way, MCF7 is clearly sensitive in these two studies as well as my own data (Figure 4.2. This is of interest since, if Arango et al is correct, this is one of the few ER+ breast cancer cell lines to respond to KPT-330 inhibition.

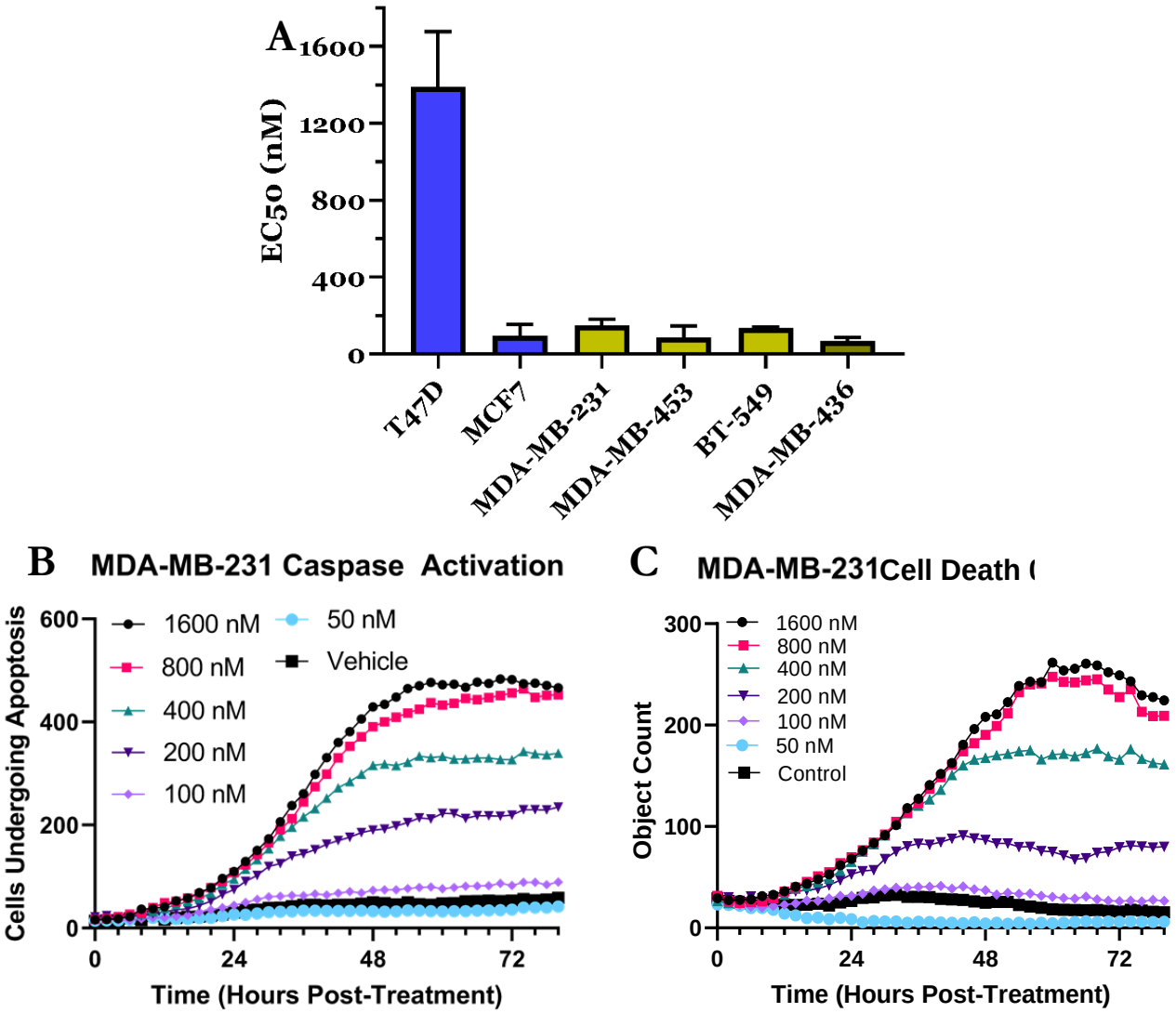


Figure 4.2 Tested Efficacy of KPT-330 in Breast Cancer Cell Lines and Time/Dose Dependencies. (**A**). Effects of KPT-330 on cell proliferation *in vitro*. 6 breast cancer cell lines (2 ER+, blue, and 4 TNBC, yellow) were treated with KPT-330 at 8 concentrations based on a 2 fold dilution series (range 25-3200 nM). Viability was measured using Cell Titer Glo 72 hours after treatment. 4 parameter log logistic dose curves were fit using Graphpad Prism and the EC50 value is graphed. (**B-C**) Time course analysis of MDA-MB-231 cell response to KPT-330. MDA-MB-231 was plated with 6 concentrations of KPT-330, a cell permeable and caspase activated dye (B) and red dye to monitor cell membrane integrity (C) and the cells were monitored every 2 hours via Incucyte. Flourescent object counts were taken and plotted over time and by dose. B-C represent one biological replicate, similar results were seen in other experiments (not shown).

To determine the kinetics and cell death mechanism of KPT-330 induced cell death, time-course apoptosis experiments were conducted with MDA-MB-231 cell (Figure 4.2 B-C). MDA-MB-231 cells show both a dose-dependent and time-dependent increase in apoptosis after

treatment that mirrors the cell death monitoring, indicating MDA-MB-231 cells are dying via apoptosis after treatment with KPT-330. Regarding dosing, cell death and apoptosis induction is most noticeable at doses above 200 nM. Regarding timing, apoptosis induction begins at 12-24 hours post treatment and reaches a plateau at approximately 48 hours. This will later inform the design of the RNA-Seq experiments as any transcriptomic changes after 24 hours are likely to be driven simply by apoptosis machinery and unlikely to yield interesting molecular insights into the mechanism of XPO1 inhibition-mediated cell death.

Survivin gene expression was shown to change after XPO1 inhibition previously in the breast cancer cell lines (Cheng *et al.*, 2014). Therefore, to verify these results and determine appropriate time-points for the RNA-Seq analysis, RT-qPCR experiments were performed to determine the extent of expressional changes at 2, 8 and 24 hours (Figure 4.3) in MDA-MB-231 and MCF-7 cells. By 24 hours, KPT-330 treated MCF7 and MDA-MB-231 cells showed a significant decrease in Survivin gene expression compared to vehicle control. Expression of Survivin in MDA-MB-231 cells is significantly decreased by 8 hours but not in MCF-7 cells. This indicates that XPO1 inhibition can affect gene expression (at least of Survivin) by at least 8 hours after treatment. Given the controversy regarding the Survivin-driven mechanism suggested by Cheng et al (as discussed in the introduction to this chapter), it is perhaps unsurprising that Survivin expression changes aren't occurring at earlier time-points. Particularly in MCF-7 cells, the gene expression levels aren't changing in the first 8 hours after treatment and in MDA-MB-231 cells Survivin gene expression is only reduced by about 50% at 8 hours. This of course doesn't rule out a Survivin-driven mechanism entirely, but this along with the reasons discussed in the introduction indicate that an unbiased and wholistic approach is needed to further explore the mechanism of action of KPT-330.

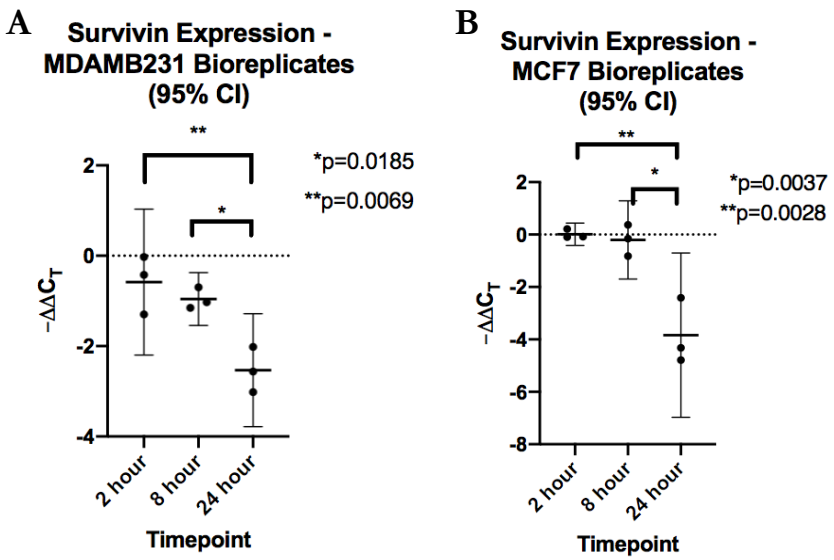


Figure 4.3 Survivin (BIRC5) gene expression changes after XPO1 Inhibition with KPT-330
MDA-MB-231 (**A**) and MCF-7 (**B**) cells were treated with vehicle (DMSO) or 200nM KPT-330 for the amount of time indicated. Then cells were lysed, and RT-qPCR was performed. Change in C_T between Survivin and B2M (as a control) compared to treated and vehicle conditions ($-\Delta\Delta C_T$) is plotted with 95% CI. Change over time is also compared via t-test, with p-values indicated. Comparisons between 2- and 8-hour values are not significant and thus not indicated on the graph.

Imputation-Based Insights into XPO1 Inhibition in Breast Cancer

My interest in XPO1 inhibition as a potential therapeutic for breast cancer began with the identification of leptomycin B as the second most significant hit from the differential imputing drug response analysis shown in the volcano plot in Figure 3.3A. Similar to the imputed results for AZD-1775, leptomycin B was predicted to be more effective in TNBC compared to every other breast cancer subtype regardless of use of clinical (Figure 4.4A) or PAM50 (Figure 4.4B) classification. This is in line with Arango et al that demonstrated that XPO1 inhibition is particularly effective in TNBC. Figure 4.1 shows a heterogeneous response profile across breast cancer subtype, but similarly there was a difference in that only 66% of ER+ breast cancer cell lines responded compared to a 94% response rate in TNBC. Combining this with the data from Arango et al, the ER+ breast cancer cell line had between a 1/7 (Arango) to a 8/12 (CTRP) response rate whereas the TNBC cell lines had between a 14/14 (Arango) and 16/17 (CTRP) response rate. In general, then, the imputed drug response values (which are built

using pan-cancer information) do seem able to recapitulate the increased efficacy of XPO1 inhibition in TNBC compared to HR+ breast cancers.

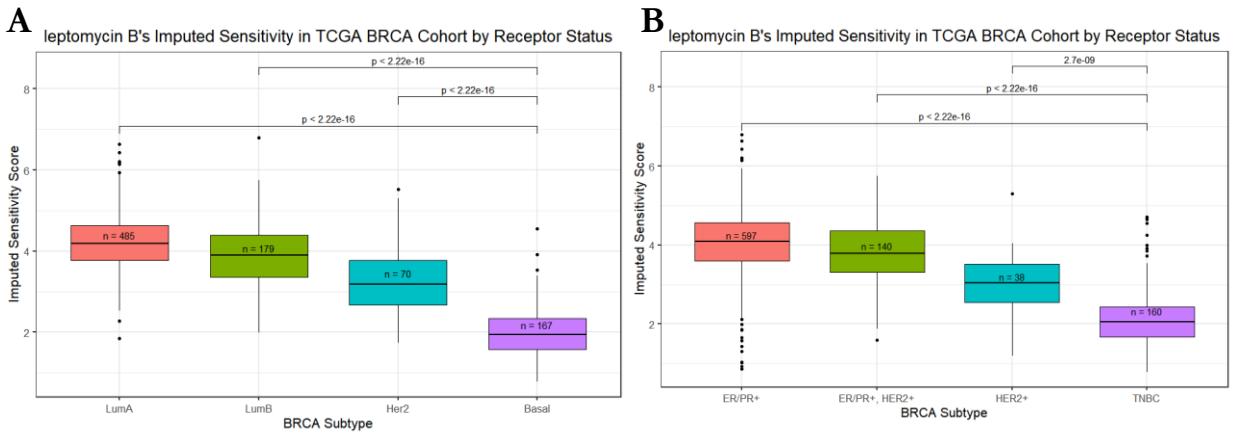


Figure 4.4. Leptomycin B Imputed Drug Sensitivity by Breast Cancer Subtype
(**A-B**) Leptomycin B imputed sensitivity in TCGA breast cancer tumors by receptor status (A) and PAM50 subtyping (B). Boxplots summarize results of each tumor sample's imputed sensitivity score to Leptomycin B in the TCGA breast cancer cohort by subtype. The n values indicate the number of patients in each group and p-values shown are adjusted for multiple testing. Lower values on the y-axis indicate increased predicted sensitivity. Dataset Abbr: TCGA, The Cancer Genome Atlas

In Chapter 3, I performed biomarker discovery with imputed drug response values as well as gene-set enrichment analysis to determine whether AZD-1775 imputed drug response values were consistent with the mechanism of action of Wee1 inhibition. To explore the mechanism of action of XPO1 inhibition, I performed similar IDWAS and gene-set enrichment experiments. After performing the IDWAS methodology using leptomycin B imputed sensitivity, there were only two significant gene associations that were also above the median effect size: NOTCH3 (FDR = 0.02, effect size = -1.7) and API5 (FDR = 0.04, effect size = -2.1). NOTCH3 is part of the NOTCH family of cell surface receptors that are commonly mutated in cancer, particularly in TNBC (Giuli *et al.*, 2019). API5 protein suppresses the transcription factor E2F1-induced apoptosis and also interacts with, and negatively regulates Acinus, a nuclear factor involved in apoptotic DNA fragmentation (Morris *et al.*, 2006). Both effect sizes indicate that XPO1 inhibition is more effective in the mutated setting, which would

indicate a situation where aberrant NOTCH signaling is present and E2F-induced apoptosis is being repressed. NOTCH downstream signaling is often associated with Myc, E2F, and Cyclin D3 overexpression (Giuli *et al.*, 2019), indicating some biological overlap among these two mutations that associate with XPO1 imputed response.

I also performed GSEA on the TCGA breast cancer patients to determine what gene sets are enriched with imputed leptomycin B sensitivity. Tables 4.2 and 4.3 show the significantly associated pathways of the Hallmark and Oncogenic Signature gene sets. Both analyses indicate that Myc and E2F/Rb signaling pathways are associated with imputed sensitivity, as they comprise 3/4 Hallmark pathways and 3/6 Oncogenic Signature pathways. This also agrees with the IDWAS results described in the previous paragraph as NOTCH3 and API5 have direct relations to Myc and E2F signaling respectively. It should be noted that E2F is a Myc target gene, and as such overlap between these two pathways should be expected and further experiments would be needed to determine the relative importance of each. Overall, though, the imputation data alone strongly indicate Myc and/or E2F signaling as particularly relevant to the response of XPO1 inhibitor leptomycin B.

Hallmark Pathway	**NES**	**FDR q-val**	**FWER p-val**
HALLMARK_MYC_TARGETS_V1	-2.21289	5.68E-04	0.001
HALLMARK_E2F_TARGETS	-2.16205	2.84E-04	0.001
HALLMARK_MYC_TARGETS_V2	-2.13598	4.15E-04	0.002
HALLMARK_G2M_CHECKPOINT	-2.07389	0.001377253	0.004

Table 4.2 Hallmark gene sets that associate with predicted sensitivity to XPO1 inhibition via GSEA Imputed drug response values to leptomycin B based on CTRP data were calculated for all TCGA breast cancer patients. Then using the imputed drug response values as a continuous phenotype label, GSEA enrichment was performed using the MSigDB Hallmark signature gene set. The top 4 pathways are shown here. NES: Normalized Enrichment Score. FWER (familywise-error rate) is more conservative than FDR, but both correct for multiple testing.

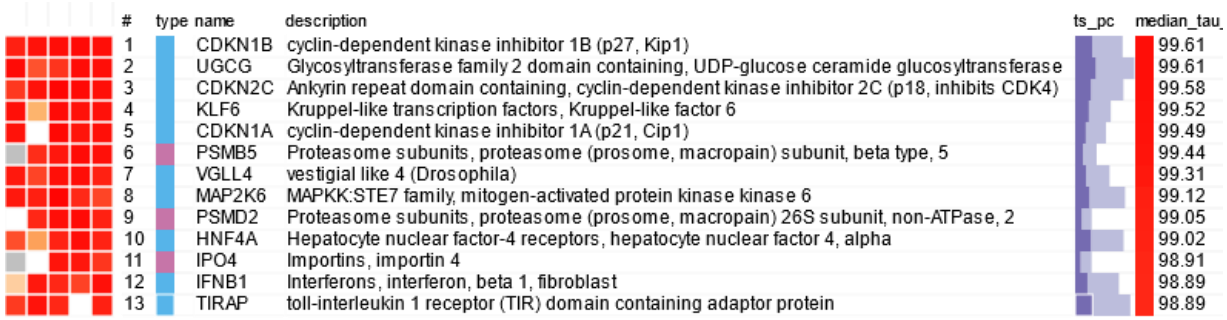
Oncogenic Signature Set	NES	FDR q-val	FWER p-val
CSR_LATE_UP.V1_UP	-2.0081375	0	0.004256156
MYC_UP.V1_UP	-1.9385983	0.006134969	0.008014669
RPS14_DN.V1_DN	-1.8072472	0.008810572	0.021875368
RB_P107_DN.V1_UP	-1.7099109	0.019823788	0.047505993
PRC2_EZH2_UP.V1_DN	-1.6948231	0.007317073	0.04268487
E2F1_UP.V1_UP	-1.641358	0.035634745	0.057982318

Table 4.3 Oncogenic Signature gene sets that associate with predicted sensitivity to XPO1 inhibition via GSEA

Same as table 4.2 except GSEA was performed with the Oncogenic Signature gene set from MSigDB. The top pathways with an FDR of less than 0.05 are shown here. NES: Normalized Enrichment Score. FWER (familywise-error rate) is more conservative than FDR, but both correct for multiple testing.

To determine if the Myc and E2F/Rb signaling pathways were consistent with other publicly available data, I investigated the Broad Institute's Connective Map (cMAP) resource (https://clue.io/, date accessed: 04/03/2021). This resource takes before and after treatment microarray gene expression data for a set of cell lines and compares these values with other perturbations to determine which perturbations are most similar to the perturbation of interest. cMAP contained one XPO1 inhibitor molecule 7b-cis (aka BRD-K61829047). When comparing 7b-cis induced transcriptional changes to gene knockdown and overexpression, the expression signature matched that of CDK inhibitors p27, p21, and p18 (Figure 4.5). These CDK inhibitor associations make up 3 of the top 5 associations. This would be consistent with the pathways noted from the imputation only response as E2F/Rb pathways and Myc pathway. To overview the CDK/Rb/E2F pathway, Rb binds to and inhibits the transcriptional activity of E2F proteins, which promote the cell cycle. CDKs (particularly CDK4, CDK6, and CDK2) put an inhibitory phosphate group on Rb. In a hyperphosphorylated state, Rb is inactive and degraded and thus allows for E2F signaling. Myc, as stated earlier, is also related in that two Myc target genes are E2F and Cyclin D. Thus, taken as a whole, the imputed results identifying E2F and

Myc gene signatures as relevant for leptomycin B response appears quite consistent with the cMAP associations found upon comparing transcriptional changes induced by 7b-cis.



#	type	name	description	ts_pc	median_tau
1		CDKN1B	cyclin-dependent kinase inhibitor 1B (p27, Kip1)		99.61
2		UGCG	Glycosyltransferase family 2 domain containing, UDP-glucose ceramide glucosyltransferase		99.61
3		CDKN2C	Ankyrin repeat domain containing, cyclin-dependent kinase inhibitor 2C (p18, inhibits CDK4)		99.58
4		KLF6	Kruppel-like transcription factors, Kruppel-like factor 6		99.52
5		CDKN1A	cyclin-dependent kinase inhibitor 1A (p21, Cip1)		99.49
6		PSMB5	Proteasome subunits, proteasome (prosome, macropain) subunit, beta type, 5		99.44
7		VGLL4	vestigial like 4 (Drosophila)		99.31
8		MAP2K6	MAPKK:STE7 family, mitogen-activated protein kinase kinase 6		99.12
9		PSMD2	Proteasome subunits, proteasome (prosome, macropain) 26S subunit, non-ATPase, 2		99.05
10		HNF4A	Hepatocyte nuclear factor-4 receptors, hepatocyte nuclear factor 4, alpha		99.02
11		IPO4	Importins, importin 4		98.91
12		IFNB1	Interferons, interferon, beta 1, fibroblast		98.89
13		TIRAP	toll-interleukin 1 receptor (TIR) domain containing adaptor protein		98.89

Figure 4.5 cMAP Data Shows Genetic Alterations that Induce Similar Transcriptional Changes as XPO1 Inhibition.
Graph was generated from cMAP online portal (clue.io) using the connections ("/CONN") between genetic perturbations and XPO1 inhibitor 7b-cis. The 5 red columns indicate the association for 4 cell lines (HCC515, HT29, MCF7, and VCAP) and the fifth column indicates the level of aggregate association. The overall Tau indicates a measure similar to that done in Gene Set Enrichment and is expressed as a number between -100 and 100, and can be read as a measure of specificity or uniqueness of the perturbation having the same effect as (in this case) XPO1 inhibition. Under type, blue represents overexpression and pink represents gene knockdown.

Overall, the imputation and cMAP data appear concordant and indicate a mechanism that involves Rb/E2F or Myc signaling. This is consistent with other literature that have shown that XPO1 inhibition can increase the nuclear accumulation of Rb and p27 in bladder cancer (Baek *et al.*, 2018) and p53, IκB, p21 and p27 in multiple myeloma (Tai *et al.*, 2014). Additionally, as shown in Table 4.1, XPO1 target proteins and that XPO1 target proteins include: E2F7, p21, and cyclin D1.

<u>RNA-Seq Based Assessment of XPO1 Inhibitor-Induced Transcriptomic Changes in Two Breast Cancer Cell Lines</u>

To further assess the mechanism of XPO1 inhibitor-mediated cell death, I designed and performed an RNA-Seq experiments to assess the transcriptomic changes induced after XPO1 inhibition. For this, I performed the assay at 2, 8, and 24 hours. These times were chosen as they are common times to assay transcriptomic changes but also because apoptosis is steadily induced by 24 hours, so assaying at these timepoints would hopefully give us snapshots into the

processes happening between XPO1 inhibition and cell death. I chose two cell lines, MCF-7 cells and MDA-MB-231. While the imputed drug response indicated that XPO1 inhibition was more sensitive in TNBC than RPBC, using an ER+ breast cancer cell line is of interest still since these cells are driven by distinct cancer drivers. Thus, investigating both an ER+ cell line and TNBC cell line would help to assess whether the heterogenous response seen in the CTRP data was due to distinct biological mechanisms or if XPO1 inhibition is inducing the same transcriptomic changes. Additionally, the mechanism of action I proposed from the imputed drug response results was Myc and E2F related, but these results were based on drug sensitivity in TNBC. To see if this mechanism is consistent even in an ER+ case, MCF-7 cells were used.

The sequencing was performed on 36 samples (2 cell lines, 2 treatment conditions, 3 time points) and carried out on an Illumina NovaSEQ 6000. Samples were either treated with 200 nM KPT-330 or vehicle control for all time points indicated. The raw reads were trimmed using Trimmomatic, mapped with HISAT2, and counted using featureCounts. A table summarizing some of the quality control metrics obtained from Picard, samtools, and HISAT2 is provided (Table 4.4). The RNA-Seq results appear quality as the average mapping rate is 98.6%, with around 28,000,000 reads being assigned to around 16,575 genes for each sample. Additionally, the small differences between minimum and maximum values of Table 4.4 indicate consistent results across all the samples.

METRIC	Average	Minimum	Max
Total Fragments	30221888	28962503	31677758
Unmapped	400905	350195	464019
Uniquely Mapped	28193633	26860250	29649909
Multiply Mapped	1280216	1012321	1619226
Discordantly Mapped	347132	189656	497571
Mapping Rate	0.9867	0.9848	0.9887
PCR Duplicate	13320904	12245270	14166386
MAPQ = 0	41661	33080	50166
Max Read Length	101	101	101
Average Read Length	99	99	99
Average Read Quality	36.49	36.4	36.5
High Quality Rate	0.945	0.93	0.952
Median Insert Size	215	210	225
Mean Insert Size	231.54	225.534	243.493
Insert Size StdDev	97.355	90.988	106.894
Genes Detected	16575	16149	17345
Expression Profiling Efficiency	0.851	0.825	0.86

Table 4.4 RNA-Seq Quality Control Metrics
RNA-Sequencing was performed on 36 samples. Raw reads were mapped/aligned using HISAT2 to GRCh38. Quality control metrics were obtained from Picard, samtools, and HISAT2. The mean, minimum and maximum value for the indicated metric are shown in the table above.

As a further quality control, principle component analysis (PCA) was performed on all the samples (Figure 4.6A). The first PC, which accounts for approximately 82% of the variation, separates the samples by cell line, which matches the transcriptional differences expected for two biologically distinct cell lines.

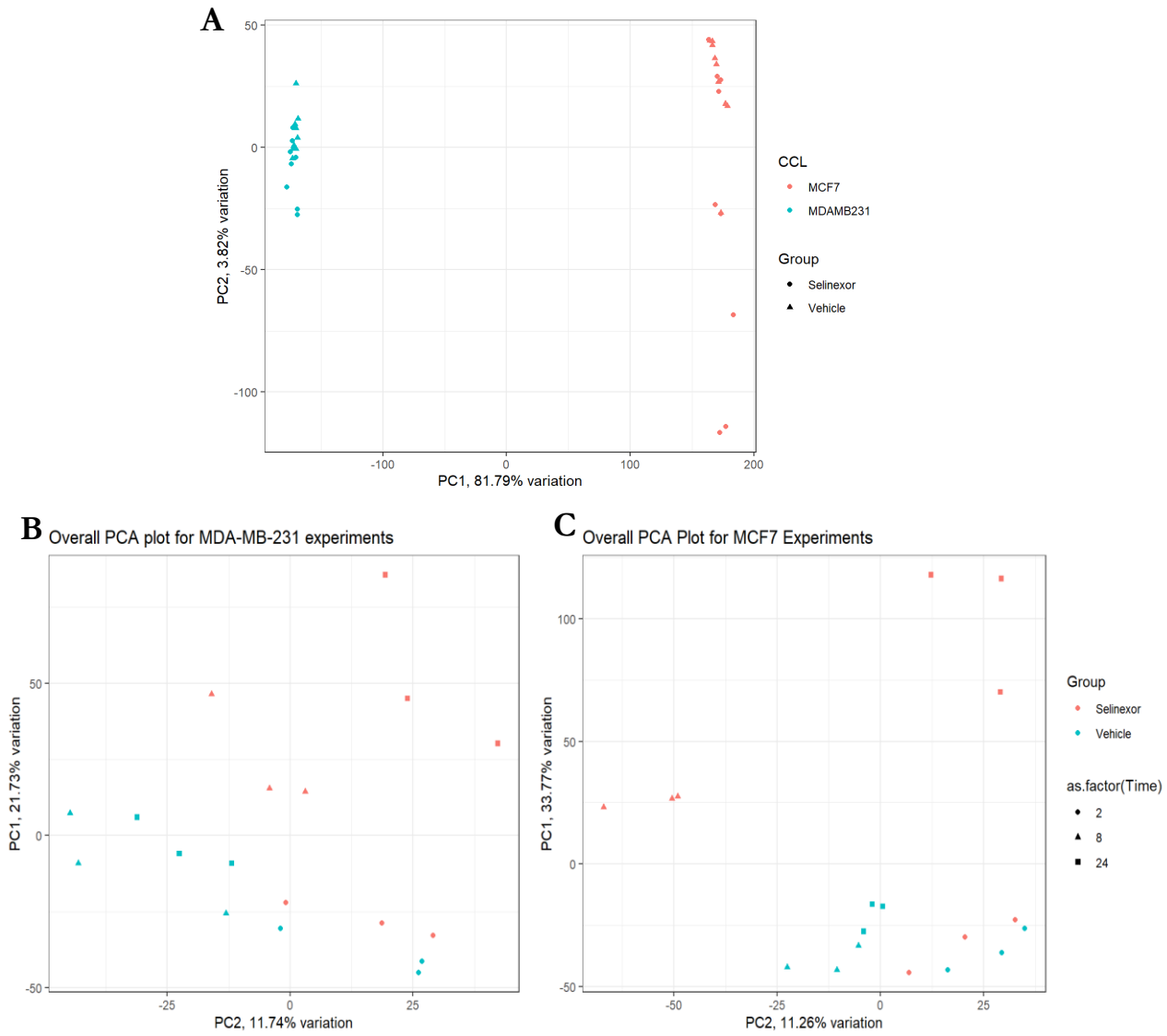


Figure 4.6 PC Analysis for MDA-MB-231 and MCF-7 Transcriptional Changes After KPT-330 Treatment

Raw counts for the 36 gene expression values were converted to CPM value and then PCA analysis was performed for all the samples (**A**), the MDA-MB-231 samples (**B**), and the MCF-7 samples (**C**). Note the meaning of the colors change and shapes change between plots A and B-C. In A, the color separate samples by CCL used whereas in B-C color separates the treated/control samples. In A, shape is used for treatment whereas shape is used to represent the timepoint of the samples in B-C.

When the PCA is performed on just MDA-MB-231 (Figure 4.6B) or MCF-7 (Figure 4.6C), the samples begin to cluster by treatment and time. For both cell lines, samples that were treated with KPT-330 for 8 and 24 hours are shifted to larger positive values on the PC1 axis.

PC1 accounts for 22% and 34% of the variation for MDA-MB-231 and MCF7 respectively. This indicates that the largest proportion of the variation seen in these samples is explained by KPT-330 treatment. The vehicle treated samples cluster together on PC1 and PC2 regardless of time point. The 2-hour KPT-330 treated samples cluster with the vehicle treated samples, indicating at 2 hours there are not many differentially expressed genes.

Differential gene expression analysis was then performed using the DESeq2 framework. The volcano plots in Figure 4.7 show a number of significantly differentiated genes at each time-point for both cell lines. As seen in the graphs, the number of differentially expressed genes increases as well as the magnitude of the fold change. That the 2-hour samples are the most similar to the control treated samples.is consistent with the PC analysis of Figure 4.6. Overall, the MCF-7 cell lines show more differentially expressed genes at the 8- and 24-hour time points than the MDA-MB-231, both in terms of the number of significantly differentially expressed genes as well as the fold-change. In both cases, however, it is clear that KPT-330 treatment is significantly changing the transcriptome of these cell lines. There are 128, 1766, 2249 differentially expressed genes for MDA-MB-231 and 89, 3528, 5578 differentially expressed genes for MCF-7 for the 2-, 8-, and 24-hour time points respectively.

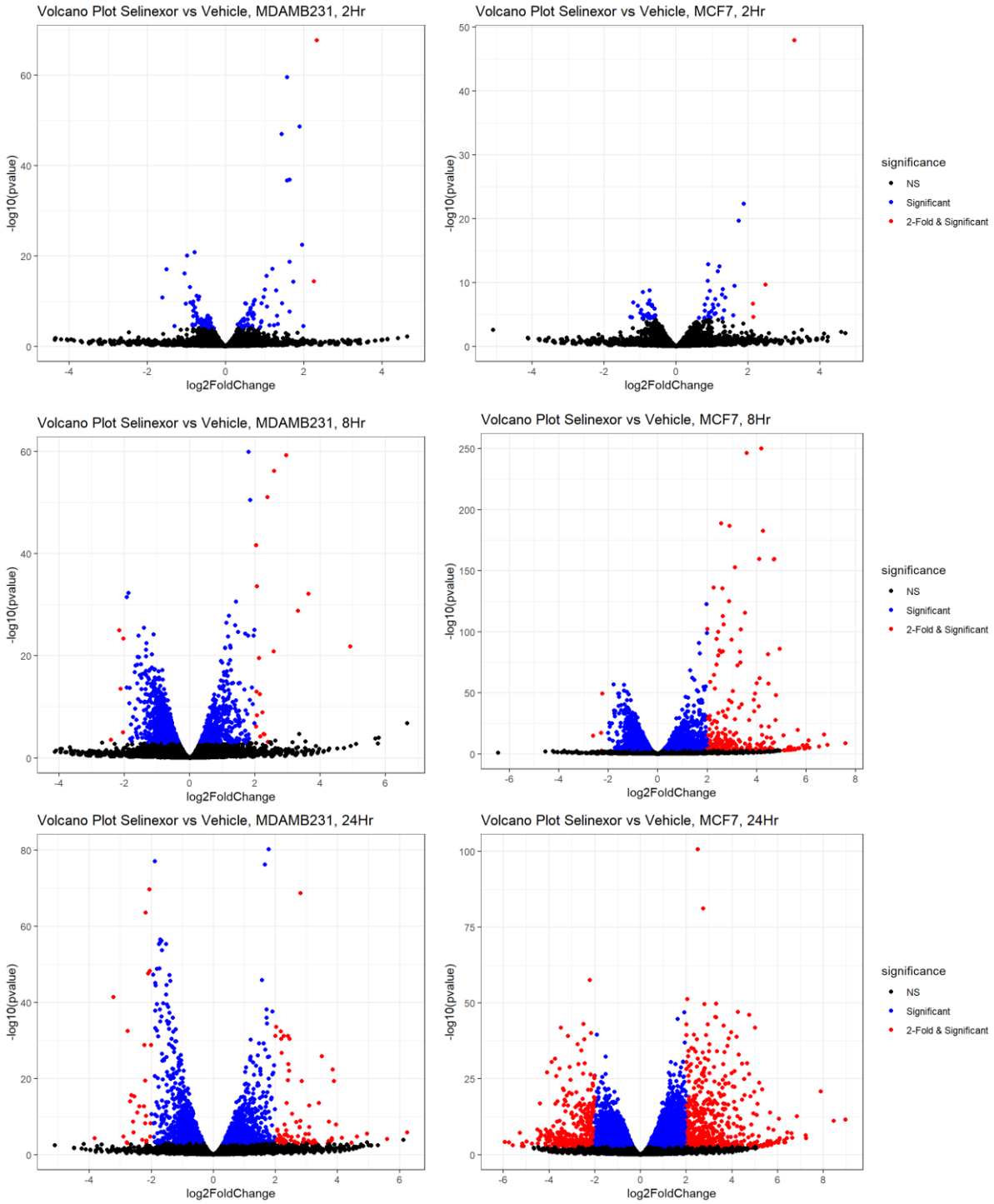


Figure 4.7. Volcano Plots of Differentially Expressed Genes Over Time.
For MDA-MB-231 (first column) and MCF-7 (second column), differential gene expression analysis was performed between KPT-330 and vehicle treated samples via DESeq2. For each time point indicated, the log2 fold change (x-axis) is plotted against the log10 of the adjusted p-value (y-axis). Blue indicates genes that achieve a significance of less than 0.10 adjusted p-value and red points are both significant and have an absolute log2 fold change greater than 2 (i.e. a 4-fold change). Note the x- and y-axes are free for each graph.

There are some commonalities among the differentially expressed genes at each time point. For example, at the 2-hour time-point there are 31 genes that are differentially expressed in both the MCF-7 and MDA-MB-231 cell lines. At 8 hours, there are 737 differentially expressed genes in common between the two cell lines and at 24-hours there are 993 genes in common. However, as can be seen in the upcoming sections, the top genes and pathways that are differentially expressed are varied between the two cell lines. Therefore, the next sections will look at the genes and pathways that are significantly altered for MCF-7 and MDA-MB-231 individually. Also, for simplicity the 8-hour timepoints will be the focus since as can be seen from the volcano plots, this is the earliest time-points to have a substantial change in the transcriptome after KPT-330 treatment.

MCF-7 Differentially Expressed Genes and Gene-Set Enrichment Analysis

Of the differentially expressed genes in MCF-7, 57 (of 89 possible) genes are consistently differentially expressed between the 2 and 8-hour timepoints and exactly 1000 (of 3489 possible) genes are differentially expressed at both 8 and 24-hour timepoints. To begin getting at which of the differentially expressed genes could help explain the XPO1 inhibition-mediated cell death, the list of differentially expressed genes were filtered for genes consistent with the Cancer Gene Census (CGC), a list of cancer-relevant genes provided by the Broad institute (Futreal *et al.*, 2004). Table 4.5 shows the differentially expressed genes with CGC entries at the 8-hour timepoint. All the genes shown here are also significantly differentially expressed at the 24-hour timepoint, indicating consistent and stable gene expression changes among these top genes at both time-points.

CGC Genes Differentially Expressed in MCF-7 at 8 Hours						
Gene Symbol	**log2(FC)**	**padj**	**Name**	**Tier**	**Hallmark**	**Role**
FAS	4.188	1.16E-246	Fas cell surface death receptor	1	Yes	TSG
MDM2	2.562	1.49E-185	Mdm2 p53 binding protein homolog	1	NA	oncogene
DDB2	2.007	7.32E-100	damage-specific DNA binding protein 2	1	Yes	oncogene, TSG
XPC	1.674	1.99E-88	xeroderma pigmentosum, complementation group C	1	NA	TSG
PPM1D	1.7	2.89E-80	protein phosphatase, Mg2+/Mn2+ dependent 1D	1	Yes	oncogene
CDKN1A	4.128	6.60E-60	cyclin dependent kinase inhibitor 1A	2	NA	oncogene, TSG
ANK1	4.004	5.52E-56	ankyrin 1	2	NA	NA
MSH6	-1.351	1.66E-54	mutS homolog 6 (E. coli)	1	NA	TSG
NOTCH1	1.679	4.69E-34	Notch homolog 1, translocation-associated (Drosophila) (TAN1)	1	Yes	oncogene, TSG, fusion
BAX	1.235	2.66E-33	BCL2 associated X, apoptosis regulator	1	NA	TSG

Table 4.5 Cancer Gene Census Genes Differentially Expressed in MCF-7 at 8 Hours
Differentially expressed genes were determined using DESeq2 with an expression change of +/- 50% (2 fold change) and adjusted p-value of 0.05. This gene list was then filtered for genes that appear in the Cancer Gene Census (CGC) to filter for cancer-relevant genes. Tier, Hallmark, and Role information all come from the CGC. Tier indicates the level of evidence (with 1 being substantial evidence and 2 being strong indications). Hallmark indicates the relation of the gene to a hallmark of cancer that are manually defined by the CGC.

From this list of genes, we can see many TP53 target genes (FAS, MDM2, BAX) and DNA-repair genes (DDB2, XPC, MSH6) being highly differentially expressed in MCF-7 cells. To examine the transcriptome more holistically, KEGG pathway analysis was performed using the differentially expressed gene list at 8 and 24 hours. The only significantly enriched pathway for the 8-hour gene set was the KEGG p53 signaling pathway (Figure 4.8). This is both consistent with the gene targets seen in Table 4.5 but also show that the gene changes are specifically associated with p53 pathway activation. At 24-hours, there are more pathways that are differentially expressed, but many of these are related to p53 signaling. The differentially expressed KEGG pathways at 24 hours in order of significance are: cell cycle, DNA replication,

p53 signaling, Fanconi Anemia pathway, homologous recombination, and the MAPK signaling pathway.

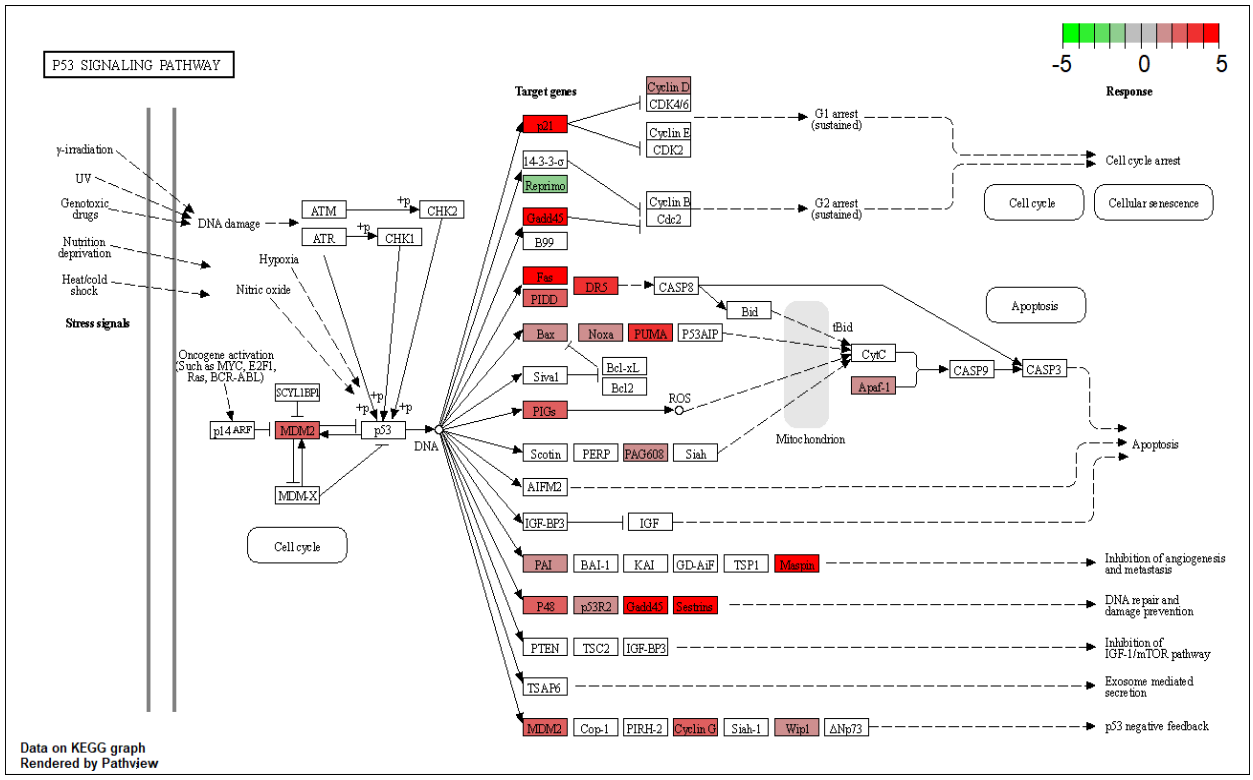


Figure 4.8 P53 Signaling Pathway After 8 Hour KPT-330 Treatment in MCF-7 Cells
KEGG pathways were enriched for using the differentially expressed genes obtained from DESeq2 for MCF-7 cells at 8 hours after treatment. The results are plotted on this KEGG pathway diagram for p53 signaling. Red indicates an increase in expression, green a decrease in expression, and gray indicates that these genes were not significantly differentially expressed at the 8 hour timepoint.

It is well established that MCF-7 cells express wildtype p53, but that the wildtype protein is excluded to the cytoplasm of the cell (Moll, Riou and Levine, 1992; Stommel *et al.*, 1999). Not only this, but it has been previously shown that inhibition of nuclear export by leptomycin B results in the retention of p53 in the nucleus in MCF7 cells (Takahashi *et al.*, 1993; Lu *et al.*, 2000). The pathway analysis and genes seen differentially expressed are consistent with these observations and indicate that changes in p53 signaling can explain the largest proportion of gene expression changes happening in MCF7 cells post XPO1 inhibition (see discussion).

MDA-MB-231 Differentially Expressed Genes and Gene-Set Enrichment Analysis

For MDA-MB-231, the number of differentially expressed genes was lower at all time points compared to MCF7. 86 (of 128) genes that are differentially expressed at the 2-hour timepoint are still differentially expressed at the 8-hour timepoint. Of the 1756 genes differentially expressed at the 8-hour timepoint, 721 of these genes are also differentially expressed at the 24-hour timepoint. However, looking at the genes themselves, there were only 12 of these 721 genes that overlap between the differentially expressed genes at 8 hours and the genes that make up the cancer gene census (Table 4.6). These genes are also not consistently differentially expressed, but this may be expected (statistically) compared to MCF-7 since the magnitude of the p-adjust is much lower in the MDA-MB-231 cells.

CGC Genes Differentially Expressed in MDA-MB-231 at 8 Hours						
Gene Symbol	**log2(FC)**	**padj**	**Name**	**Tier**	**Hallmark**	**Role**
CDKN2C	-1.157	2.27E-18	cyclin-dependent kinase inhibitor 2C (p18, inhibits CDK4)	1	NA	TSG
SGK1	1.295	4.26E-13	serum/glucocorticoid regulated kinase 1	2	NA	oncogene
BUB1B	-1.222	3.61E-10	BUB1 budding uninhibited by benzimidazoles 1 homolog beta (yeast)	1	Yes	TSG
BARD1*	-1.027	4.61E-10	BRCA1 associated RING domain 1	1	NA	TSG
JUN*	-1.376	7.66E-10	jun oncogene	1	NA	oncogene
CIITA	-1.85	1.49E-09	class II, major histocompatibility complex, transactivator	1	Yes	TSG, fusion
BCL6	1.077	1.22E-07	B-cell CLL/lymphoma 6	1	NA	oncogene, fusion
KNL1	-1.209	4.24E-07	cancer susceptibility candidate 5	1	Yes	TSG, fusion
LCP1*	1.003	3.17E-04	lymphocyte cytosolic protein 1 (L-plastin)	2	NA	fusion
PRDM1	1.255	3.31E-03	PR domain containing 1, with ZNF domain	1	NA	TSG
ACKR3*	1.424	0.00347	atypical chemokine receptor 3	1	Yes	oncogene, fusion
NCOR2*	-1.225	0.0376	nuclear receptor corepressor 2	1	Yes	TSG

Table 4.6 All Cancer Gene Census Genes Differentially Expressed in MDA-MB-231 at 8 Hours
Similar to table 4.5, differentially expressed genes were determined using DESeq2 with a fold-change cut off of 1 and adjusted p-value of 0.05. This gene list was then filtered for genes that appear in the Cancer Gene Census (CGC), which provides the Tier, Hallmark, and Role information. *Indicate this gene is no longer significantly differentially expressed at the 24 hour timepoint.

Pathway analysis was performed with these sets of genes to determine if there are any more consistent global transcriptional changes in these samples. There were no significantly enriched pathways in the 2-hour data. Here, I focus on the 8-hour samples since these samples are the first time points with significant differences in the transcriptome. Additionally, the pathways that are differentially expressed at 8 hours are also differentially expressed at 24 hours. KEGG gene enrichment was performed. Cell cycle and some DNA-Repair pathways were significantly downregulated after treatment with KPT-330 (Table 4.7, Figure 4.9-4.10). An important note is that unlike the MCF-7 cells, the only DNA-repair pathways that are downregulated in the MDA-MB-231 cells are homologous recombination and Fanconi Anemia pathways (Figure 4.10, which include the BRCA genes). These pathways are specifically involved in the recombination repair of DNA damage and are most active in the S and G2 phase of the cell cycle. However, as can be seen in Figure 4.9, the cell cycle is also highly downregulated. Thus, cell cycle downregulation could then explain the downregulation seen in the Fanconi anemia, homologous recombination, and DNA replication pathways seen.

ID	Description	GeneRatio	BgRatio	pvalue	qvalue
hsa04110	Cell cycle	38/658	124/8105	1.93E-13	5.31E-11
hsa03460	Fanconi anemia pathway	19/658	54/8105	1.80E-08	2.48E-06
hsa03440	Homologous recombination	13/658	41/8105	1.23E-05	0.001133
hsa03030	DNA replication	11/658	36/8105	8.48E-05	0.005751
hsa04068	FoxO signaling pathway	24/658	131/8105	0.000119	0.005751
hsa04218	Cellular senescence	27/658	156/8105	0.000125	0.005751
hsa05202	Transcriptional misregulation in cancer	31/658	192/8105	0.000154	0.006078

Table 4.7 KEGG Pathways Enriched in the MDA-MB-231 Cell Lines at 8 Hours
Differential gene expression analysis was performed on gene expression of MDA-MB-231 cells between 8-hour vehicle and KPT-330 treatment. The genes that were significantly differentially expressed were then used in an enrichment analysis with the KEGG gene pathways. The 7 most significant results are displayed above. Gene ratio is the number of differentially expressed genes in the pathway over the number of genes given, BgRatio is the ratio of genes in the pathway to the total number of KEGG genes. These ratios are compared in an enrichment analysis to generate the p values listed. The qvalue is the FDR corrected p-value.

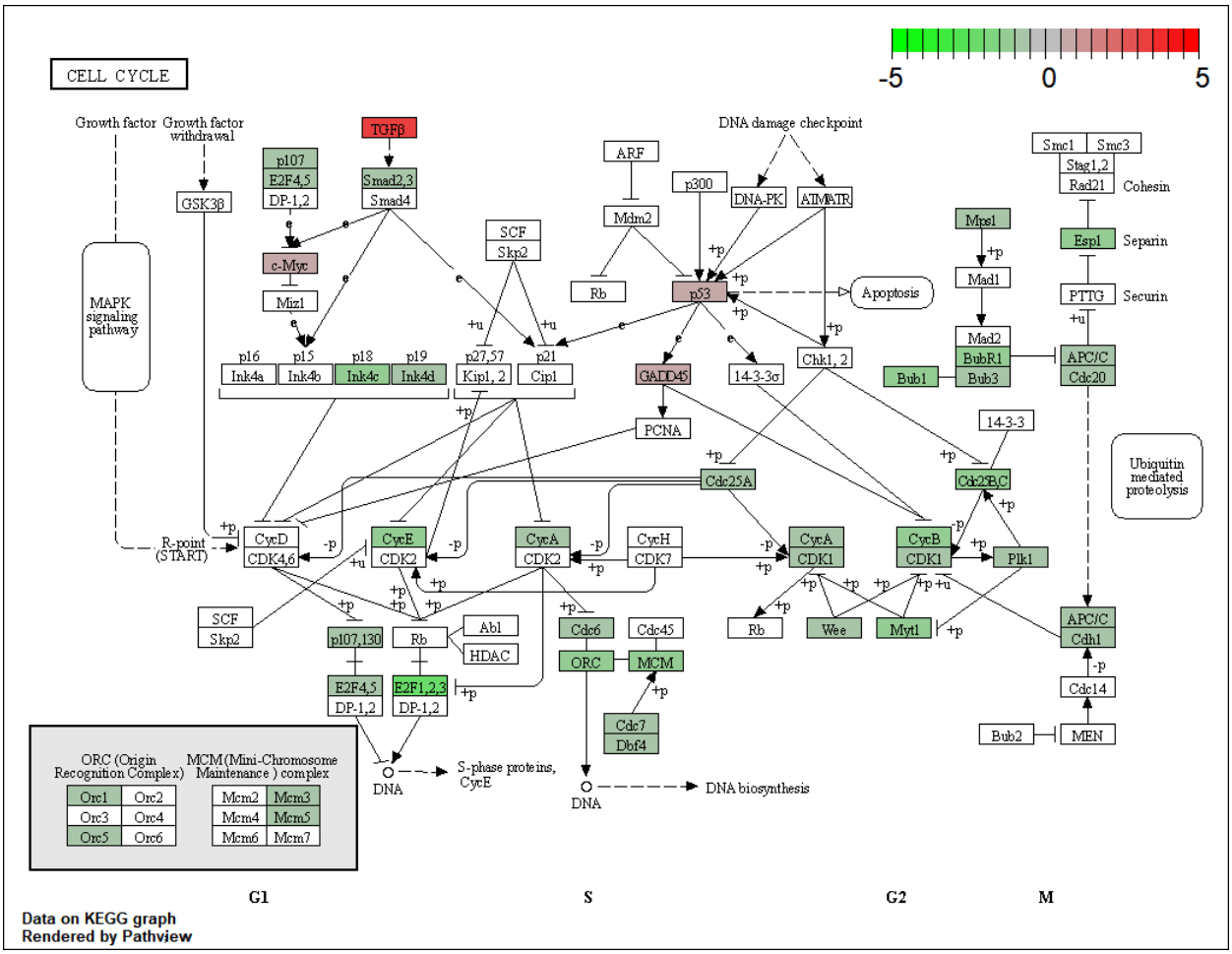


Figure 4.9 Cell Cycle Signaling Pathway After 8 Hour KPT-330 Treatment in MDA-MB-231 Cells
KEGG pathways were enriched for using the differentially expressed genes obtained from DESeq2 for MDA-MB-231 cells at 8 hours after treatment. The cell cycle pathway was the most significantly enriched for pathway. The fold change of the significantly expressed genes are color mapped onto the KEGG cell cycle diagram. Red indicates an increase in expression, green a decrease in expression, and white indicates that these genes were not significantly differentially expressed at the 8-hour timepoint.

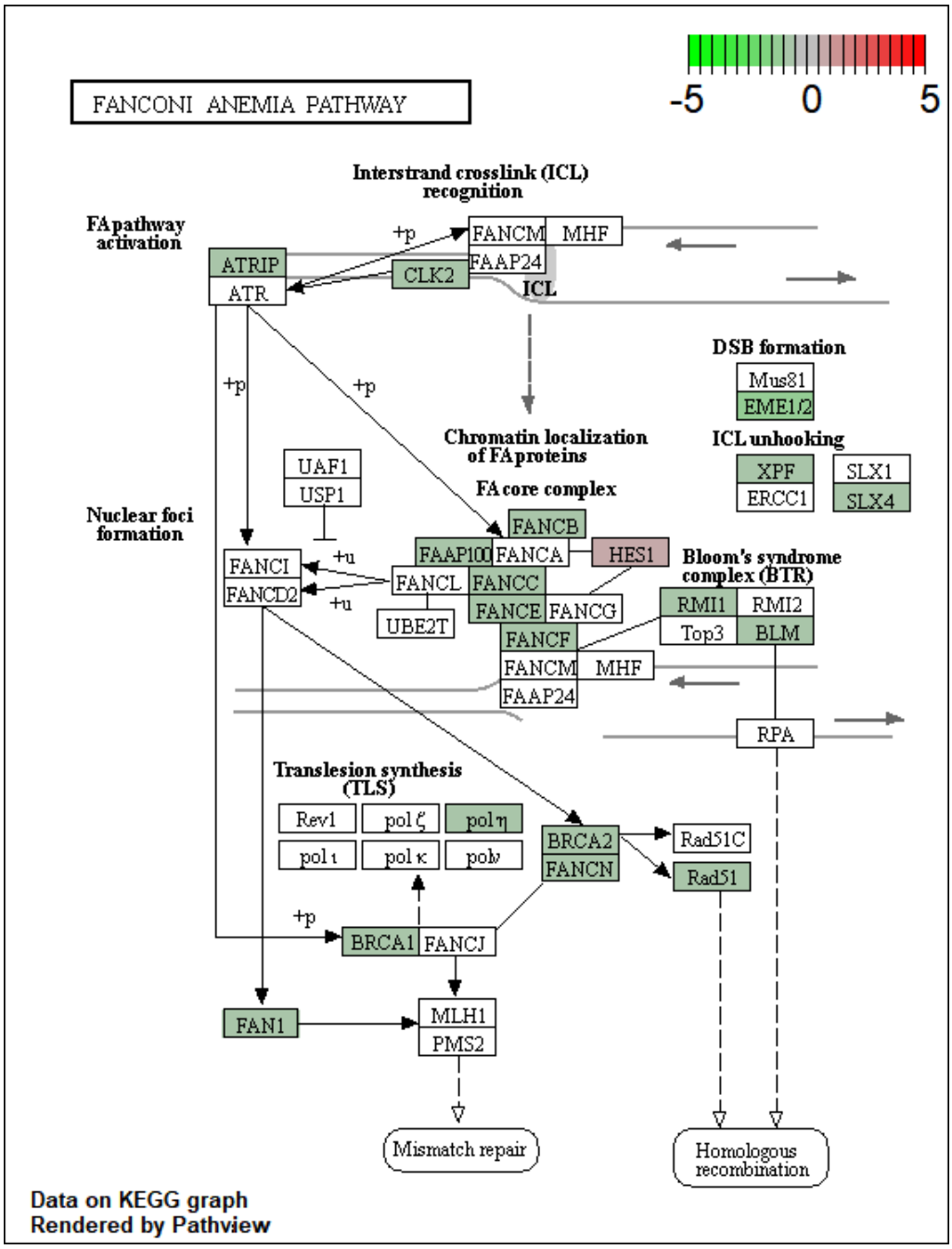


Figure 4.10 Fanconi Anemia Pathway After 8 Hour KPT-330 Treatment in MDA-MB-231 Cells
Similar to Figure 4.9, the fold change of the significantly expressed genes are color mapped onto the KEGG Fanconi Anemia pathway diagram. Red indicates an increase in expression, green a decrease in expression, and white indicates that these genes were not significantly differentially expressed at the 8-hour timepoint. This is the second most differentially expressed pathway in MDA-MB-231 cells after 8-hour treatment with KPT-330.

Since the top 4 pathways could be explained through changes in cell cycle signaling, this pathway was investigated further. Of the cell cycle genes downregulated at the 8-hour timepoint, the most significantly differentially expressed gene is E2F2. At 24-hours, the most downregulated cell cycle genes are cyclin A and cyclin B, which would be consistent with G1/S cell cycle arrest and E2F2 downregulation as E2Fs drive the G1/S transition.

To further probe the gene expression results, I performed gene set enrichment analysis using the MSigDB Hallmark gene set. The results of this can be seen in Table 4.8. E2F targets is the most significantly enriched for pathway from this analysis, consistent with the KEGG analysis. Overall, the RNA-Seq analysis in MDA-MB-231 cells has an enrichment of cell cycle pathways that have been downregulated. This is consistent with previous literature reports of G1/S cell cycle arrest as well as is consistent with the imputation and computational analysis performed previously (see discussion).

Hallmark Gene Set ID	#genes	Enrichment Score	NES	pvalue	p.adjust	qvalues
E2F_TARGETS	138	-0.51971	-2.75545	1.00E-10	2.50E-09	1.95E-09
G2M_CHECKPOINT	156	-0.52926	-2.86167	1.00E-10	2.50E-09	1.95E-09
MITOTIC_SPINDLE	121	-0.37625	-1.95279	9.88E-06	0.000165	0.000128
MYC_TARGETS_V2	45	-0.46013	-1.95408	0.000319	0.00388	0.003022
MYC_TARGETS_V1	81	-0.37576	-1.82669	0.000388	0.00388	0.003022
INFLAMMATORY_RESPONSE	115	0.365115	1.587659	0.003616	0.030133	0.023472

Table 4.8 Gene Sets Enrichment Analysis in MDA-MB-231 Cells 8 Hours Post KPT-330 Treatment
Fold change gene expression was calculated for every gene between KPT-330 and vehicle treated MDA-MB-231 cells. All genes were then ordered by the fold-change expression and GSEA analysis was performed using the MSigDB Hallmark gene sets. NES stands for normalized enrichment score and adjusts the enrichment score based on the gene set size. Negative enrichment scores indicate that the genes in this pathway are downregulated after KPT-330 treatment.

4.4 DISCUSSION

In this analysis, I was able to show that XPO1 inhibition is capable of decreasing viability in a heterogeneous population of breast cancer cells and shows highly dichotomous sensitive/resistant response profile for both Leptomycin B as well as KPT-330. KPT-330 activates p53 signaling pathways in MCF-7 cells and highly downregulates the cell cycle of MDA-MB-231 cells. This latter analysis is also highly consistent with the computation analysis that indicated Myc and/or E2F signaling pathways were associated with imputed leptomycin B response. This work is aimed to be exploratory with the end goal being to determine what vulnerability or vulnerabilities exist in breast cancer that enables XPO1 inhibition to be effective in a variety of heterogeneous breast cancer models. To this end, it seems nuclear exclusion of p53 as well as aberrant E2F signaling may enable XPO1 inhibition-mediated cell death.

In Arango et al (2017), the CTRP dataset, and my own data, MCF-7 cells were highly responsive to XPO1 inhibition. Investigation into the gene expression changes indicated that p53 signaling is being reactivated in the MCF7 cells after KPT-330 treatment at the 8- and 24-hour timepoints. This is consistent with previous literature that show that inhibition of nuclear export by leptomycin B results in the retention and signaling of p53 in the nucleus in MCF7 cells (Takahashi *et al.*, 1993; Lu *et al.*, 2000). However, Arango et al indicated that XPO1 inhibition is independent of TP53 mutational status in breast cancer. While this may be generally true that breast cancer cells can respond to XPO1 regardless of TP53 mutation status, the evidence presented here indicates wildtype p53 may be important for the response of MCF-7 cells. In the cell lines tested by Arango et al, only one (1/8) ER+ cell lines responded to XPO1 inhibition, which is inconsistent with the CTRP data (Figure 4.1). However, the only ER+ cell line to have WT TP53 status in Arango et al (MCF-7), did respond to therapy. From the CTRP data, 5/8

ER+ cell lines that responded to leptomycin B were TP53 wildtype while of the 4 resistant ER+ cell lines, only one was TP53 wildtype.

While by no means conclusive, it may be that ER+ breast cancers with wildtype TP53 (which comprises most of ER+ breast cancers) may still be susceptible to XPO1 inhibition. P53 has been described as either mutated or nuclear excluded in breast cancer (Moll, Riou and Levine, 1992). Additionally, TP53 is mutated in approximately 80% of TNBC and only 20% of other breast cancers (Berger, Qian and Chen, 2013; Li *et al.*, 2019, p. 53). Also, XPO1 inhibition-mediated p53 activation has been shown to be a more general phenomenon in other cancers including NSCLC, melanoma, renal cancer, AML, NHL, MM, pancreatic cancers (Wang and Liu, 2019) Thus, XPO1 inhibition-based activation of p53 may also be applicable to breast cancer, and in particular ER+ breast cancers that maintain wildtype p53. While it is clear that p53 activation is associated with XPO1 response in MCF-7 cells and that this is consistent with the literature, further analysis is needed to determine if p53 activity is sufficient or necessary for XPO1 inhibition induced cell death (see future directions).

MDA-MB-231 cells on the other hand are TP53 mutated and unsurprisingly don't show evidence for TP53 pathway activation. Instead, in MDA-MB-231 we see downregulation of the cell cycle. The parsimonious explanation for the most significantly differential expressed pathways is that E2F signaling is being strongly downregulated. Consistent with this E2F2 is the most significantly downregulated gene in the cell cycle pathway in the MDA-MB-231 cells at 8-hours post treatment. E2F2 is a transcription factor that autoregulates itself, meaning E2F2 is its own transcriptional target and that inhibiting E2F activity would also lead to a decrease in E2F gene expression. This suggests that Rb, the major protein regulator of E2F transcriptional

activity, may become activated in these cells. Rb is not mutated in MDA-MB-231 cells. An important part of the G1/S phase of the cell cycle is the hyperphosphorylation of Rb by CDKs, particularly CDK4/6. This allows E2Fs to become active and transcribe genes that promote the cell cycle transition from G1 to S. Thus, seeing changes in E2F signaling could be consistent with other cancers and known XPO1 targets. The CDK inhibitor protein p21 is an established XPO1 target protein (Table 4.1) and has been shown to increase in nuclear accumulation in bladder cancers, non-Hodgkins lymphoma, and multiple myeloma (Azmi, Al-Katib, *et al.*, 2013; Tai *et al.*, 2014; Baek *et al.*, 2018).

There is considerable consistency between the pathways associated with imputed XPO1 inhibition and measured gene expression changes in MDA-MB-231 cells. Impressively, the hallmark gene sets experimentally enriched in MDA-MB-231 shown in Table 4.8 include all four gene sets that were identified based solely on imputation results (Table 4.2). The pathways identified via RNA-Seq for MDA-MB-231, the imputation analysis, and the cMAP analysis are all associated with E2F signaling and decrease in cell cycle signaling. The consistency among these pathways may give more credence to the generalizability of the MDA-MB-231 RNA-Seq results to other TNBC settings since the imputed response analysis was more broadly TNBC focused. As mentioned already, inhibition of the G1/S cell cycle checkpoint has been implicated in XPO1 response in other cancer settings and is consistent with the cMAP associations that comprise 5 pan-cancer cell lines. Overall, the mechanism identified in MDA-MB-231 may be applicable to TNBC and other cancers more generally, but more investigation is necessary to prove this association.

5. SUMMARY AND FUTURE DIRECTIONS

Throughout my dissertation, I have explored the utility of drug imputation methods for drug discovery. Drug response is a complicated biological phenomenon. Terabytes of -omics information have been generated in molecular assays trying to assess the biological and pharmacological diversity in cancer. Drug response modeling of cell lines drug sensitivity data is able to connect the transcriptional dysregulations of cancer to drug efficacy, which has the potential to translate directly into patient care. It allows for the generation of a virtual screen in patients against hundreds of drugs, greatly simplifying the *in vitro* to *in vivo* translational process. Applying this process to TNBC has led to the computational identification of AZD-1775 and KPT-330, experimental validation of Wee1 inhibition efficacy, and exploration of XPO1 inhibition in TNBC.

In Chapter 2, I overviewed the potential scope of drug response modeling by surveying the cancer cell line datasets available for model generation. In total, there are thousands of potential drug models that can be built, but only a select few datasets have screened a large number of cell lines per compound: GDSC, CTRP, and PRISM. The PRISM drug repurposing dataset, having only been released last year, is not used in this dissertation but provides many additional compounds for future drug response modeling. Additionally, the study design differences (for example the dosing range differences between GDSC and CTRP) have implications for the validity of the drug response models. For some of the drugs in GDSC, the drug sensitivity estimates are mostly if not entirely driven by noise since none of the cell lines respond to the compound at the maximum tested concentration. Finally, there are the implications of cell diversity. Both the majority of cell lines as well as the most commonly used cell lines come from white/Caucasian backgrounds. For several cancers, such as liver cancer, the

majority of the cell lines are male. As we continue to study the role of ethnicity and sex on cancer biology, it becomes clear that these factors can play an important role in drug response. Therefore, the lack of diversity in cell lines has implications not only for the translational ability of drug response models but also for the translation potential of all the experiments and analyses using these cell lines.

Traditional drug discovery is needed to identify novel compounds and targets for the treatment of cancer. But traditional drug development is slow, time-consuming, and prone to failure. These methods are typically evaluating a single cancer target, and historically, we know that most cancer targets turn out not to be effective drug targets. Instead, methods are needed for determining which of the 1000+ cancer therapeutics may be effective in a particular subgroup of cancer patients. In chapter 3, I show that drug response modeling can be a novel method for this form of drug discovery. All the compounds identified from this approach (Table 3.1) had reasonable mechanisms of action for targeting triple-negative breast cancer and literature support. Investigating further, the drug response modeling of AZD-1775 associated with TP53, a known biomarker of response for this drug. Not only was the drug computationally shown to associate with TNBC, I demonstrated that AZD-1775 was able to inhibit the growth of *in vitro* and *in vivo* models of TNBC, further validating the utility of computational drug response modeling in TNBC.

Furthermore, the imputed drug response of AZD-1775 significantly enriched for gene sets that were consistent with the mechanism of action of AZD-1775. This was further explored in Chapter 4 with the XPO1 inhibitors leptomycin B and KPT-330, which have unknown mechanisms of action in breast cancer. Using the same differential imputed response patterns

from Chapter 3, I was able to identify XPO1 inhibition as being particularly effective in TNBC and validate sensitivity of many TNBC cell lines to XPO1 inhibition. When performing a gene set enrichment analysis and imputed drug response association study, the Myc and E2F pathways were significantly associated with XPO1 inhibition imputed response. This was highly consonant with the cMAP analysis and RNA-Seq analysis performed in the MDA-MB-231 TNBC cell line. Not only did this work show the potential of drug response modeling for drug repurposing and drug discovery, but also for mechanism of action identification. Further work is needed to further evaluate this potential of drug response modeling.

FUTURE DIRECTIONS

Many avenues exist for extending this research whether exploring the utility of AZD-1775 and KPT-330 in TNBC, testing the utility of other compounds identified by this approach, further evaluation of drug response modeling, or developing novel applications of drug response modeling. Here, I will outline four specific future research projects that could build upon the findings presented here.

Validating XPO1 Inhibition-mediated Cell Death Mechanisms of Action

The aim of Chapter 4 was to be an introductory exploration of XPO1 inhibition, particularly in the context of the drug response modeling results. My results clearly evinces that p53 and cell cycle signaling are being modulated in MCF-7 and MDA-MB-231 cells after XPO1 inhibition. However, further molecular studies are needed to confirm that the mechanism of action of XPO1 inhibition in MCF-7 cells relies on p53 signaling and that the mechanism of action in MDA-MB-231 cells relies on the inhibition/downregulation of E2F.

In MCF-7 cells, probing the dependence of XPO1 inhibition on p53 signaling should be relatively straightforward. CRISPR/Cas9 can be used to introduce mutations in TP53 of MCF-7

cells followed by the assessment of XPO1 inhibition in these mutated cells. If the cells are no longer sensitive, p53 signaling is needed to get response to XPO1 inhibitors. While many p53 target genes were more highly expressed after XPO1 inhibition, a p53 reporter assay could be used to further validate p53 signaling activation. In this assay, a p53 responsive promoter would control the expression of a luciferase reporter. If luciferase is produced (as monitored by luminescence) after XPO1 inhibition, it can be concluded that p53 is becoming transcriptionally active after XPO1 inhibition. Since XPO1 exports proteins from inside the nucleus to the cytoplasm and p53 is a XPO1 target, changes in subcellular localization of p53 should next be assessed. Assessing nuclear and cytoplasmic fractions via western blot before and after treatment with XPO1 inhibitors could determine if p53 changes subcellular localization. Another experiment for this could be to tag p53 with GFP or perform fluorescence *in situ* hybridization to visually assess whether p53 subcellular localization is affected after XPO1 inhibition. Finally, it is expected that XPO1 inhibition would change the subcellular localization of many proteins and that these other proteins may be involved in XPO1 inhibition-mediated cell death. To assess whether changes to p53 localization exclusively can result in the cell death of MCF-7 cells, an inducible p53 expression system could be introduced such that the p53 induced lacks a nuclear export signal and maintains nuclear localization. It would be important for this inducible system to be tuned such that the amount of p53 produced is biologically consistent with the levels of nuclear p53 seen in MCF-7 cells after XPO1 inhibition. If the nuclear-bound p53 is able to kill MCF-7 alone, then this would prove that XPO1 inhibition-induced p53 signaling is sufficient to explain the mechanism of cell death. If not, then other potential targets need to be evaluated and a CRISPR/Cas9 screen could be used. A CRISPR screen could mutate a number of gene targets and determine which mutation(s) lead to

resistance of XPO1 inhibition and therefore are important in the response to XPO1. This study could also be important simply to determine potential mechanisms of resistance and biomarkers of resistance.

As indicated in the text of Chapter 4, the activation of p53 was seen in MCF-7 cells even though a previously published paper (Arango *et al.*, 2017) indicated that p53 status was not relevant for XPO1 inhibitor response in general. However, nuclear exclusion of p53 is common in breast cancer (Moll, Riou and Levine, 1992), presumably particularly common in the ER+ setting where wildtype p53 is much more common. Arango et al only tested one p53 wildtype ER+ cell line (MCF-7) and this was the only ER+ cell line that responded. Based on the leptomycin B response data provided by CTRP, several other ER+ cell lines responded to XPO1 inhibition, many of these with wildtype p53. Therefore, it is possible that this mechanism is more generalizable to other ER+ breast cancer cell lines and patients. Screening KPT-330 in additional ER+ cell lines could easily determine if ER+ cell lines were sensitive to XPO1 inhibition. Then similar studies as described for MCF-7 could be performed to determine if the response is due to p53 activity. If death in other ER+ cell lines are p53 associated, this could be generalized and the studies mentioned above done in more than one cell line. Further experiments could lead to biomarkers and rationale combinations as discussed later.

For MDA-MB-231, the mechanism of action implicated by the RNA-Seq analysis was inhibition of the G1/S of the cell cycle, likely through the inhibition of E2F signaling. Cell cycle analysis could be performed using flow cytometry or monitored in live cells via fluorescence ubiquitination cell cycle indicators (FUCCI) assay to determine if G1/S cell cycle checkpoint is becoming activated in these cells. To further evaluate E2F signaling and prove that it is

downregulated in MDA-MB-231 cells after KPT-330 treatment, luciferase reporter assay could be used again, but this time the XPO1 treatment should reduce the amount of luciferase produced by an E2F reporter gene. Additionally, the cause of E2F downregulation could be further determined. Western blotting for Rb could show an increase in accumulation of Rb and a decrease in phosphorylation. Western blotting to probe the nuclear and cytoplasmic extract for increasing levels of p27 as well as other CDK inhibitors after inhibition of XPO1 could be performed to determine if these proteins are changing localization. If these results are consistent with restoration of Rb activity, CRISPR could be used to introduce mutations in Rb, which should diminish XPO1 inhibition-mediated cell death if Rb is necessary. Knockdown of E2F may also be used to determine if simply reducing E2F levels similarly decreases viability of MDA-MB-231 cells. Similarly, to what was suggested for MCF-7 cells, a CRISPR screen could be used to see what mutations reduce cell sensitivity to XPO1 inhibition to look for resistance mechanisms and other pathways that may play a critical role in XPO1 inhibition response.

The imputed drug response gene set associations (as well as the cMAP data) indicated that downregulation of the E2F may be a more general mechanism of action for TNBC response. However, this would be important to assess in additional models of TNBC. Similar analyses as described in the previous paragraph should be applied to not only MDA-MB-231, but other TNBC cell lines. Additionally, determining if XPO1 inhibition is able to kill breast cancer cells that are both Rb and p53 mutated would indicate that additional pathways may be involved in the response to XPO1 inhibitors in breast cancer.

The SINE KPT-330 is effective in a panel of breast cancer cell lines at less than 200 nM concentrations. However, since this activity is highly specific to only a subset of cell lines, drug

combination and biomarkers would greatly help when translating the drug into the clinic. Assuming that p53 is playing an important role in XPO1 response in ER+ breast cancers, a number of potential biomarkers come immediately to mind. Of course, p53 mutational status but also cytoplasmic protein levels of p53 could be checked via IHC as potential biomarkers of response for this population. In addition to XPO1 inhibition, use of MDM2 inhibitors like Nutlin-3 may synergize to fully activate the p53 pathway. Additionally, since XPO1 inhibition lead to increases in the pro-apoptotic factors Bax and PUMA, BCL-2 inhibitors may also synergize with XPO1 inhibition. Assuming Rb upregulation and E2F downregulation is facilitating XPO1 inhibition-mediated cell death in the TNBC setting, Rb mutational status could be an important biomarker. Regarding combination therapy, CDK4/6 inhibitors may help further prevent Rb phosphorylation though they also may be redundant with XPO1 inhibition since XPO1 inhibition is able to downregulate E2F signaling sufficiently on its own. E2Fs also transcribe various apoptosis related proteins such as BCL2 and Bid, therefore, once again inhibiting BCL-2 may be helpful to further drive the cells into apoptosis. It should be noted that traditional chemotherapy drugs used in TNBC may be less effective with XPO1 inhibition since the traditional view is that chemotherapies cause DNA damage in replicating cells, but XPO1 induces cell cycle arrest. One final combination could be to combine XPO1 inhibition with senolytics. Senolytics are compounds that can induce senescence or push senescent cancer cells to die, often by activating the NRF2 pathway. Senescent signaling is upregulated in MDA-MB-231 cells after treatment with XPO1 inhibition (Table 4.7), suggesting senolytics could be a rationale combination with XPO1 inhibition.

Evaluating Drug Response Models Based on Biological Meaningfulness of Pathway and IDWAS Analyses

In Chapter 3, I evaluated the performance of the 496 drug response models via in sample cross-validation. In this case, I removed the 69 worst performing models based on spearman correlation between imputed and measured results leaving 427 drug response models that were further assessed. Other metrics could be used; however, the field of drug response modeling has yet to identify a good way to validate models. The assumption regarding cross-validation is that the samples used to build the drug response models are the same as the samples to which the model is being applied. However, that is not true and never will be true in translational models of drug response. Therefore, it can be difficult to assess model fit and look for general trends for successful drug prediction.

For example, tamoxifen does not perform well in cross-validation analysis. Tamoxifen has a spearman correlation of 0.11, which ranks 413th of all 496 drug response models. However, as noted in Chapter 3, tamoxifen imputed drug sensitivity still significantly associates with ER+ breast cancers and appears biologically meaningful. The opposite example can also be found. Olaparib, for example, is a compound currently approved for the treatment of BRCA deficient TNBC tumors, but this compound was not significantly associated with TNBC despite performing better (by cross validation) compared to tamoxifen (a spearman correlation of 0.43, which ranks 160th among all CTRP drugs). The takeaway is that it is hard to discern the difference between a model not performing well and the drug not being effective for a subtype. Additionally, as mentioned in chapter 2, it is very likely that for many of the drug models may be driven by noise since they are not effective at the tested doses, especially drug response models built in the PRISM and GDSC datasets.

Fortunately, bad models are unlikely to obscure meaningful results from well-performing models. If a model is driven by noise, the imputations should be random and uninformative. Thus, so long as the subtypes chosen are biologically distinct (non-random) it is unlikely that the model noise would be significantly associated with the biological phenomenon of interest. While having reliable metrics of drug model quality likely doesn't affect the drug discovery performed in this analysis, there are many interesting questions as to what makes a good model. For example, how many cell lines are needed to screen to build an accurate model of drug response? How many cell lines need to respond to a compound for the modeling to work? What are the best transformations to apply to the data before modeling? How should the gene expression data be filtered and homogenized? Which types of regression models should be used? Are there general rules for which models should be used when the cell lines show continuous or dichotomous (responder/non-responder) response to a drug?

Patient drug response data to evaluate the performance of multiple drug response models is generally lacking. A recent review of drug response modeling was only able to identify 5 patient datasets that had clinical response information and pretreated RNA-Seq information in a clinical trial setting (Schätzle, Esfahani and Schuppert, 2020). That means only 5 drug response models could be thoroughly assessed through these means. However, the analysis performed here was able to recapitulate many expected associations of drug response. I have shown here that tamoxifen associates with ER+ breast cancer and a number of compounds that have been investigated for use in TNBC associate with TNBC. In our IDWAS manuscript, we were also able to identify meaningful markers, such as lapatinib associates with HER2+ disease and nutlin-3 response associates with TP53 mutation status among many others. Therefore, instead of using the models for novel drug discovery, we can look for models that are able to recapitulate

known biological phenomenon. These models then could be assessed and manipulated to answer the questions posed earlier.

For example, to determine the number of cell lines needed to obtain an accurate model of drug response, I could take the training data for models that are known to perform well, such as the lapatinib drug response model, and take random samples of this training dataset. Then I could perform associations between the output of numerous lapatinib drug response models and determine at what sample size does the drug response no longer associate with HER2 copy number. A similar process could be done with many drug response models to get a more global consensus as to the amount of data needed to build biologically meaningful models of drug response. In addition, comparisons between drug response models that have the same mechanism of action could be useful to probe similar questions. For example, there are 22 HDAC inhibitors in the CTRP dataset, but when all these models are used to impute drug sensitivity in a cancer that is known to be responsive to HDAC inhibition only 10 of these inhibitors were significantly associated with response. Comparing the number of cell lines available for each drug response model, the relative activity of each inhibitor in the training data, among other characteristics of the drug response models (such as cross-validation results) could help in identifying rules that help identify well-performing drug response models.

Even further, I presented evidence here that gene-set enrichment analysis performed using imputed drug response as a continuous phenotype variable was able to recapitulate the mechanism of action of drug response. Further testing to determine if this is generalizable trait of drug response models should be caried out. If this is a general phenomenon, then imputed drug response modeling has the potential to be used to determine the unknown mechanism of

action of compounds. This information could also be leveraged to test a model's biological significance, similar to what was described in the preceding paragraph for gene-drug associations. This may enable the validation of drug response models for compounds that don't have well established biomarkers of response to be evaluated as well. This could enable the situation where a chemical probe is screened against a set of cell lines and imputed drug response is able to identify a subset of cancers the compound may be effective against, a potential mechanism of action for the compound, and biomarkers that may affect the compounds efficacy *in vivo*.

In summation, the evidence that drug response models can pick up on biological meaningful associations may enable both identifying mechanism of action's for compounds as well as be a way to further evaluate the parameters that lead to accurate models of drug response. Further investigating modeling accuracy will be needed to continue to push drug response modeling forward.

Performing Imputation-Based Drug Discovery in Other Cancer Types

This dissertation was focused on breast cancers, specifically triple-negative breast cancers, but as alluded to in the discussion for Chapter 3, this methodology could easily be applied to other cancer types. The drug response models are built using pan-cancer data and there is no reason that prevents their application to other cancer types. So long as there are enough patients with gene expression data and relevant subtypes that can be compared, similar imputation-based drug discovery could be performed to identify compounds predicted to be efficacious in a cancer population of interest. Three cancer settings come readily to mind: TNBC subtypes, small cell lung cancer, and castration-resistant prostate cancer.

TNBC is a heterogenous disease and additional studies have tried to further subtype TNBC into 3-4 subgroups. One of these subtypes in particular, LAR or luminal androgen receptor positive TNBC has particularly poor 5-year overall survival and tends to be quite distinct from the other TNBC tumor types. While my analysis looked at TNBC in general, the TNBC subtypes, being transcriptionally defined, could make for an interesting comparison. Indeed, an early analysis that I performed indicated that almost all the compounds identified in my analysis (Table 3.1) were predicted to be the least effective in the LAR subtype. This analysis was limited, there are only about 130 TNBC breast cancer patients with between 29-60 patients per TNBC subtype. Use of additional datasets that are TNBC focused may increase the sample size and perhaps additional compounds that target the unresponsive LAR subpopulation could be identified. Additionally, LAR and TNBC subtypes are traditionally not defined clinically and the number of LAR models are limited to test any potential hits, which may make translation of the results more difficult. However, increases in model diversity as well as our ability to link imputed drug sensitivity directly to markers of drug response may enable the validation and translation of these findings.

Small Cell Lung Cancer (SCLC) is analogous to TNBC in many ways. SCLC comprises 10-15% of lung cancer patients and the patients that present with SCLC generally have poor outcomes. SCLC cancers are marked by high proliferation and propensity for metastatic spread and most SCLCs are treated with cytotoxic chemotherapies (Bernhardt and Jalal, 2016; Rudin *et al.*, 2021). With so many similarities to TNBC, it is easy to imagine performing a similar analysis in SCLC. Drug response models could be built and then applied to a patient dataset containing both non-small cell lung cancer (NSCLC) and SCLC tumor gene expression data. Then similar statistical tests could be applied to determine which drugs are predicted to be particularly

effective in SCLC. SCLC is also highly associated with smoking and another analysis could be carried out where the imputed drug sensitivity is compared within a set of SCLC patients between smokers and non-smokers. These analyses may be helpful in less toxic and more effective agents for the treatment of SCLC.

Since the drug response modeling is built on pan-cancer cell line data, we have often speculated that drug response modeling could be particularly helpful for cancer settings that lack readily available models for the evaluation of compounds. Prostate cancer, which as seen in chapter 2, has a relatively few available cell lines. In prostate cancer, many patients respond well to androgen deprivation therapy (as indicated by the quote that began this dissertation). However, it is possible for prostate cancer to not respond to or recur from androgen deprivation. This is commonly referred to as castration-resistant prostate cancer (CRPC). While not a traditional cancer subtype, there are well established biological differences between CRPC and prostate cancer. Using drug response modeling, we may be able to associate these biological differences directly with treatment response and identify compounds that work in the CRPC setting.

These are just 3 examples of opportunities where applying our drug response modeling-based drug discovery approach may lead to the identification of clinically viable treatments. One of the advantages of our method is that it is highly flexible, and any two patient subgroups could be used. The subgroups could be defined clinically, defined transcriptionally, or defined by single mutation or mutational signatures. So long as there are expected transcriptional differences between the two groups, drug response modeling could be used to identify targetable vulnerabilities in one patient subgroup verse the other.

Imputing Drug Response in Subpopulation of Cells Using Single Cell RNA-Seq data

On the same topic of transcriptional differences, a recent advance in molecular biology has been the creation of single-cell RNA-seq (scRNA-seq) technologies that isolate and evaluate the transcriptome of individual cells. While this is a great tool to explore intratumoral heterogeneity, directly translating findings from scRNA-Seq experiments into drug response is not straightforward. In theory, scDNA-Seq or scRNA-Seq could be used to identify subclonal populations that are expected to be resistant to therapy. Mutations, for instance, may be present subclonally that would render an inhibitor ineffective in a subclonal population that can then repopulate a tumor after treatment. For example, scRNA-Seq has been used to show therapy-induced evolution in lung cancer patients (Maynard *et al.*, 2020).

My hypothesis is that, similar to identifying druggable vulnerabilities based on gene expression differences in patients, drug response modeling has the potential to translate subclonal molecular differences directly into drug response predictions. There are several important technical considerations that would first need to be considered to allow for translational models bult on bulk RNA-Seq to be used for scRNA-Seq. scRNA-Seq data is generally on a different scale and depth as RNA-Seq information and often contains many dropout genes (i.e. genes with no expression information). We have developed ways to translate drug response models built with microarray gene expression data to RNA-Seq, and similar scaling approaches likely could be used to make the data comparable. To address dropout genes, we would have to ensure that the models will work when they are fit to with fewer features (i.e. genes). This could potentially be tested using the validation approaches outlined in the previous future direction. The major challenge of drug response modeling on scRNA-Seq would be that there are no standards to compare with. Unlike the analyses in Chapter 3, there would likely be

no proof-of-concept or expected results. Additionally, scRNA-Seq necessarily destroys the cells, so even though we know the subclonal structure of the population the population no longer exists for us to probe with drugs. However, models that track heterogeneity could be used if the subclonal populations are consistent. For example, molecular barcoding could be done in a tumor sample that is then grown in organoid culture. Part of the culture could be sent to scRNA-Seq and drug could be added to the other culture. The molecular barcodes could be sequenced to determine which subclonal populations are affected by the treatment and if they match our predictions. While this may indicate our ability to predict drug response in subclonal populations, definitive proof that a drug predicted to target a particular subclonal population killed specifically that population is impractical if not impossible. This is because this procedure would have to be performed for multiple drugs to be able to generalize the results since some predictions from drug models would be expected to fail. Additionally, finding drug prediction scenarios where one subclone is expected to be sensitive and the others are completely resistant may be difficult. Finally, drug response variations (since the drug has the potential of affecting multiple subclonal populations) and the population dynamics among the subclones may be difficult to isolate the effect to the treatment itself.

There are many hurdles and difficulties proving the modeling in scRNA-Seq data would be effective, but there is also great potential for the rational identification of drug combinations. One major resistance mechanism in cancer is the presence of resistant subclones that is able to repopulate a tumor after treatment (Zhao, Hemann and Lauffenburger, 2014; Lim and Ma, 2019; Ramón y Cajal *et al.*, 2020). Being able to link tumor subclonal populations directly with a targeted therapy would enable the selection of drugs that work on different subclonal populations. Combining these therapies then has the potential of directly circumventing this

issue of intratumoral heterogeneity and resistant subclonal populations. Additionally, while proving drug response modeling is accurate in the single cell setting for many drugs may be difficult, testing drug combinations in a heterogenous tumor model is straightforward and has the most translational impact. Given the potential impact, I am particularly excited for this future direction and look forward to working on it for some time.

Conclusion

This is an exciting time for modeling drug response. Computational resources continue to become available that make forming new and increasingly complex models of drug response models more and more efficient. Additionally, as reviewed in Chapter 2, the amount of data needed for building models of drug response continues to increase. It has been a privilege to contribute to the field and show that the utility of drug response models is also capable of being expanded. Drug response modeling can be used successfully for drug discovery, biomarker identification, and perhaps even mechanism of action identification. These methods were able to successfully identify a Wee1 and XPO1 inhibitor for the treatment of TNBC, which (with further analysis) may lead to the more efficient treatment of breast cancer patients. The approach I utilized is flexible and can be adopted for other cancer contexts and may even be able to facilitate the identification of rationale drug combinations. The employment of these computational tools is quick, efficient, and allows for contextualizing cell line response in any patient population. I hope that cancer biologists and oncologists are able to make use of this framework for drug discovery and that this ultimately enables the creation of new and effective treatment options for patients.

6. REFERENCES

Abubakar, M. *et al.* (2019) 'Clinicopathological and epidemiological significance of breast cancer subtype reclassification based on p53 immunohistochemical expression', *npj Breast Cancer*, 5(1), pp. 1–9. doi: 10.1038/s41523-019-0117-7.

Aguirre, A. J. *et al.* (2016) 'Genomic Copy Number Dictates a Gene-Independent Cell Response to CRISPR/Cas9 Targeting', *Cancer Discovery*, 6(8), pp. 914–929. doi: 10.1158/2159-8290.CD-16-0154.

Albiges, L. *et al.* (2014) 'Chk1 as a new therapeutic target in triple-negative breast cancer', *Breast (Edinburgh, Scotland)*, 23(3), pp. 250–258. doi: 10.1016/j.breast.2014.02.004.

Arango, N. P. *et al.* (2017) 'Selinexor (KPT-330) demonstrates anti-tumor efficacy in preclinical models of triple-negative breast cancer', *Breast cancer research: BCR*, 19(1), p. 93. doi: 10.1186/s13058-017-0878-6.

Ashburn, T. T. and Thor, K. B. (2004) 'Drug repositioning: identifying and developing new uses for existing drugs', *Nature Reviews Drug Discovery*, 3(8), pp. 673–683. doi: 10.1038/nrd1468.

Aysola, K. *et al.* (2013) 'Triple Negative Breast Cancer – An Overview', *Hereditary genetics : current research*, 2013(Suppl 2). doi: 10.4172/2161-1041.S2-001.

Azmi, A. S., Aboukameel, A., *et al.* (2013) 'Selective inhibitors of nuclear export block pancreatic cancer cell proliferation and reduce tumor growth in mice', *Gastroenterology*, 144(2), pp. 447–456. doi: 10.1053/j.gastro.2012.10.036.

Azmi, A. S., Al-Katib, A., *et al.* (2013) 'Selective inhibitors of nuclear export for the treatment of non-Hodgkin's lymphomas', *Haematologica*, 98(7), pp. 1098–1106. doi: 10.3324/haematol.2012.074781.

Azuaje, F. (2017) 'Computational models for predicting drug responses in cancer research', *Briefings in Bioinformatics*, 18(5), pp. 820–829. doi: 10.1093/bib/bbw065.

Baek, H. B. *et al.* (2018) 'XPO1 inhibition by selinexor induces potent cytotoxicity against high grade bladder malignancies', *Oncotarget*, 9(77), pp. 34567–34581. doi: 10.18632/oncotarget.26179.

Balafoutas, D. *et al.* (2013) 'Cancer testis antigens and NY-BR-1 expression in primary breast cancer: prognostic and therapeutic implications', *BMC cancer*, 13, p. 271. doi: 10.1186/1471-2407-13-271.

Baller, J. *et al.* (2019) 'CHURP: A Lightweight CLI Framework to Enable Novice Users to Analyze Sequencing Datasets in Parallel', in *Proceedings of the Practice and Experience in Advanced Research Computing on Rise of the Machines (learning)*. New York, NY, USA: Association for Computing Machinery (PEARC '19), pp. 1–5. doi: 10.1145/3332186.3333156.

Barretina, J. *et al.* (2012) 'The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity', *Nature*, 483(7391), pp. 603–607. doi: 10.1038/nature11003.

Barrett, T. *et al.* (2012) 'BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata', *Nucleic Acids Research*, 40(Database issue), pp. D57–D63. doi: 10.1093/nar/gkr1163.

Ben-David, U. *et al.* (2018) 'Genetic and transcriptional evolution alters cancer cell line drug response', *Nature*, 560(7718), pp. 325–330. doi: 10.1038/s41586-018-0409-3.

Berger, C., Qian, Y. and Chen, X. (2013) 'The p53-Estrogen Receptor Loop in Cancer', *Current molecular medicine*, 13(8), pp. 1229–1240. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3780397/ (Accessed: 12 April 2021).

Bernhardt, E. B. and Jalal, S. I. (2016) 'Small Cell Lung Cancer', *Cancer Treatment and Research*, 170, pp. 301–322. doi: 10.1007/978-3-319-40389-2_14.

Boehm, J. S. and Golub, T. R. (2015) 'An ecosystem of cancer cell line factories to support a cancer dependency map', *Nature Reviews Genetics*, 16(7), pp. 373–374. doi: 10.1038/nrg3967.

Bouhaddou, M. *et al.* (2016) 'Drug response consistency in CCLE and CGP', *Nature*, 540(7631), pp. E9–E10. doi: 10.1038/nature20580.

Broad Institute Cancer Cell Line Encyclopedia (CCLE) (no date). Available at: https://portals.broadinstitute.org/ccle/data (Accessed: 9 March 2021).

Bryant, C., Rawlinson, R. and Massey, A. J. (2014) 'Chk1 inhibition as a novel therapeutic strategy for treating triple-negative breast and ovarian cancers', *BMC cancer*, 14, p. 570. doi: 10.1186/1471-2407-14-570.

Cancer feature: TP53_mut - Cancerrxgene - Genomics of Drug Sensitivity in Cancer (no date). Available at: https://www.cancerrxgene.org/feature/TP53_mut/289/volcano (Accessed: 9 March 2021).

Cancer Genome Atlas Network (2012) 'Comprehensive molecular portraits of human breast tumours', *Nature*, 490(7418), pp. 61–70. doi: 10.1038/nature11412.

Cancer Genome Atlas Research Network *et al.* (2013) 'The Cancer Genome Atlas Pan-Cancer analysis project', *Nature Genetics*, 45(10), pp. 1113–1120. doi: 10.1038/ng.2764.

Chen, S. *et al.* (2020) 'SPTBN1 and cancer, which links?', *Journal of Cellular Physiology*, 235(1), pp. 17–25. doi: 10.1002/jcp.28975.

Chen, X. *et al.* (2018) 'Cyclin E Overexpression Sensitizes Triple-Negative Breast Cancer to Wee1 Kinase Inhibition', *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research*, 24(24), pp. 6594–6610. doi: 10.1158/1078-0432.CCR-18-1446.

Cheng, Y. *et al.* (2014) 'XPO1 (CRM1) inhibition represses STAT3 activation to drive a survivin-dependent oncogenic switch in triple-negative breast cancer', *Molecular Cancer Therapeutics*, 13(3), pp. 675–686. doi: 10.1158/1535-7163.MCT-13-0416.

Cheung, H. W. *et al.* (2011) 'Systematic investigation of genetic vulnerabilities across cancer cell lines reveals lineage-specific dependencies in ovarian cancer', *Proceedings of the National Academy of Sciences*, 108(30), pp. 12372–12377. doi: 10.1073/pnas.1109363108.

Clausse, V. *et al.* (2016) 'Wee1 inhibition potentiates Wip1-dependent p53-negative tumor cell death during chemotherapy', *Cell Death & Disease*, 7, p. e2195. doi: 10.1038/cddis.2016.96.

Colaprico, A. *et al.* (2016) 'TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data', *Nucleic Acids Research*, 44(8), p. e71. doi: 10.1093/nar/gkv1507.

Cole, K. A. *et al.* (2020) 'Phase I Clinical Trial of the Wee1 Inhibitor Adavosertib (AZD1775) with Irinotecan in Children with Relapsed Solid Tumors: A COG Phase I Consortium Report (ADVL1312)', *Clinical Cancer*

Research: An Official Journal of the American Association for Cancer Research, 26(6), pp. 1213–1219. doi: 10.1158/1078-0432.CCR-19-3470.

Corno, C. *et al.* (2018) 'FoxO-1 contributes to the efficacy of the combination of the XPO1 inhibitor selinexor and cisplatin in ovarian carcinoma preclinical models', *Biochemical Pharmacology*, 147, pp. 93–103. doi: 10.1016/j.bcp.2017.11.009.

Corsello, S. M. *et al.* (2017) 'The Drug Repurposing Hub: a next-generation drug library and information resource', *Nature Medicine*, 23(4), pp. 405–408. doi: 10.1038/nm.4306.

Corsello, S. M. *et al.* (2020) 'Discovering the anti-cancer potential of non-oncology drugs by systematic viability profiling', *Nature Cancer*, 1(2), pp. 235–248. doi: 10.1038/s43018-019-0018-6.

Costa, R. L. B. and Gradishar, W. J. (2017) 'Differences Are Important: Breast Cancer Therapy in Different Ethnic Groups', *Journal of Global Oncology*, 3(4), pp. 281–284. doi: 10.1200/JGO.2017.009936.

Cuneo, K. C. *et al.* (2016) 'Wee1 Kinase Inhibitor AZD1775 Radiosensitizes Hepatocellular Carcinoma Regardless of TP53 Mutational Status Through Induction of Replication Stress', *International Journal of Radiation Oncology, Biology, Physics*, 95(2), pp. 782–790. doi: 10.1016/j.ijrobp.2016.01.028.

Daemen, A. *et al.* (2013) 'Modeling precision treatment of breast cancer', *Genome Biology*, 14, p. R110. doi: 10.1186/gb-2013-14-10-r110.

DeSantis, C. E. *et al.* (2019) 'Breast cancer statistics, 2019', *CA: A Cancer Journal for Clinicians*, 69(6), pp. 438–451. doi: https://doi.org/10.3322/caac.21583.

Diab, A. *et al.* (2019) 'Multiple Defects Sensitize p53-Deficient Head and Neck Cancer Cells to the WEE1 Kinase Inhibition', *Molecular cancer research: MCR*, 17(5), pp. 1115–1128. doi: 10.1158/1541-7786.MCR-18-0860.

Do, K. *et al.* (2015) 'Phase I Study of Single-Agent AZD1775 (MK-1775), a Wee1 Kinase Inhibitor, in Patients With Refractory Solid Tumors', *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology*, 33(30), pp. 3409–3415. doi: 10.1200/JCO.2014.60.4009.

Drug Approval Package: XPOVIO (2019). Available at: https://www.accessdata.fda.gov/drugsatfda_docs/nda/2019/212306Orig1s000TOC.cfm (Accessed: 8 April 2021).

Drug Download Page - Cancerrxgene - Genomics of Drug Sensitivity in Cancer (no date). Available at: https://www.cancerrxgene.org/downloads/bulk_download (Accessed: 9 March 2021).

Dutil, J. *et al.* (2019) 'An Interactive Resource to Probe Genetic Diversity and Estimated Ancestry in Cancer Cell Lines', *Cancer research*, 79(7), pp. 1263–1273. doi: 10.1158/0008-5472.CAN-18-2747.

Fabregat, A. *et al.* (2018) 'The Reactome Pathway Knowledgebase', *Nucleic Acids Research*, 46(Database issue), pp. D649–D655. doi: 10.1093/nar/gkx1132.

Forbes, S. A. *et al.* (2017) 'COSMIC: somatic cancer genetics at high-resolution', *Nucleic Acids Research*, 45(D1), pp. D777–D783. doi: 10.1093/nar/gkw1121.

Foulkes, W. D., Smith, I. E. and Reis-Filho, J. S. (2010) 'Triple-negative breast cancer', *The New England Journal of Medicine*, 363(20), pp. 1938–1948. doi: 10.1056/NEJMra1001389.

Futreal, P. A. *et al.* (2004) 'A CENSUS OF HUMAN CANCER GENES', *Nature reviews. Cancer*, 4(3), pp. 177–183. doi: 10.1038/nrc1299.

Gandhi, U. H. *et al.* (2018) 'Clinical Implications of Targeting XPO1-mediated Nuclear Export in Multiple Myeloma', *Clinical Lymphoma, Myeloma & Leukemia*, 18(5), pp. 335–345. doi: 10.1016/j.clml.2018.03.003.

Gao, H. *et al.* (2015) 'High-throughput screening using patient-derived tumor xenografts to predict clinical trial drug response', *Nature Medicine*, 21(11), pp. 1318–1325. doi: 10.1038/nm.3954.

Geeleher, P. *et al.* (2015) 'Predicting Response to Histone Deacetylase Inhibitors Using High-Throughput Genomics', *Journal of the National Cancer Institute*, 107(11). doi: 10.1093/jnci/djv247.

Geeleher, P. *et al.* (2016) 'Consistency in large pharmacogenomic studies', *Nature*, 540(7631), pp. E1–E2. doi: 10.1038/nature19838.

Geeleher, P. *et al.* (2017) 'Discovering novel pharmacogenomic biomarkers by imputing drug response in cancer patients from large genomics studies', *Genome Research*, 27(10), pp. 1743–1751. doi: 10.1101/gr.221077.117.

Geeleher, P., Cox, N. J. and Huang, R. S. (2014) 'Clinical drug response can be predicted using baseline gene expression levels and in vitrodrug sensitivity in cell lines', *Genome Biology*, 15, p. R47. doi: 10.1186/gb-2014-15-3-r47.

Geenen, J. J. J. and Schellens, J. H. M. (2017) 'Molecular Pathways: Targeting the Protein Kinase Wee1 in Cancer', *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research*, 23(16), pp. 4540–4544. doi: 10.1158/1078-0432.CCR-17-0520.

Giuli, M. V. *et al.* (2019) 'Notch Signaling Activation as a Hallmark for Triple-Negative Breast Cancer Subtype', *Journal of Oncology*, 2019, p. 8707053. doi: 10.1155/2019/8707053.

Gruener, R. F. *et al.* (2021) 'Facilitating Drug Discovery in Breast Cancer by Virtually Screening Patients Using In Vitro Drug Response Modeling', *Cancers*, 13(4), p. 885. doi: 10.3390/cancers13040885.

Haibe-Kains, B. *et al.* (2013) 'Inconsistency in large pharmacogenomic studies', *Nature*, 504(7480), pp. 389–393. doi: 10.1038/nature12831.

Haverty, P. M. *et al.* (2016) 'Reproducible pharmacogenomic profiling of cancer cell line panels', *Nature*, 533(7603), pp. 333–337. doi: 10.1038/nature17987.

Hirai, H. *et al.* (2009) 'Small-molecule inhibition of Wee1 kinase by MK-1775 selectively sensitizes p53-deficient tumor cells to DNA-damaging agents', *Molecular Cancer Therapeutics*, 8(11), pp. 2992–3000. doi: 10.1158/1535-7163.MCT-09-0463.

Hoadley, K. A. *et al.* (2014) 'Multiplatform Analysis of 12 Cancer Types Reveals Molecular Classification within and across Tissues of Origin', *Cell*, 158(4), pp. 929–944. doi: 10.1016/j.cell.2014.06.049.

Howlader, N. *et al.* (2018) 'Differences in Breast Cancer Survival by Molecular Subtypes in the United States', *Cancer Epidemiology, Biomarkers & Prevention: A Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology*, 27(6), pp. 619–626. doi: 10.1158/1055-9965.EPI-17-0627.

https://ocg.cancer.gov/programs/ctd2/data-portal (no date) *Office of Cancer Genomics*. Available at: https://ocg.cancer.gov/programs/ctd2/data-portal (Accessed: 9 March 2021).

Huang, L. *et al.* (2020) 'PDX-derived organoids model in vivo drug response and secrete biomarkers', *JCI Insight*, 5(21). doi: 10.1172/jci.insight.135544.

Huggins, C. (1979) *Experimental Leukemia and Mammary Cancer: Induction, Prevention, Cure*. University of Chicago Press.

Hurle, M. R. *et al.* (2013) 'Computational drug repositioning: from data to therapeutics', *Clinical Pharmacology and Therapeutics*, 93(4), pp. 335–341. doi: 10.1038/clpt.2013.1.

Hussain, M. R. M., Hoessli, D. C. and Fang, M. (2016) 'N-acetylgalactosaminyltransferases in cancer', *Oncotarget*, 7(33), pp. 54067–54081. doi: 10.18632/oncotarget.10042.

Hwang, S.-Y., Park, S. and Kwon, Y. (2019) 'Recent therapeutic trends and promising targets in triple negative breast cancer', *Pharmacology & Therapeutics*, 199, pp. 30–57. doi: 10.1016/j.pharmthera.2019.02.006.

Jiang, Y.-Z. *et al.* (2019) 'Genomic and Transcriptomic Landscape of Triple-Negative Breast Cancers: Subtypes and Treatment Strategies', *Cancer Cell*, 35(3), pp. 428-440.e5. doi: 10.1016/j.ccell.2019.02.001.

Jin, J. *et al.* (2018) 'Combined Inhibition of ATR and WEE1 as a Novel Therapeutic Strategy in Triple-Negative Breast Cancer', *Neoplasia (New York, N.Y.)*, 20(5), pp. 478–488. doi: 10.1016/j.neo.2018.03.003.

Johnson, W. E., Li, C. and Rabinovic, A. (2007) 'Adjusting batch effects in microarray expression data using empirical Bayes methods', *Biostatistics (Oxford, England)*, 8(1), pp. 118–127. doi: 10.1093/biostatistics/kxj037.

Kanehisa, M. *et al.* (2017) 'KEGG: new perspectives on genomes, pathways, diseases and drugs', *Nucleic Acids Research*, 45(D1), pp. D353–D361. doi: 10.1093/nar/gkw1092.

Keenan, T. *et al.* (2015) 'Comparison of the Genomic Landscape Between Primary Breast Cancer in African American Versus White Women and the Association of Racial Differences With Tumor Recurrence', *Journal of Clinical Oncology*, 33(31), pp. 3621–3627. doi: 10.1200/JCO.2015.62.2126.

Lapalombella, R. *et al.* (2012) 'Selective inhibitors of nuclear export show that CRM1/XPO1 is a target in chronic lymphocytic leukemia', *Blood*, 120(23), pp. 4621–4634. doi: 10.1182/blood-2012-05-429506.

Leijen, S., van Geel, R. M. J. M., Pavlick, A. C., *et al.* (2016) 'Phase I Study Evaluating WEE1 Inhibitor AZD1775 As Monotherapy and in Combination With Gemcitabine, Cisplatin, or Carboplatin in Patients With Advanced Solid Tumors', *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology*, 34(36), pp. 4371–4380. doi: 10.1200/JCO.2016.67.5991.

Leijen, S., van Geel, R. M. J. M., Sonke, G. S., *et al.* (2016) 'Phase II Study of WEE1 Inhibitor AZD1775 Plus Carboplatin in Patients With TP53-Mutated Ovarian Cancer Refractory or Resistant to First-Line Therapy Within 3 Months', *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology*, 34(36), pp. 4354–4361. doi: 10.1200/JCO.2016.67.5942.

Lewis, C. W. *et al.* (2017) 'Prolonged mitotic arrest induced by Wee1 inhibition sensitizes breast cancer cells to paclitaxel', *Oncotarget*, 8(43), pp. 73705–73722. doi: 10.18632/oncotarget.17848.

Li, H. *et al.* (2015) 'Integrated Analysis of Transcriptome in Cancer Patient-Derived Xenografts', *PLoS ONE*, 10(5). doi: 10.1371/journal.pone.0124780.

Li, J. *et al.* (2019) 'Association of p53 expression with poor prognosis in patients with triple-negative breast invasive ductal carcinoma', *Medicine*, 98(18), p. e15449. doi: 10.1097/MD.0000000000015449.

Li, X. *et al.* (2017) 'Triple-negative breast cancer has worse overall survival and cause-specific survival than non-triple-negative breast cancer', *Breast Cancer Research and Treatment*, 161(2), pp. 279–287. doi: 10.1007/s10549-016-4059-6.

Liberzon, A. *et al.* (2015) 'The Molecular Signatures Database Hallmark Gene Set Collection', *Cell Systems*, 1(6), pp. 417–425. doi: 10.1016/j.cels.2015.12.004.

Lim, Z.-F. and Ma, P. C. (2019) 'Emerging insights of tumor heterogeneity and drug resistance mechanisms in lung cancer targeted therapy', *Journal of Hematology & Oncology*, 12. doi: 10.1186/s13045-019-0818-2.

Ling, A. *et al.* (2018) 'More than fishing for a cure: The promises and pitfalls of high throughput cancer cell line screens', *Pharmacology & Therapeutics*, 191, pp. 178–189. doi: 10.1016/j.pharmthera.2018.06.014.

Lok, S. W. *et al.* (2019) 'A Phase Ib Dose-Escalation and Expansion Study of the BCL2 Inhibitor Venetoclax Combined with Tamoxifen in ER and BCL2-Positive Metastatic Breast Cancer', *Cancer Discovery*, 9(3), pp. 354–369. doi: 10.1158/2159-8290.CD-18-1151.

Love, M. I., Huber, W. and Anders, S. (2014) 'Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2', *Genome Biology*, 15(12), p. 550. doi: 10.1186/s13059-014-0550-8.

Lu, W. *et al.* (2000) 'Nuclear exclusion of p53 in a subset of tumors requires MDM2 function', *Oncogene*, 19(2), pp. 232–240. doi: 10.1038/sj.onc.1203262.

Luo, W. and Brouwer, C. (2013) 'Pathview: an R/Bioconductor package for pathway-based data integration and visualization', *Bioinformatics (Oxford, England)*, 29(14), pp. 1830–1831. doi: 10.1093/bioinformatics/btt285.

Ma, C. X. *et al.* (2012) 'Targeting Chk1 in p53-deficient triple-negative breast cancer is therapeutically beneficial in human-in-mouse tumor models', *The Journal of Clinical Investigation*, 122(4), pp. 1541–1552. doi: 10.1172/JCI58765.

Maynard, A. *et al.* (2020) 'Therapy-Induced Evolution of Human Lung Cancer Revealed by Single-Cell RNA Sequencing', *Cell*, 182(5), pp. 1232-1251.e22. doi: 10.1016/j.cell.2020.07.017.

Méndez, E. *et al.* (2018) 'A Phase I Clinical Trial of AZD1775 in Combination with Neoadjuvant Weekly Docetaxel and Cisplatin before Definitive Therapy in Head and Neck Squamous Cell Carcinoma', *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research*, 24(12), pp. 2740–2748. doi: 10.1158/1078-0432.CCR-17-3796.

Milacic, M. *et al.* (2012) 'Annotating Cancer Variants and Anti-Cancer Therapeutics in Reactome', *Cancers*, 4(4), pp. 1180–1211. doi: 10.3390/cancers4041180.

Mizuarai, S. *et al.* (2009) 'Discovery of gene expression-based pharmacodynamic biomarker for a p53 context-specific anti-tumor drug Wee1 inhibitor', *Molecular Cancer*, 8, p. 34. doi: 10.1186/1476-4598-8-34.

Moll, U. M., Riou, G. and Levine, A. J. (1992) 'Two distinct mechanisms alter p53 in breast cancer: mutation and nuclear exclusion', *Proceedings of the National Academy of Sciences of the United States of America*, 89(15), pp. 7262–7266. doi: 10.1073/pnas.89.15.7262.

Morris, E. J. *et al.* (2006) 'Functional Identification of Api5 as a Suppressor of E2F-Dependent Apoptosis In Vivo', *PLoS Genetics*, 2(11). doi: 10.1371/journal.pgen.0020196.

Mpindi, J. P. *et al.* (2016) 'Consistency in drug response profiling', *Nature*, 540(7631), pp. E5–E6. doi: 10.1038/nature20171.

MSigDB gene sets R package (no date). Available at: https://igordot.github.io/msigdbr/index.html (Accessed: 27 March 2021).

Newman, L. A. *et al.* (2015) 'The 2014 Society of Surgical Oncology Susan G. Komen for the Cure Symposium: Triple-Negative Breast Cancer', *Annals of Surgical Oncology*, 22(3), pp. 874–882. doi: 10.1245/s10434-014-4279-0.

News - Cancerrxgene - Genomics of Drug Sensitivity in Cancer (2021). Available at: https://www.cancerrxgene.org/news (Accessed: 2 April 2021).

Pappano, W. N. *et al.* (2014) 'Genetic inhibition of the atypical kinase Wee1 selectively drives apoptosis of p53 inactive tumor cells', *BMC cancer*, 14, p. 430. doi: 10.1186/1471-2407-14-430.

Parker, L. L. and Piwnica-Worms, H. (1992) 'Inactivation of the p34cdc2-cyclin B complex by the human WEE1 tyrosine kinase', *Science (New York, N.Y.)*, 257(5078), pp. 1955–1957. doi: 10.1126/science.1384126.

Pozdeyev, N. *et al.* (2016) 'Integrating heterogeneous drug sensitivity data from cancer pharmacogenomic studies', *Oncotarget*, 7(32), pp. 51619–51625. doi: 10.18632/oncotarget.10010.

Prat, A. *et al.* (2013) 'Molecular characterization of basal-like and non-basal-like triple-negative breast cancer', *The Oncologist*, 18(2), pp. 123–133. doi: 10.1634/theoncologist.2012-0397.

Pushpakom, S. *et al.* (2019) 'Drug repurposing: progress, challenges and recommendations', *Nature Reviews Drug Discovery*, 18(1), pp. 41–58. doi: 10.1038/nrd.2018.168.

Quevedo, R. *et al.* (2020) 'Assessment of Genetic Drift in Large Pharmacogenomic Studies', *Cell Systems*, p. S2405471220303239. doi: 10.1016/j.cels.2020.08.012.

Ramón y Cajal, S. *et al.* (2020) 'Clinical implications of intratumor heterogeneity: challenges and opportunities', *Journal of Molecular Medicine (Berlin, Germany)*, 98(2), pp. 161–177. doi: 10.1007/s00109-020-01874-2.

Ranganathan, P. *et al.* (2012) 'Preclinical activity of a novel CRM1 inhibitor in acute myeloid leukemia', *Blood*, 120(9), pp. 1765–1773. doi: 10.1182/blood-2012-04-423160.

Rubin, J. B. *et al.* (2020) 'Sex differences in cancer mechanisms', *Biology of Sex Differences*, 11. doi: 10.1186/s13293-020-00291-x.

Rudin, C. M. *et al.* (2021) 'Small-cell lung cancer', *Nature Reviews Disease Primers*, 7(1), pp. 1–20. doi: 10.1038/s41572-020-00235-0.

Sanai, N. *et al.* (2018) 'Phase 0 Trial of AZD1775 in First-Recurrence Glioblastoma Patients', *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research*, 24(16), pp. 3820–3828. doi: 10.1158/1078-0432.CCR-17-3348.

Schätzle, L.-K., Esfahani, A. H. and Schuppert, A. (2020) 'Methodological challenges in translational drug response modeling in cancer: A systematic analysis with FORESEE', *PLOS Computational Biology*, 16(4), p. e1007803. doi: 10.1371/journal.pcbi.1007803.

Scherf, U. *et al.* (2000) 'A gene expression database for the molecular pharmacology of cancer', *Nature Genetics*, 24(3), pp. 236–244. doi: 10.1038/73439.

Seashore-Ludlow, B. *et al.* (2015) 'Harnessing Connectivity in a Large-Scale Small-Molecule Sensitivity Dataset', *Cancer Discovery*, 5(11), pp. 1210–1223. doi: 10.1158/2159-8290.CD-15-0235.

Segal, E. *et al.* (2004) 'A module map showing conditional activity of expression modules in cancer', *Nature Genetics*, 36(10), pp. 1090–1098. doi: 10.1038/ng1434.

Sekine, I. *et al.* (2008) 'Emerging ethnic differences in lung cancer therapy', *British Journal of Cancer*, 99(11), pp. 1757–1762. doi: 10.1038/sj.bjc.6604721.

Serrano-Gómez, S. J., Fejerman, L. and Zabaleta, J. (2018) 'Breast Cancer in Latinas: A Focus on Intrinsic Subtypes Distribution', *Cancer Epidemiology and Prevention Biomarkers*, 27(1), pp. 3–10. doi: 10.1158/1055-9965.EPI-17-0420.

Shafique, M. *et al.* (2019) 'A Phase II Trial of Selinexor (KPT-330) for Metastatic Triple-Negative Breast Cancer', *The Oncologist*, 24(7), pp. 887-e416. doi: 10.1634/theoncologist.2019-0231.

Shoemaker, R. H. (2006) 'The NCI60 human tumour cell line anticancer drug screen', *Nature Reviews Cancer*, 6(10), pp. 813–823. doi: 10.1038/nrc1951.

Siddharth, S. and Sharma, D. (2018) 'Racial Disparity and Triple-Negative Breast Cancer in African-American Women: A Multifaceted Affair between Obesity, Biology, and Socioeconomic Determinants', *Cancers*, 10(12). doi: 10.3390/cancers10120514.

Siegel, R. L. *et al.* (2021) 'Cancer Statistics, 2021', *CA: A Cancer Journal for Clinicians*, 71(1), pp. 7–33. doi: https://doi.org/10.3322/caac.21654.

Siegel, R. L., Miller, K. D. and Jemal, A. (2017) 'Cancer statistics, 2017', *CA: A Cancer Journal for Clinicians*, 67(1), pp. 7–30. doi: 10.3322/caac.21387.

Smirnov, P. *et al.* (2018) 'PharmacoDB: an integrative database for mining in vitro anticancer drug screening studies', *Nucleic Acids Research*, 46(D1), pp. D994–D1002. doi: 10.1093/nar/gkx911.

Stommel, J. M. *et al.* (1999) 'A leucine-rich nuclear export signal in the p53 tetramerization domain: regulation of subcellular localization and p53 activity by NES masking', *The EMBO Journal*, 18(6), pp. 1660–1672. doi: 10.1093/emboj/18.6.1660.

Subramanian, A. *et al.* (2005a) 'Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles', *Proceedings of the National Academy of Sciences*, 102(43), pp. 15545–15550. doi: 10.1073/pnas.0506580102.

Subramanian, A. *et al.* (2005b) 'Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles', *Proceedings of the National Academy of Sciences of the United States of America*, 102(43), pp. 15545–15550. doi: 10.1073/pnas.0506580102.

Subramanian, A. *et al.* (2017) 'A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles', *Cell*, 171(6), pp. 1437-1452.e17. doi: 10.1016/j.cell.2017.10.049.

Tai, Y.-T. *et al.* (2014) 'CRM1 inhibition induces tumor cell cytotoxicity and impairs osteoclastogenesis in multiple myeloma: molecular mechanisms and therapeutic implications', *Leukemia*, 28(1), pp. 155–165. doi: 10.1038/leu.2013.115.

Takahashi, K. *et al.* (1993) 'Protein synthesis-dependent cytoplasmic translocation of p53 protein after serum stimulation of growth-arrested MCF-7 cells', *Molecular Carcinogenesis*, 8(1), pp. 58–66. doi: 10.1002/mc.2940080112.

Teicher, B. A. *et al.* (2015) 'Sarcoma Cell Line Screen of Oncology Drugs and Investigational Agents Identifies Patterns Associated with Gene and microRNA Expression', *Molecular Cancer Therapeutics*, 14(11), pp. 2452–2462. doi: 10.1158/1535-7163.MCT-15-0074.

The Cancer Cell Line Encyclopedia Consortium and The Genomics of Drug Sensitivity in Cancer Consortium (2015) 'Pharmacogenomic agreement between two cancer cell line data sets', *Nature*, 528(7580), pp. 84–87. doi: 10.1038/nature15736.

The UniProt Consortium (2021) 'UniProt: the universal protein knowledgebase in 2021', *Nucleic Acids Research*, 49(D1), pp. D480–D489. doi: 10.1093/nar/gkaa1100.

Thompson, R. and Eastman, A. (2013) 'The cancer therapeutic potential of Chk1 inhibitors: how mechanistic studies impact on clinical trial design', *British Journal of Clinical Pharmacology*, 76(3), pp. 358–369. doi: 10.1111/bcp.12139.

Ueda, A. *et al.* (2019) 'Therapeutic potential of PLK1 inhibition in triple-negative breast cancer', *Laboratory Investigation; a Journal of Technical Methods and Pathology*, 99(9), pp. 1275–1286. doi: 10.1038/s41374-019-0247-4.

van de Wetering, M. *et al.* (2015) 'Prospective Derivation of a Living Organoid Biobank of Colorectal Cancer Patients', *Cell*, 161(4), pp. 933–945. doi: 10.1016/j.cell.2015.03.053.

Wang, A. Y. and Liu, H. (2019) 'The past, present, and future of CRM1/XPO1 inhibitors', *Stem Cell Investigation*, 6, p. 6. doi: 10.21037/sci.2019.02.03.

Webster, P. J. *et al.* (2017) 'AZD1775 induces toxicity through double-stranded DNA breaks independently of chemotherapeutic agents in p53-mutated colorectal cancer cells', *Cell Cycle (Georgetown, Tex.)*, 16(22), pp. 2176–2182. doi: 10.1080/15384101.2017.1301329.

Wickham, H. (2009) *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer-Verlag (Use R!). doi: 10.1007/978-0-387-98141-3.

Williams, S. P. and McDermott, U. (2017) 'The Pursuit of Therapeutic Biomarkers with High-Throughput Cancer Cell Drug Screens', *Cell Chemical Biology*, 24(9), pp. 1066–1074. doi: 10.1016/j.chembiol.2017.06.011.

Witkiewicz, A. K. *et al.* (2018) 'Targeting the Vulnerability of RB Tumor Suppressor Loss in Triple-Negative Breast Cancer', *Cell Reports*, 22(5), pp. 1185–1199. doi: 10.1016/j.celrep.2018.01.022.

Wong, C. H., Siah, K. W. and Lo, A. W. (2019) 'Estimation of clinical trial success rates and related parameters', *Biostatistics (Oxford, England)*, 20(2), pp. 273–286. doi: 10.1093/biostatistics/kxx069.

'XPO1 Inhibitor Approved for Multiple Myeloma' (2019) *Cancer Discovery*, 9(9), pp. 1150–1151. doi: 10.1158/2159-8290.CD-NB2019-085.

Xu, D., Grishin, N. V. and Chook, Y. M. (2012) 'NESdb: a database of NES-containing CRM1 cargoes', *Molecular Biology of the Cell*, 23(18), pp. 3673–3676. doi: 10.1091/mbc.e12-01-0045.

Yang, W. *et al.* (2013) 'Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells', *Nucleic Acids Research*, 41(Database issue), pp. D955-961. doi: 10.1093/nar/gks1111.

Yin, Y. *et al.* (2018) 'Wee1 inhibition can suppress tumor proliferation and sensitize p53 mutant colonic cancer cells to the anticancer effect of irinotecan', *Molecular Medicine Reports*, 17(2), pp. 3344–3349. doi: 10.3892/mmr.2017.8230.

Yu, G. *et al.* (2012) 'clusterProfiler: an R package for comparing biological themes among gene clusters', *Omics: A Journal of Integrative Biology*, 16(5), pp. 284–287. doi: 10.1089/omi.2011.0118.

Zhao, B., Hemann, M. T. and Lauffenburger, D. A. (2014) 'Intratumor heterogeneity alters most effective drugs in designed combinations', *Proceedings of the National Academy of Sciences*, 111(29), pp. 10773–10778. doi: 10.1073/pnas.1323934111.

Zhao, S. *et al.* (2020) 'Molecular subtypes and precision treatment of triple-negative breast cancer', *Annals of Translational Medicine*, 8(7). doi: 10.21037/atm.2020.03.194.