MACHINE LEARNING FOR DRUG DEVELOPMENT: INTEGRATING GENOMIC, CHEMICAL, AND CLINICAL DATA TO IDENTIFY DRUG TARGETS, EFFICACIES, ADVERSE EVENTS, AND COMBINATIONS

A Dissertation
Presented to the Faculty of the Weill Cornell Graduate School
of Medical Sciences
in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy

By
Neel S. Madhukar
August 2017

MACHINE LEARNING FOR DRUG DEVELOPMENT: INTEGRATING GENOMIC, CHEMICAL, AND CLINICAL DATA TO IDENTIFY DRUG TARGETS, EFFICACIES, ADVERSE EVENTS, AND COMBINATIONS

Neel Srivastava Madhukar, Ph.D.

Cornell University 2017

Despite recent technological advances, drug development has remained a challenging and inefficient process. Machine learning methods have the potential to accelerate this process by using information from past drug successes and failures to decipher the mechanisms and activities of new compounds. This will become even more crucial in the age of “precision medicine” where thorough mechanistic knowledge will be needed to properly position compounds. The purpose of this dissertation is to address this through the development of methods for drug target identification, biomarker identification, indication selection, and adverse event prediction.

First we introduce BANDIT to accelerate the process of drug target identification/deconvolution. BANDIT integrates multiple different data types within a Bayesian network to predict the targets for both new and approved small molecules. We found that BANDIT was able to accurately recover a large number of known drug-target interactions, identify new drugs for a common cancer target, and identify DRD2 as the target for ONC201 – a first-in-class molecule in clinical development. Our work on ONC201 led us to ask how we could integrate known information on DRD2 with gene expression profiling and BANDIT to better select analogs and indications for ONC201. We found that we could accurately rank analogs based on measured efficacy,

select new cancer types where ONC201 was likely to be efficacious, and identified DRD5 and cancer stem cell genes as biomarkers for ONC201 activity. Following our work on ONC201 and drug target identification, we asked whether these methods could be applied to predict specific adverse events for a specific drug. Building off previous work published by our lab, we developed MAESTER, a data-driven machine learning approach that integrates properties on a compound's structure and targets, with tissue wide gene expression profiling and known biological networks to calculate the probability of a compound presenting with a set of tissue specific adverse events in the clinic. We found that MAESTER could accurately identify known side effects of approved drugs and could even pinpoint the adverse events of drugs that were approved and later withdrawn for tissue specific toxicities.

Altogether this work demonstrates how challenging problems in drug development could be addressed through the integration of diverse datasets. These approaches have the potential to transform the current drug development pipeline by focusing experimental efforts, and identifying new compounds with therapeutic potential, and choosing optimal indication and patient populations – all which could have a direct impact on patient care.

BIOGRAPHICAL SKETCH

Neel S. Madhukar graduated from the University of Tennessee at Knoxville in 2013 with a Bachelors of Science in Biochemistry & Cellular and Molecular Biology. During his undergraduate career he worked at the Center for Prostate Disease Research, analyzing expression data from patients with Prostate Cancer to identify biomarkers of progression and response. He joined the Tri-Institutional Training Program in Computational Biology and Medicine Program in 2013 where he rotated in the labs of Drs. Jason Locasale and Haiyuan Yu researching metabolism and protein dynamics in cancer respectively. He became a member of the Elemento lab in 2014 where his work has focused on the development of predictive methods for identifying drug targets, toxicities, efficacies, and combinations.

This dissertation is dedicated to C. and my friends who kept me smiling even on the hardest days; to Dolla for giving me the motivation to always make her proud; to Mamma for her unwavering love and support; to Pappa for giving me my ambition and keeping me inquisitive; and to my Nana for always being there, no matter where he is, whenever I needed him.

ACKNOWLEDGEMENTS

I would like to thank Olivier for helping me become the scientist I am today. I joined CBM to work with you, and since that day you have given me the freedom to pursue goals that many thought were untenable and provided the support and guidance to help prove them wrong. You were part of my journey from the very beginning and I look forward to continuing this journey together.

I would like to thank Evi and the Giannakakou lab and Josh and the Oncoceutics team for their partnership and for showing me the impact my work could have on real problems, and my committee– John Chodera, Marcin Imielinski, and Paraskevi Giannakakou – as well as the Mezey and Elemento lab members for the insightful feedback and discussions.

I would like to thank David, Margie, and the Tri-Institutional PhD Program in Computational Biology & Medicine, for their constant support.

I would like to thank my labmates for their close collaboration, without which this dissertation would not have been possible, my friends for providing a space where I could discuss ideas and for forcing me the defend my thoughts, and to Priyanka Vijay for her advice and input on problems big and small and for keeping me going forward even when a problem seemed insurmountable.

This work was supported by the CAREER grant from National Science Foundation (DB1054964), NIH grant R01CA194547, the Starr Cancer Foundation, as well as by startup funds from the Institute for Computational Biomedicine. Support was also provided by the PhRMA Foundation Pre Doctoral Informatics Fellowship and by the Tri-Institutional Training Program in Computational Biology and Medicine (via NIH training grant T32GM083937).

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER ONE

INTRODUCTION

Over the past 20 years there has been an emergence of technologies and databases that facilitate that facilitate the application of “Big-Data” and machine learning technologies to healthcare. Already we’ve seen these technologies being applied to diverse areas such as the interpretation of electronic medical records or image analysis [1,2]. However, drug development has remained relatively unchanged over the past decade despite these new advancements. In fact, R&D productivity in most major pharmaceutical corporations has actually fallen in the past few years with drug costs continuing to rise [3]. This is often attributed to the length of the drug development process (with most drugs taking an average of 12 years and $2.6 billion to reach approval stages) and the fact that most steps are driven by costly, tedious, and case-specific experimentation [4]. With this in mind, there are many opportunities for machine learning methodologies to accelerate the process such as during target deconvolution, biomarker identification, analog selection, toxicity profiling, and repurposing.

The majority of candidate identification efforts for further development fall into one of two main categories: 1) target-based screening where compounds are designed with a specific protein target in mind and 2) phenotypic screening, where a large number of compounds are screened to determine which can induce a phenotype of interest [5]. Recent reports have shown how phenotypic screening based methods are actually more efficient in terms of generating novel small molecule compounds, however one major challenge for compounds identified this way is target deconvolution – identification of the binding targets of a given compound that cause a certain phenotypic effect [6,7]. Current experimental efforts are often trial-and-error based without any efficient ways to broadly survey the massive number of potential targets. This becomes even more complicated once we consider that, on average, a drug binds to six different targets, shattering the traditional "one drug, one target" reasoning [8]. Current computational approaches to target deconvolution have great utility, however one limiting factor is that they often require prior information on known binding targets for a given drug, a complex 3D structure of all target proteins, or demand extensive computational resources [9-11]. Additionally, most published methods only integrate one or two of the available data types for a given compound, yet as more data becomes publically available, it now more tractable than ever for *in silico* target deconvolution methods to integrate a number of diverse and orthogonal data types into a single prediction.

In the age of precision medicine, the emphasis has shifted from designing a drug that will be effective for all patients in all indications to instead identifying the optimal indications and patients for a given compound. Even once a target has been identified for a compound, this can be a challenging endeavor as often multiple factors can contribute to efficacy of a compound in a given patient, such as tissue specific gene expression profiles and complex biological networks. In oncology, indication selection has traditionally consisted of screening a compound against hundreds of different cell lines to determine which class best responds, or by running multiple, costly clinical trials to approximate the best responding patients [12]. A number of studies have shown how integrating target information with large scale sequencing efforts (such as TCGA) and known protein-protein or pathway interactions can pinpoint indications and biomarkers to focus experiment or clinical efforts toward [13,14]. Additionally retrospective analyses of efficacy screening results with cell line genomic information can help identify predictive biomarkers of response that could be used to direct future experiments.

Moving past pre-clinical development into clinical trials, one of the largest, and most costly, causes of drug failure is unexpected toxicities or adverse events [15,12]. Currently drug likeness methods are used to filter out compounds with undesirable features that are correlated with toxicity issues, such as poor bioavailability, however we had previously found that drug likeness measures were insufficient to flag toxic compounds [16,17]. Additionally, these methods are designed to detect broad toxicities, whereas for clinical development it is crucial to understand which specific adverse event or tissue-specific toxicities to expect. In vitro screens and animal studies are another tool used to identify specific adverse events prior to entering human trials, however these are slow and costly, and many results may not translate to effects in humans [18]. Integrative computational methods have the potential to not only expediete this process, but could also help us understand the specific features that contribute to adverse events. This information could then be used to ensure better drug design in the future.

In this these I hope to address a subset of these challenges in the drug development pipeline. I shall being with a review of current approaches to predict drug efficacy and indications. From there I will describe different approaches to address drug target deconvolution, analog optimization, indication selection, biomarker identification, and the prediction of specific adverse events. The first method is a machine learning approach that can integrate multiple diverse data types to predict targets for novel compound and identify new compounds targeting a specific protein. From there I will describe how I used the aforementioned method and genomic analyses to identify optimal analogs, indications, and pathway-specific biomarkers for a novel class of anti-cancer small molecules. The last method discussed is a machine that integrates features on a compound's structure and targets with tissue wide expression profiles and known biological networks to predict whether a compound is likely to cause specific adverse events (such as neutropenia or heart attacks).

CHAPTER TWO

BIOINFORMATICS APPROACHES TO PREDICT DRUG RESPONSES FROM GENOMIC SEQUENCING[*]

PREAMBLE

This chapter is an edited version of review chapter that is in press at *Cancer Systems Biology*. The idea for the chapter was conceived in partnership with Dr. Olivier Elemento. All writing was done myself with input from Dr. Elemento.

INTRODUCTION

One of the greatest challenges in the current paradigm of medicine is how to deal with patient heterogeneity – both across different diseases and even within patients diagnosed with the same disease. Over the past 50 years there have been many studies showing that patients with the same disease have completely different responses when treated with the same drug [19-22]. The prevailing hypothesis to explain the heterogeneous response is each patient's specific genetic profile. Precision medicine involves using this patient-specific genomic information to guide drug treatment, with the expectation that this will ultimately improve clinical outcomes [23]. With the decrease in sequencing costs over the past decade, it is now possible to obtain genomic information for patients prior to determining a specific treatment regimen. In addition, there has been an emergence of bioinformatics methods to interpret this sequencing data and come up with actionable strategies for precise drug choices. These methods not only allow for the identification of specific genetic traits that confer

[*] Madhukar NS and Elemento OE. "Bioinformatics Approaches To Predict Drug Responses From Genomic Sequencing." Cancer Systems Biology. 2017 (In Press)

susceptibility or resistance to drug treatment, but can also combine genetic markers with gene ontologies and biological networks to predict precise response levels. In this chapter we provide an overview of these bioinformatics methods, review the basic premises for each type of method, and discuss some of the current problems and future challenges that need to be solved. While we tend to focus on cancer, the databases and methods we described are often applicable to other diseases, as well.

DATABASES

In recent years there have been a number of community efforts to generate and publicly release datasets that could be used to improve drug response prediction **(Table 3.1)**. In this review we will cover what we believe to currently be the best-suited and most popular public resources for aiding drug response prediction:

NCI60 Drug Sensitivity Database

The National Cancer Institute's (NCI) 60 cell line drug screen is a database of in vitro drug efficacies (either in terms of GI50, LD50, or TGI) for over 50,000 compounds screened against the NCI60 panel of cancer cell lines [24]. With 60 cancer cell lines from 9 distinct tumor types – leukemia, colon, lung, central nervous system, renal, melanoma, ovarian, breast, and prostate – the NCI60 collection aims to provide information on a broad set of genetic conditions and tumor types. The NCI60 panel has itself been profiled using a variety of assays from genomic to gene expression and proteomics [25-28]. The profiling data can be used in conjunction with the Developmental Therapeutics

Program's (DTP) drug screening database to identify genetic signatures indicative of a certain response pattern.

Table 2.1 – List of abbreviations and websites referenced

Abbreviation	Full Description	Website
GI50	Concentration of a compound that leads to a 50% inhibition of cell proliferation	
IC50	Concentration of a compound that leads to a 50% decrease in the desired activity	
LD50	Concentration of a compound that leads to 50% cell death	
TGI	Total growth inhibition	
GWAS	Genome Wide Association Study	
SNP	Single Nucleotide Polymorphism	
DREAM	Dialogue on Reverse Engineering Assessment and Methods	
NCI60-DTP	Drug screen of 60 cancer cell lines by the National Cancer Institute's (NCI) Developmental Therapeutics Program	https://dtp.cancer.gov
CCLE	Cancer Cell Line Encyclopedia	http://www.broadinstitute.org/ccle/home
CMap	Connectivity Map	http://www.broadinstitute.org/cmap
GDSC	Genomics of Drug Sensitivity in Cancer	http://www.cancerrxgene.org
TCGA	The Cancer Genome Atlas	http://cancergenome.nih.gov
GTEx	Genotype-Tissue Expression	http://www.gtexportal.org/home/

Cancer Cell Line Encyclopedia

The Cancer Cell Line Encyclopedia (CCLE) [29,30] is a database of 947 different human cancer cell lines encompassing 36 different tumor types that have been genetically profiled – gene expression, copy number, mutations, etc. Furthermore, 24 known anticancer drugs were profiled against approximately 500 of these cell lines. Though the number of compounds profiled is smaller than the NCI60 drug screen, the greater number of cell lines tested allows for more precise identification of genetic predictors of sensitivity for the drugs measured.

Genomics of Drug Sensitivity in Cancer

Hosted by the Wellcome Trust Sanger Institute, the Genomics of Drug Sensitivity in Cancer (GDSC) database is a massive drug screen project similar to the NCI60 and CCLE. In their initial release, investigators screened a set of 138 known anti-cancer compounds against over 1000 different cancer cell lines (on average 525 cell lines tested per compound). Each cell line also was subjected to thorough expression and copy number profiling along with targeted mutation data for a set of 75 cancer genes. This dataset constitutes another great resource for the identification of genomic markers of drug responses.

Connectivity Map/LINCS

Released by the Broad Institute, the Connectivity Map (CMap) seeks to find connections between small molecules, physiological processes, and disease states [31]. Using mRNA expression (measured by DNA microarrays) as the “language” of cellular response, the CMap measures how a panel of cancer cell lines responds transcriptionally to a variety of different drug treatments. This approach had previously been successful in identifying drug mechanisms in yeast but had never been applied to cancer cells [32]. The investigators profiled 4 different cancer cell lines before and after treatment with a panel of more than 1000 small molecules. The LINCS database is an updated version of this profiling system with a much larger number of drugs and cell lines. This database makes use of the LINC1000 expression profiling system where the expression of 1000 key genes is measured and used to infer the global gene expression profile. From these transcriptional changes it is possible to explore a drug's mechanisms of action. These could be used to successfully repurpose drugs for specific diseases or genetic states [33,34].

IDENTIFICATION OF GENOMIC MARKERS OF DRUG RESPONSE

A key first step to any drug response prediction effort involves the identification of genomic markers that can impact efficacy. Identifying those markers makes response prediction a much simpler task. Once a polymorphism, gene expression pattern, or pathway has been identified, all new samples can simply be screened for that marker and, using known correlations with drug response, a prediction of drug susceptibility can be made. Here we focus on a variety of approaches that can be used to identify genomic markers indicative of drug response.

Using Genome Wide Associate Studies to Identify Polymorphisms related to Drug Response

Genome Wide Associate Studies (GWAS) have classically been used to detect genetic variations associated with specific disease phenotypes.

However, in recent years, the use of GWAS has proved to be a powerful method to identify polymorphisms that can affect drug efficacy and toxicity [35]. Unlike approaches focusing on known drug targets or candidate gene lists, GWAS provides a hypothesis-free method that can systematically test a large number of variants [36,37]. In order to run a GWAS one must provide a measure of response or toxicity for a large number of samples, as well as a thorough genotyping of each sample.

GWA studies typically fall into two main categories depending on whether the provided response measure is categorical (such as case/control, responder/non-responder, adverse reaction/no reactions, etc.) or quantitative (such as IC50 or a measure of side effect severity). Recently there have been a series of developments improving the traditional GWAS, such as taking into account a gene's functional information [38], epistasis [39], or missing data [40]. Here we review the basic premise of the categorical and quantitative GWA studies:

1. **Categorical** – The goal of a categorical GWAS is to identify SNPs that are highly predictive of which category a given sample will be assigned to. To begin, samples are assigned to one of the two categories based on either their response to a given drug or the observation of a given adverse effect. For each observed SNP, we count the number of samples where that SNP is present (or absent). This data is then used to populate what is known as a contingency table. For instance, if in a dataset with 100 responders and 500 non-responders we observe 90 responders with a certain SNP and 15 non-responders with that same SNP (**Table 2.2**). A statistical test is then run on each contingency table to measure the

deviation from the null-hypothesis, which assumes that there is no association between the SNP and categorical classes. The most common test used is either the chi-squared test (or the related Fishers exact test). This approach has successfully identified variants related to interferon beta [41] and anti-TNF treatment efficacy [42] as well as variants predictive of statin-induced myopathy [43].

Table 2.2 – Sample contingency table showing how we can use the number of responders with a certain SNP to test whether it is related to drug efficacy.

	Responders	Non-Responders
SNP Present	90	15
SNP Absent	10	485

2. **Quantitative** – Instead of using a contingency table test to detect significantly associated SNPs, a quantitative GWAS traditionally uses a generalized linear model (GLM), such as an Analysis of Variance (ANOVA) – a variant of a linear regression analysis – to identify SNPs that are highly correlated to the variable of interest (such as drug IC50)[44]. Though more complicated than the categorical case, there exist a number of public bioinformatics software packages such as PLINK[45] or SNPTEST[46] that can run quantitative GWAS and output a p-value for each polymorphism. While these analyses are less common for drug response prediction because of the difficulty in measuring quantitative response values, various groups have successfully used them to identify SNPs associated with susceptibility to chemotherapeutic drugs[47] or ACE inhibitors[48].

Regardless of the type of GWAS used, the output is a set of p-values, one for each polymorphism tested. One important caveat is that all p-values must be corrected for multiple hypothesis testing (MHT) to account for the large number of statistical tests being performed. The most commonly used methods for MHT are the Bonferroni or Benjamini-Hochberg corrections. Adjusted P-values are then visualized using a Manhattan plot where the genomic position of each SNP is plotted against the negative log of its p-value (**Figure 2.1**). Using the Manhattan plot one can visually identify genomic regions or particular SNPs that are significantly associated with the given response feature.

Using Gene Expression To Find Response Signatures and Predict Response

While GWA studies aim to find a set of mutations or polymorphisms that are predictive of how a patient will respond to a drug, another popular approach is using gene expression data to find an expression signature associated with a positive (or negative) response. Different transcriptional profiles can often lead to different levels of drug efficacy, and differential expression analyses can help pinpoint the specific genes or pathways that drive the heterogeneous drug response and can be used to predict response levels.

The classic approach involves treating a cohort of mice or patients, or patient samples or cell lines with a given drug and measuring the degree of response in each sample. Similar to a GWAS, the response rate can be measured either

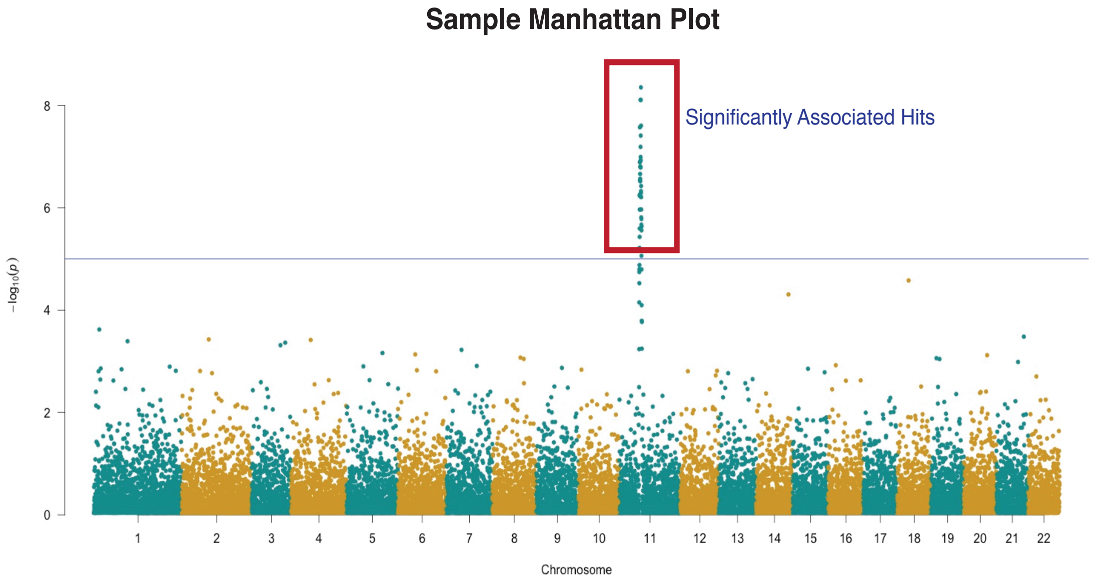

Figure 2.1 – Sample Manhattan plot showcasing how one can use the output of GWAS calculation to find SNPs related to drug efficacy. Boxed hits represent those that pass the significant p value cutoff and thus may be relevant to treatment response.

categorically (responder/non-responder) or as a continuous variable. Using either sequencing data from before treatment or differential gene expression (comparing pre and post-treatment samples) one can search for gene expression patterns that seems more prevalent in the samples that are susceptible (or resistant) to treatment (**Figure 2.2**). For instance, one would expect to see genes that confer drug resistance to be more highly expressed in samples where drug treatment shows limited effect.

A number of methods exist for detecting differential expression across a set of samples. For microarray data oftentimes statistical tests such as an ANOVA would suffice, but packages such as limma [49] use linear models that can help deal with more complicated experimental designs. For RNA-seq data the most popular methods include a limma-voom [50], DESeq2 [51], edgeR [52], and cufflinks (cuffdiff) [53]. DESeq2 and edgeR are currently considered the standard for differential expression analysis and both use similar underlying models (however with different dispersion estimates). However, in our experience we have found DESeq2 to be more conservative. One key difference between DESeq2/edgeR and limma-voom is that voom doesn't employ a negative binomial distribution and instead estimates the mean variance relationship. Therefore voom may be a better choice if the input data differs strongly from a negative binomial distribution. Finally, one major difference between the cuffdiff pipeline and DESeq2 is that cuffdiff acts on the level of transcripts while DESeq2 uses gene counts as inputs. Additionally, Wright et al [54] used a Bayesian predictor to automatically separate samples into subtypes based on their respective gene expression profiles, and used the output p-values to find the set of genes most predictive of subtype. This type

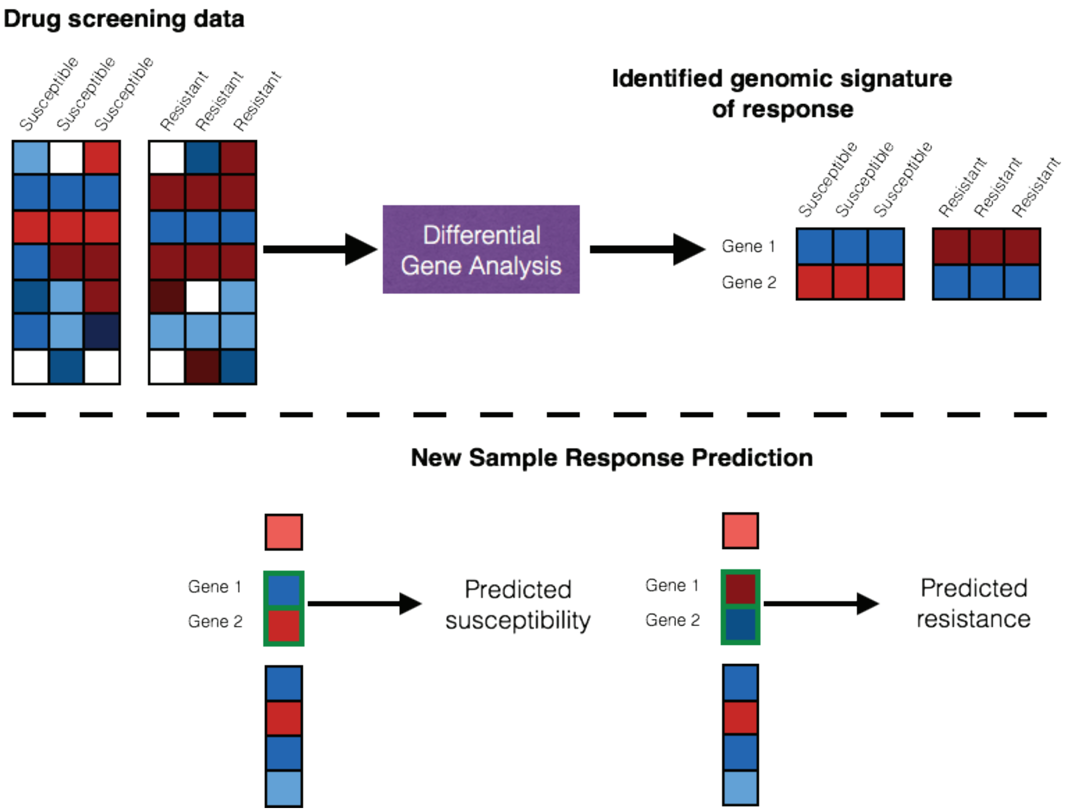

Figure 2.2 – Diagram on how gene expression patterns from responders and non-responders can be used to identify signatures related to response and how these can be used to better select new patients likely to respond.

of approach is useful for pooled sets of samples without knowledge of their subtype – for instance when one would like to determine if well-responding patients all fall into a certain disease subtype [55]. While initially tested on microarray data, this approach can be easily adapted to RNA-seq data and could generally be adapted to all types of predictive models.

Using Pathway Annotations and GSEA to Identify Differential Biological States

Often a differential gene expression analysis will have a set of genes as output, which has no obvious pattern or relevance to the type of drug being investigated. Additionally, it is quite common for a set of genes to be marked as significant in a differential gene expression analysis, but when experiments are done to perturb individual genes they seem to have little to no effect on drug response. In cases like these it is often helpful to translate the differentially expressed genes into a set of enriched biological pathways or gene sets. These can provide a broader explanation of a drug's mechanism of action and a clearer understanding on how to predict efficacy. This approach has previously been successful not only in drug response prediction, but also in the development of highly effective drugs. Overexpression of the mTOR pathway in lymphoma led to the development of inhibitors to specifically target genes in that pathway [56], and global activation of the epidermal growth factor receptor pathway was found to be predictive of erlotinib susceptibility in pancreatic cancer xenografts[57].

The basic technique to finding enriched pathways or canonical gene sets is to first annotate each gene based on the pathways/sets it falls into. A few popular resources for pathway and gene set annotation include: the Molecular

Signatures Database (MSigDB)[58], Reactome[59,60], the Kyoto Encyclopedia of Genes and Genomes (KEGG)[61], Gene Ontologies[62], and InnateDB[63,64]. Reactome, KEGG, and InnateDB group genes based on their biochemical pathways (with InnateDB focusing on pathways relating to immunity), Gene Ontologies group genes based on their biological/molecular function or cellular localization, and MSigDB is a combination of all the aforementioned databases with custom sets of “hallmark” gene sets, or important genes involved in certain processes. Following annotation, a statistical test (such as the Fishers exact test) can be used to test whether a certain pathway is enriched for up (or down) regulated genes compared to what would be expected by random chance.

Another popular method for testing pathway enrichment is Gene Set Enrichment Analysis (GSEA)[65]. GSEA tests whether genes of a certain pathway/set are differentially expressed between the cases. It does this by computing an enrichment score for each gene set – increase in score if genes in set are differentially expressed, decrease in score if not – and using a number of permutations (number can be set by the user) it tests whether that enrichment score is significantly different than what would be expected by chance. Packaged with the MSigDB gene sets, GSEA has demonstrated success at identifying common biological pathways in independent lung cancer datasets while single-gene differential analyses could not [66].

IDENTIFYING DRUG TARGETS AND MECHANISMS AND USING THEM TO IMPROVING RESPONSE

Computational Techniques to Identify Drug Targets and Mechanisms

For a small molecule in development the mechanisms of action and binding targets are often not fully understood. A number of computational methods exist that seek to predict targets for these orphan small molecules, based either on chemical structure or its down-stream effects. These methods can broadly be divided into three categories:

1. Molecular Dynamics: Using intricate mathematical models, molecular dynamics methods computationally simulate a drug's interaction with a given protein. To predict targets, an orphan small molecule is tested against a series of proteins to identify any with favorable binding results [10,67]. However, this approach requires significant computation power, complex mathematical models, and full 3D structures for each queried protein – data that is often unavailable.
2. Ligand-Based [68,69]: Using a set of known protein binding partners for a given small molecule, ligand-based approaches apply machine learning techniques to find other proteins with high enough similarity to the known targets. The proteins with high degrees of similarity are predicted to be novel binding targets. However ligand-based methods often require a large number of known binding partners for each tested small molecule, and thus can mostly be used on drugs far enough in the drug development phase.
3. Downstream Effect Based: Recently, a number of methods emerged, which use the downstream effects of a small molecule (such as induced

gene expression change [70] or side-effects [71]) to predict targets for orphan small molecules. The basic premise of these methods is to compare the effects of an orphan small molecule to the effects of drugs with known targets. If the orphan molecule has an effect very similar to a drug with a known target, one would predict this known target to also be a target of the orphan small molecule. However, most current methods only utilize a small number of the available data sources and are thus not broadly applicable to all drug types. Our lab recently developed BANDIT (See Chapter 3), a novel computational method that integrates multiple different pieces of data on small molecules to predict specific binding targets and mechanisms [72]. When tested on a set of diverse drugs, BANDIT achieved an accuracy of approximately 90% at identifying known targets (validated using a standard cross validation setup), much higher than expected from other target prediction methods.

Another popular option is to focus on a drug's broad mechanism of action rather than its specific binding targets. One way to accomplish this is to observe how a given drug changes the transcriptional profile in a sample. For example, using gene expression data following cisplatin treatment, this type of analysis identified the p53 response and other pathways to be involved in cisplatin response [73]. This approach has become more practical with the emergence of public databases such as the Connectivity Map (CMap)[74]. From the CMap database, one can calculate fold change values for each gene after drug treatment. Using GSEA or other pathway enrichment methods, the fold change values can be converted into a set of pathway scores that reveal which pathways were enriched or mobilized. Though far less precise than specific target identification, this information is easier to obtain and could provide additional information on the context in which a given drug could be used.

Using Known Drug Targets To Predict Response

Assuming one can determine the mechanisms of action of a drug – either in terms of specific binding targets or broad knowledge on the biological pathways mobilized – the task of predicting efficacies is often much simpler. For example, if a drug's main mechanism of action is to target Protein A, then one would expect different efficacies in samples based on whether there is an amplification or deletion of Protein A. This type of reasoning also applies when there are mutations in a known drug target. Examples of this are treatments involving Gefitinib or Herceptin. Gefitinib is an anti-cancer small molecule known to target the EGFR kinase, and mutations in EGFR were found to predict sensitivity of samples to gefitinib treatment [75]. Herceptin, an antibody which targets HER2, was found to improve the outcomes of cancer patients with HER2 amplifications or activating mutations [76,77]. Another example of this concept is vemurafenib – a small molecule that targets V600E BRAF mutation – that has been found to be selectively effective in cancer patients with this exact mutation, while having no beneficial effect on normal BRAF samples [78-80]. These are just a few of the many examples showing how combining known drug targets with targeted sequencing can help detect instances of differential response.

However, it is also important to note that while the alterations of a drug's target are often predictive of efficacy, this is not always the case, even if the target itself serves as biomarker [81]. Moreover, there are often cases where the predictive biomarker for a given drug is not the actual target, but rather another gene or set of genes involved in the same pathway or biological processes as drug's target. In cases like these sequencing could still prove to be a valuable tool, and we advise utilizing some of the other methods mentioned in this chapter. Drug target information could be used in combination with these methods to refine predictions and gain greater biological insights.

Sequencing-based approaches also can be very successful in positioning drugs for specific disease conditions – especially different cancer types. Using resources like the Cancer Genome Atlas (TCGA)[13] and Genotype-Tissue Expression (GTEx) project [82], one can find genes or pathways that are significantly upregulated in certain cancers or cancer types compared to either normal tissue samples or other cancer subtypes. Identifying such cancer-subtype-specific, upregulated signatures could highlight drugs known to target these signatures as particularly viable candidates for treatment. For instance, it was recently discovered that dopamine receptors were selectively upregulated in neoplastic stem cells in breast cancer. It was observed that thioridazine (a compound known to target dopamine receptors) was particularly effective against these cell populations [83].

4.3 Exploiting Genetic Interactions (SL/SDL)

One approach that has become increasingly popular is exploiting networks of synthetic lethality (SL) and synthetic dosage lethality (SDL) to predict drug efficacy. SL describes a specific type of genetic interactions involving two or more genes, where the loss of either gene individually is non-fatal, but the combined loss of all SL partner genes leads to a severe decrease in fitness or cell death. SDL describes a related genetic interaction where lethality is observed when one gene is lost while its SDL partner is overexpressed [84,85]. Both SL and SDL interactions are highly relevant to cancer biology, as most cancers have both widespread losses and gains of certain genes. Exploiting these could drastically improve patient prognosis. For instance, if Gene A and Gene B are in an SL pair and Gene A is lost in a given cancer sample, then one would expect compounds targeting Gene B to have better responses in this sample (**Figure 2.3**).

To this end there have recently been many efforts to uncover underlying SL and SDL networks in cancer. Among the most successful efforts was the data mining synthetic lethality identification pipeline DAISY[86]. DAISY uses three distinct hypotheses to detect SL pairs (with the inverse hypotheses being used for SDL pair detection):

1. Genes in an SL pair will have significantly lower raters of co-mutation or co-loss
2. Knockout/knockdown of a given gene will be more fatal in samples with under-expression or loss of its SL partner
3. Genes in an SL pair are more likely to be co-expressed

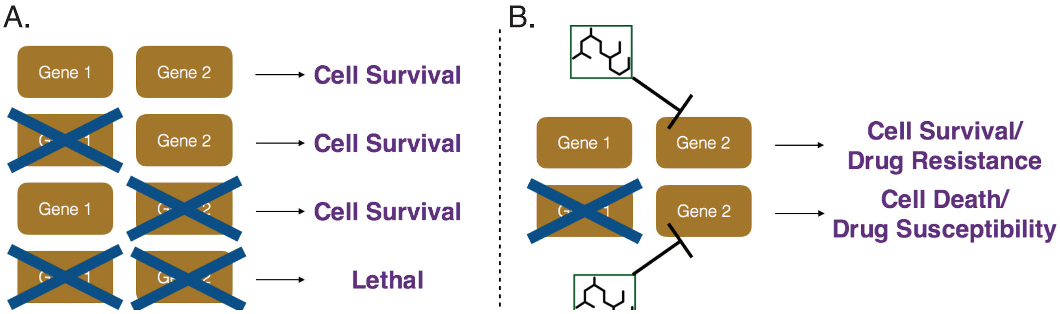

Figure 2.3 – A) Diagram highlighting the concept of synthetic lethality and how known synthetic lethal relationships can be combined with genomic information to better predict drug response. B) Using synthetic lethality to predict differential response

By scanning for gene pairs that fulfill all three hypotheses, DAISY predicted networks of SL and SDL interactions. It achieved an accuracy level of approximately 77% (measured by Area Under the Receiver Operating Curve) when compared to known SL interactions, demonstrating that DAISY could accurately infer SL and SDL genetic interactions. To translate this into predicting drug responses, the authors identified sample-specific exploitable interactions, or SDL interactions where one gene was overexpressed and SL interactions where one gene was lost. DAISY then identified drugs known to target the other gene in each exploitable interaction. For each drug DAISY ranked the most sensitive samples based on the number of exploitable interactions being targeted by each drug. They found that specific drugs were significantly more effective in cell lines predicted to be sensitive than those predicted to be resistant. Furthermore, the authors used a similar approach to predict the exact IC50 value for each drug across a set of cancer cell lines and observed a strong correlation between the predicted and observed values (R = 0.721). Taken together these results show how known genetic interactions (particularly SL and SDL interactions) can be combined with sequencing data to better predict drug sensitivities and inform treatment.

MACHINE LEARNING APPROACHES

In cases where identification of response biomarkers is too complex or the identified biomarkers do not reveal any underlying biological insight, machine learning approaches, which can combine sequencing data with information such as biological networks, are very powerful. The idea for employing machine learning approaches for drug response prediction is for the computational algorithm to learn how to combine a set of distinct features into a prediction of sensitivity. Most machine learning methods for drug sensitivity prediction are classified as supervised methods. Those supervised methods use a set of sequenced samples with known drug sensitivities to “train” the algorithm and determine how to combine features based on their predictive power (**Figure 2.4**). While the linear regression model discussed earlier can be considered the oldest form of machine learning, most popular methods currently utilize more advanced modeling to account for the complexity in genetic sequencing data. In fact, machine learning methods can often detect higher order genomic markers of drug response that other methods may have missed. One example is the use of machine learning to identify the EWS-FL11 translocation in Ewing’s sarcoma as a marker of sensitivity to PARP inhibitors [87].

Many methods seek to improve their performance by including additional information on known biological networks, genetic interactions, or drug chemical properties. For instance, Menden et al [88] found that including drug chemical information (such as weight and lipophilicity) with sequencing data improved the performance of both a neural network and random forest for sensitivity prediction. In collaboration with the NCI, the Dialogue on Reverse Engineering Assessment and methods (DREAM) project led a community effort to improve drug sensitivity predictions [89]. Through this effort, the NCI-DREAM consortium publically released drug sensitivity data for a set of breast cancer cell lines along with thorough genetic, epigenetic, and proteomic sequencing data. Individual groups each submitted different sensitivity prediction methods and the NCI-DREAM consortium analyzed each method to identify any particular method features that led to higher accuracies. I

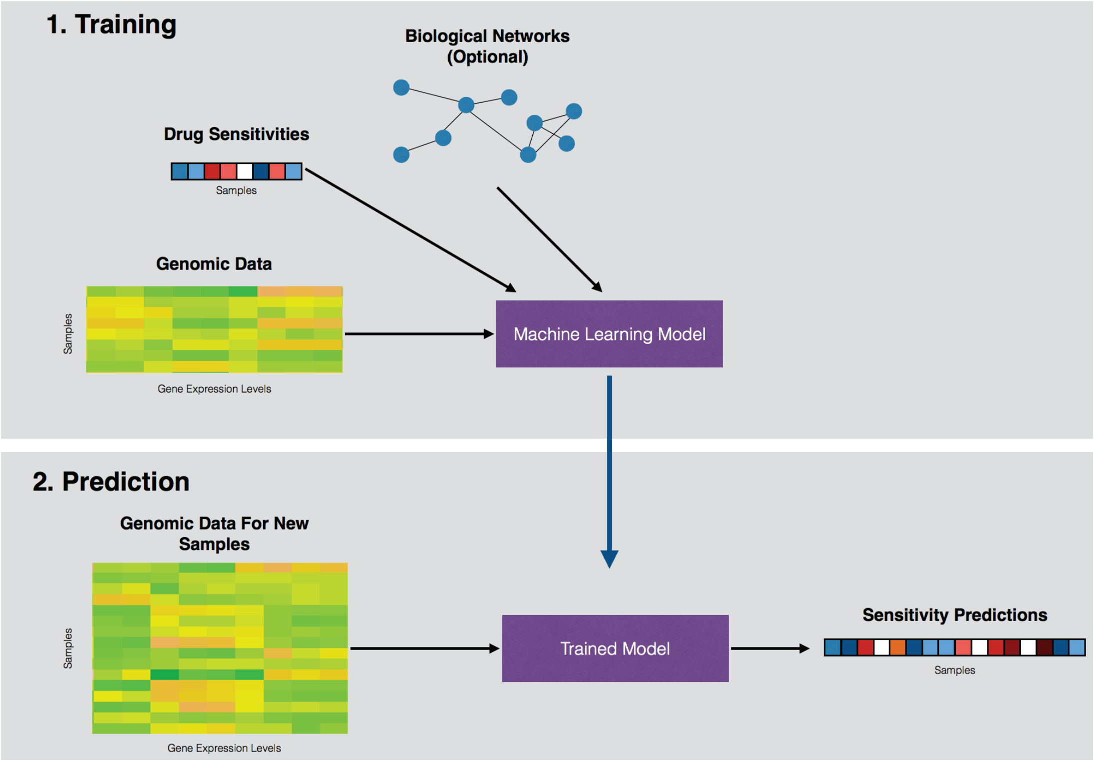

Figure 2.4 – Overview of how common machine-learning methods combine multiple data types to train a specific model that can be applied to new samples to predict sensitivity.

Interestingly, they found that the inclusion of annotated biological pathways was one of the two variables that significantly boosted performance [89]. Additionally, the consortium found that the top performing methods all utilized nonlinear modeling, indicating that in many cases the connections between individual genetic features and drug response are too complex to be understood using a strictly linear approach. Finally they observed that though sensitivity to proteasome inhibitors tended to be predicted with the most accuracy, there was a predictive signal for most of the drugs in their test set. This further indicated that machine-learning methods have the potential to significantly improve sequencing based drug response prediction.

Conclusion and outlook

In the past two decades, there have been significant advances in using genomic data and bioinformatics to better understand the heterogeneous nature of drug response. By combining data on genomic alterations and drug response with thorough statistical methods we can identify specific predictive markers. Moreover, through post-treatment genomic profiling we can gain a better understanding of the mechanism and effect of a given drug. This knowledge can then be used to better select patients or diseases where that mechanism will provide the most therapeutic benefit. Additionally, there has recently been an emergence of computational methods to identify drug targets when conventional approaches fail. However, as the amount of data generated continues to increase and drugs targeting new pathways are developed, we imagine that no single approach or method will provide high enough accuracy. Therefore we expect the field to move towards using machine learning strategies that are able to integrate a variety of different data-types into a single predictive output. We are already seeing the creation of sophisticated methods for this purpose and we anticipate this to only improve over the coming years. All together though we believe that the adoption of the methodology described in this chapter not only has the power to expand our understanding of pharmacology but can also significantly improve the current schema of patient treatment.

CHAPTER THREE

A NEW BIG-DATA PARADIGM FOR TARGET IDENTIFICATION AND DRUG DISCOVERY*

PREAMBLE

This chapter consists of a paper that has been submitted and is currently under review. The method (BANDIT) was conceived in partnership with Dr. Olivier Elemento. I implemented the method and subsequent computational analyses. Linda Huang and Katie Gayvert contributed to model deployment and follow-up analyses. The experimental follow-up for ONC201 was done by Oncoceutics Inc. (M.S. and J.A) and all microtubule experiments were done by the Giannakakou lab (PK, GG, PG). I primarily wrote the manuscript with input from PK, JA, PG, and OE.

INTRODUCTION

It typically takes 15 years and 2.6 billion dollars to go from a small molecule in the lab to an approved drug [90-92], and for natural products and phenotypic screen derived small molecules, one of the greatest bottlenecks is identifying the targets of any candidate molecules[91,93]. Proper understanding of binding targets can position drugs for ideal indications and patients, allow for better analog design, and explain observed adverse events. There exist a number of experimental approaches for target identification ranging from affinity pull-downs to genome-wide knockdown screens [93,94], but these approaches are

[*] Madhukar NS[a], Khade PK[a], Huang L, et al. A New Big-Data Paradigm For Target Identification And Drug Discovery. 2017 (Submitted/Under Review) ([a] = co first authors)

labor, resource, and time intensive, not to mention failure prone. Computational target prediction has the potential to substantially reduce the work and resources needed for drug target identification. Existing computational methods traditionally fall into three major categories: ligand-based, molecular docking, and data driven. Ligand-based approaches take known binding targets for a given drug and attempt to find other proteins that are sufficiently similar to the known targets [9,95]. These similar proteins are then predicted as novel targets. However, to achieve high predictive power they require a large input of known binding partners for each tested drug, and therefore can only be used on drugs that have prior comprehensive target information [9,95]. Because of this, these methods are often not broadly applicable, especially to orphan molecules – molecules with no known binding targets. On the other hand, molecular docking uses simulations of small molecules interacting with proteins to model if and how a drug may bind a given protein [96,11]. However, this approach requires significant computational power and complex 3D structures for each queried protein – data that is often unavailable.

Traditionally, data-driven methods have focused on a single aspect out of a small molecule's activity in a biological system. Wang et al. [97] used post-treatment gene expression changes to predict drugs with shared targets [74,98]. Another method relied on side-effect similarity between drugs with known targets to predict new drug-protein interactions [99]. However, this method was restricted to the small subset of small molecules that had been clinically tested and had thorough side effect annotation. Though each of these methods represents a significant advancement in the field, they all suffer from either lack of accuracy or broad utility – evidenced either by an inability to reliably validate target predictions, or by their limited applicability to a small subset of all small molecules. This is not surprising though, as past research has demonstrated that these individual datasets are noisy, thus, it is expected that reliance on any single data type will lead to low predictive power [100-102].

Additionally, other groups have shown how the combination of multiple types of data can improve the calculation of drug-drug similarities[103] and adverse event prediction[104], yet, this type of combinatorial approach has not been fully explored for drug-target prediction. The few reported studies using combinatorial approaches for drug-target prediction, suffer from significant limitations that minimize their impact in the field. These limitations include the use of gene-based similarity features, a method inherently biased against the discovery of diverse types of targets (favoring instead, the discovery of genes of the same class as the known drug-targets), the small number of drugs used in the study (<500), or lack of experimental target validation[105-107]. To overcome these limitations, we introduce BANDIT, a novel drug-target prediction platform. BANDIT achieves unprecedented target-identification accuracy, without any reliance on gene-based similarities (making it broadly applicable to newly discovered compounds), uncovers novel targets for the treatment of cancer, and can be used to quickly pinpoint potential therapeutics with novel mechanisms of action to accelerate drug development.

RESULTS

A novel combinatorial Big-Data Approach leads to a large increase in predictive power

In the age of “Big Data” there has been an explosion of techniques that permit genomic, chemical, clinical, and pharmacological measurements to characterize a small molecule’s mechanism. Many such measurements are either already published or are reasonably straightforward to perform. We hypothesized that integrating the multiple, independent pieces of evidence provided by each data type into a cohesive prediction framework would dramatically improve target predictions. To test this hypothesis, we developed **BANDIT**: a **B**ayesian **AN**alysis to determine **D**rug **I**nteraction **T**argets. BANDIT integrates over 20,000,000 data points from six distinct data types – drug efficacies[108], post-treatment transcriptional responses [74,98], drug structures [109,110], reported adverse effects [111], bioassay results [109,110], and known targets [112,113] – to predict drug-target interactions. This underlying database contains information on approximately 2,000 different drugs with 1,670 different known targets and over 50,000 unique orphan compounds (compounds with no known targets).

For each data type we calculate a similarity score for all drug pairs with known targets. Since each dataset uses a distinct reporting metric, the similarity calculation was specific to the data type being considered (**Figure 3.1)**. Previous approaches have argued that high similarity in one feature indicates high similarity in others, implying that only one or two data types are sufficient for target prediction since others can be inferred [114]. However, using our vastly expanded dataset, we found little overall correlation across different similarity scores (**Figure 3.2A**). These results suggest that each data type is measuring a distinct aspect of a molecule's activity and that individual features for a given drug cannot be extrapolated based on other data types. This shortcoming further supported our hypothesis that a novel approach that integrates independent data types could significantly improve target prediction accuracy.

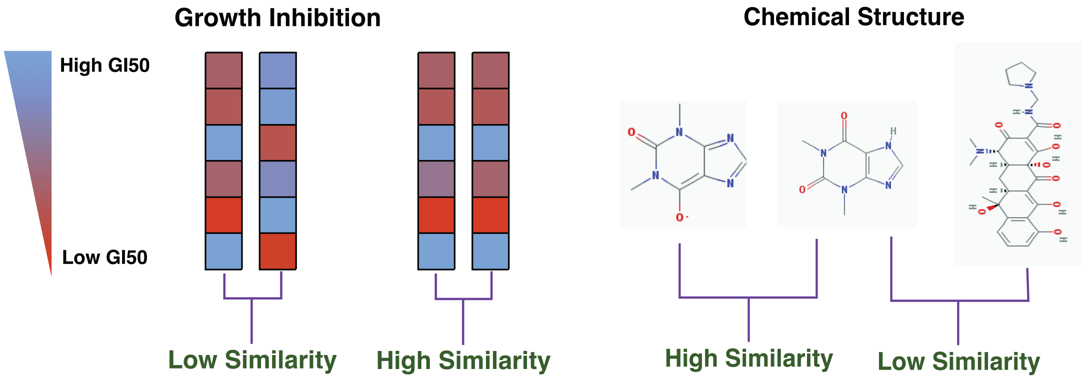

Figure 3.1 – Examples of similarity score calculations for growth inhibition data and chemical structures

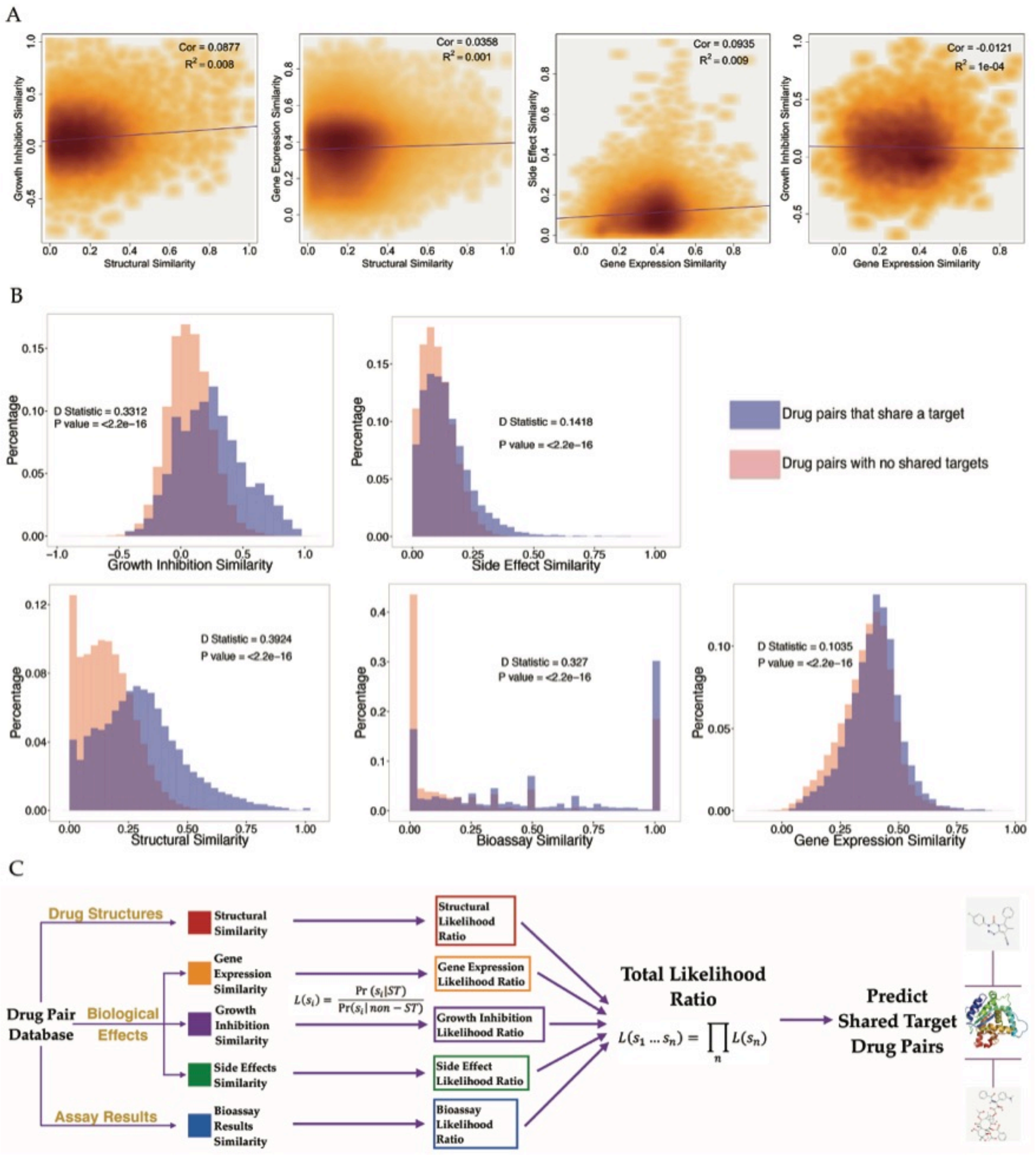

Figure 3.2 – BANDIT exploits both the independence and individual predictive powers of each data type – A) Density plots showing how various different similarity scores correlate with one another, with darker area corresponding to a higher density of values. R^2 and P value were calculated using a pearson correlation. B) Distributions of similarity scores across two sets – drug pairs known to share a target and those with no known shared targets. P values and D statistics were calculated using the Kolmogorov-Smirnov test. C) Schematic of BANDIT's method of integrating multiple data types to predict shared target drug pairs.

We next separated drug pairs into those that shared at least one known target (>34,000 pairs) and pairs with no known shared targets (>1,250,000 pairs). We applied a Kolmogorov-Smirnov test to each similarity score and used the associated D statistic to calculate the degree a given data type could separate out drug pairs that shared targets (**Figure 3.2B**). We found that all features were able to significantly separate the two classes ($P < 2e-16$), and structural similarity was found to be the most discriminative among all features evaluated ($D_{Structure} = 0.39$). Additionally, we discovered that similarity across an unbiased set of bioassays and the relatively simple NCI-60 growth inhibition screen could strongly differentiate shared target drug pairs ($D_{Bioassay} = 0.327$ & $D_{GI50} = 331$), while, surprisingly [115,97,99], transcriptional responses ($D_{TResponse} = 0.1$) and reported adverse effects ($D_{SideEffect} = 0.14$) were much weaker differentiators. This information not only identifies the strengths of each data type, but will also allow researchers to efficiently prioritize experiments when faced with limited resources.

For every drug pair, BANDIT converts each individual similarity score into a distinct likelihood ratio. These individual likelihood ratios are then combined within a Naïve Bayes framework to obtain a total likelihood ratio (TLR) that is proportional to the odds of two drugs sharing a target given all available evidence (**Figure 3.2C**). We calculated TLRs for all possible drug pairs with known targets and the output was evaluated using 5-fold cross validation. We observed an Area Under the Receiver Operating Curve (AUROC) of 0.89 – higher than any competing approach [99,114]– demonstrating that BANDIT's integrative approach can accurately identify drugs that share targets. We recomputed the AUROC while varying the number of included data types and observed an overall increase in predictive power as we added new data types

(**Figure 3.3A**). Furthermore we observed a steady increase in predictive power regardless of the addition order. This result verified the power of BANDIT's "Big Data" approach and demonstrated how separate information sources can be combined to yield predictions more powerful than those obtained from any individual source. This was confirmed using the KS test where we saw that the TLR output could better separate shared target drug pairs than any individual similarity score with a drastic increase in performance when focusing on drug pairs with all 5 data types (D_{TLR} = .69). Furthermore, we observed that BANDIT's ratio of true to false positives continually increased as we raised the cutoff value, indicating that BANDIT's TLR output is a dynamic value that estimates the strength and confidence level of a specific prediction and can effectively pick out high quality shared-target predictions (**Figure 3.3B**).

BANDIT can replicate the results of experimental screens and predict specific target interactions

We next investigated how we could use BANDIT to replicate results from published experimental screens. Peterson et al. [116] tested 178 known protein kinase inhibitors against a panel of 300 different kinases and measured the level of inhibition (in terms of percent remaining kinase activity) for each inhibitor-kinase pair. We examined all orphan molecules – molecules with no known targets – in both the Peterson kinase database and BANDIT's, and, used BANDIT to predict potential kinases targets for each orphan molecule. We observed that the kinase targets BANDIT predicted for each orphan molecule had higher levels of reported inhibition in the Peterson dataset than non-predictions ($p<1e-5$; **Figure 3.4**). This result supports using BANDIT to guide experimental screens while minimizing operational costs.

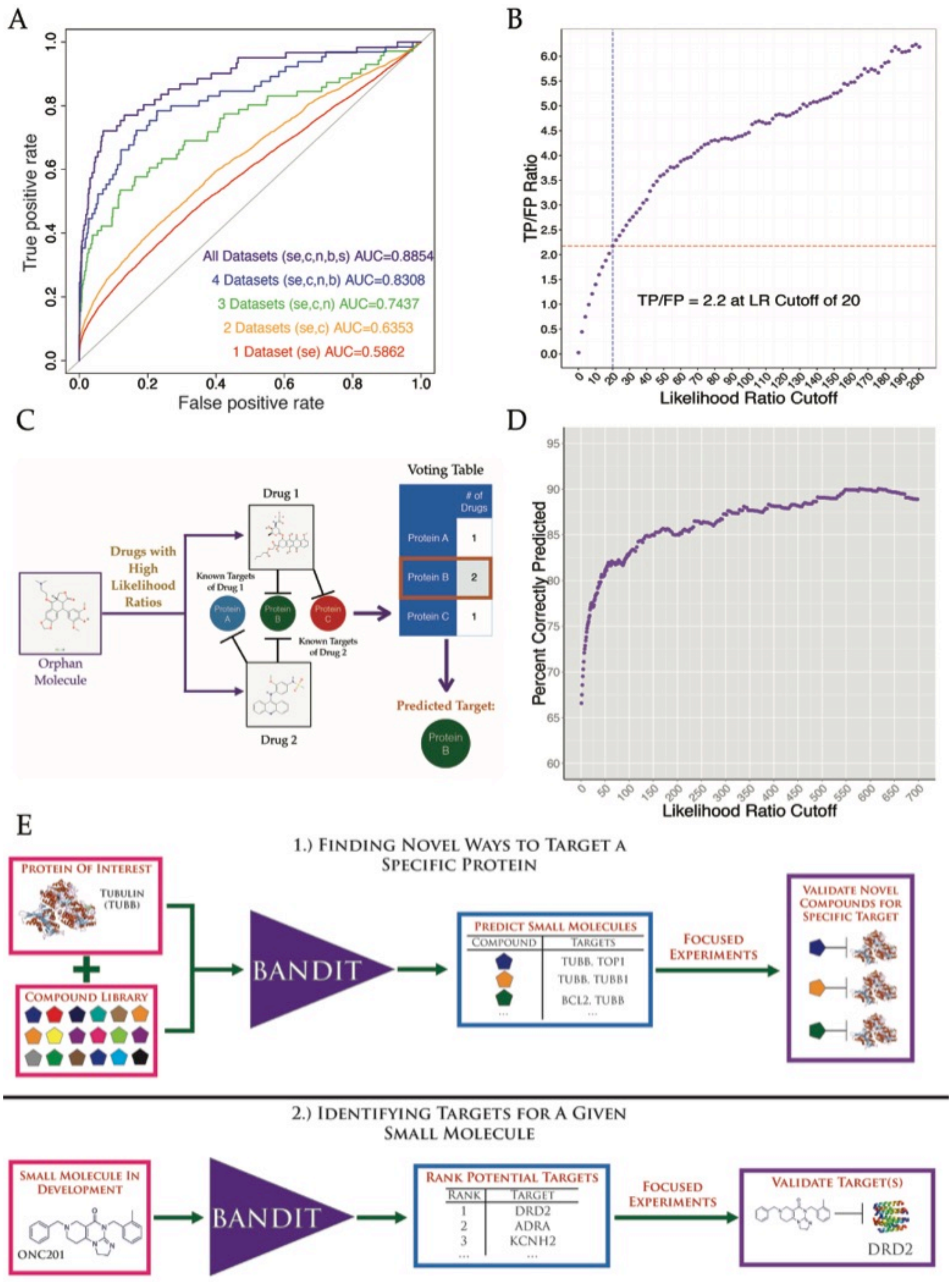

Figure 3.3 – BANDIT can accurately predict shared targets and specific target interactions – A) Area under the receiver-operating curve for different sets of data types. SE = Side effects; C = CMap; N = NCI60; B = Bioassays; S = Structure. B) Ratio of true positives to false positives at different likelihood ratio cutoffs. C) Schematic of the BANDIT voting schematic for predicting specific target interactions. D) Accuracy level of BANDIT's voting algorithm at various likelihood ratio cutoffs E) Schematic of two proposed operating scenarios for BANDIT

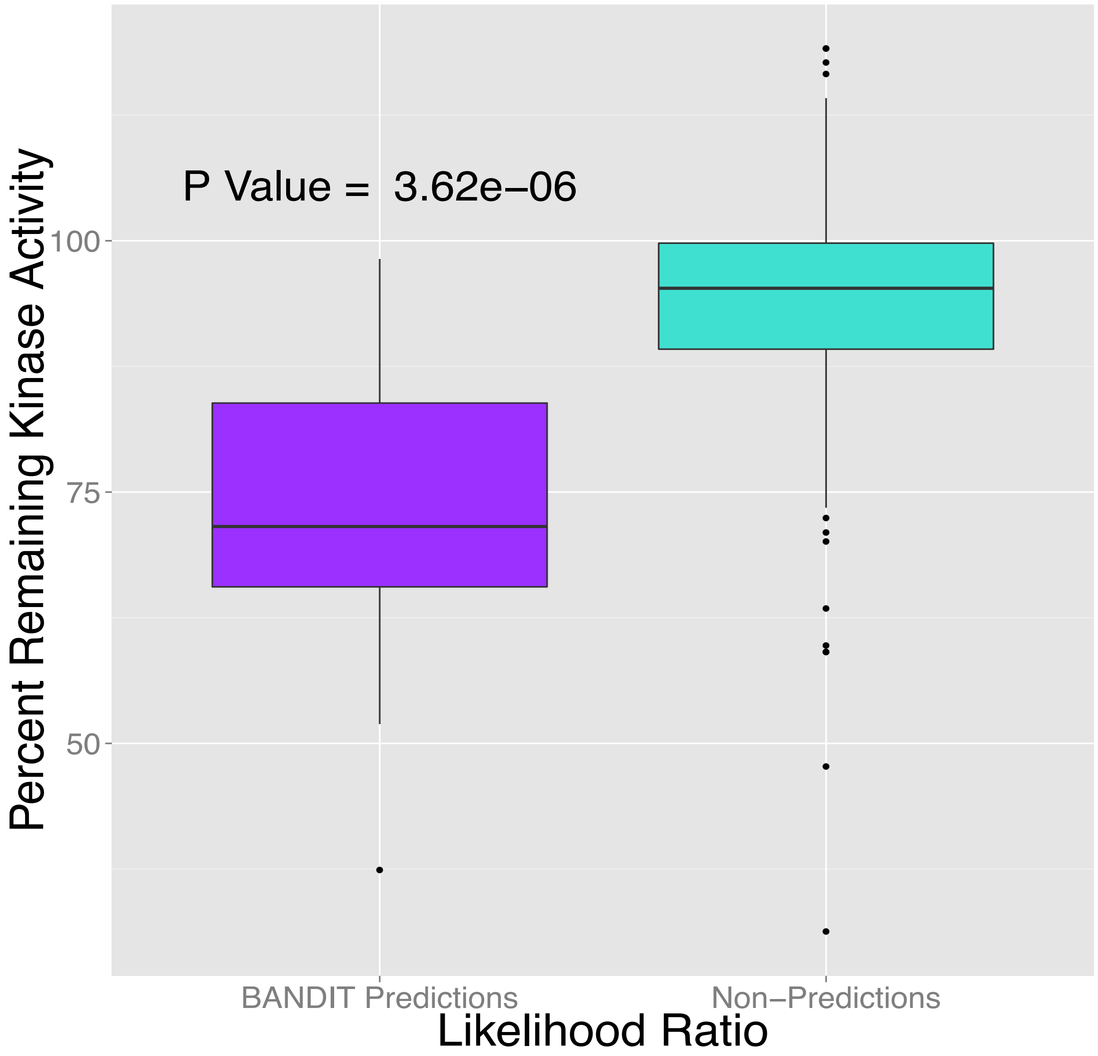

Figure 3.4 – BANDIT can replicate results from an experimental kinase screen – Boxplot showing the distributions of "% inhibition" across BANDIT predictions and non-predictions. P value was calculated using a Mann Whitney test.

Moving forward from shared-target predictions, we examined whether for a given drug BANDIT could be used to predict a specific binding target from our database of over 1,600 unique proteins. We hypothesized that if a protein appeared as a known target in a large number of shared target predictions, then it is likely a target for the tested orphan molecule. To test this hypothesis, we developed a “voting” algorithm to predict specific targets for each orphan small molecule by identifying any recurring targets (**Figure 3.3C**)**.** We applied our voting method to all drugs in our database with known targets and demonstrated that as we required more stringent TLR values for a pair of drugs to be predicted to share a target, the accuracy level – measured by whether BANDIT correctly identified a known drug target – steadily increased (**Figure 3.3D**). The accuracy level eventually reached ~90%, demonstrating that BANDIT could be used to accurately identify specific targets for a diverse set of small molecules.

We then used BANDIT to predict novel targets for 14,168 small molecules with no known targets or mechanisms of action in our database. We confidently predicted targets for 4,167 unique small molecules (30% of our original set), with predictions spanning over 560 distinct protein targets. By setting a higher TLR cutoff for predictions and requiring a higher number of “votes” for any predicted targets, we further narrowed this list to 720 high confidence target predictions. To date, this is the largest database of novel drug-target predictions (nearly double the number of drugs in DrugBank’s drug-target database) and this list can be interrogated further to discover novel therapeutics and small molecules for a target of interest. Based on this success, we envisioned two main operating scenarios for BANDIT: 1) Using BANDIT in combination with the library of orphan small molecules to identify new small molecules targeting a specific protein and 2) to integrate BANDIT directly into the drug development pipeline to predict targets and guide experiments for drugs currently in development (**Figure 3.3E)**.

Discovery of Novel Microtubule-Targeting Compounds Capable of Overcoming Drug Resistance

Beginning with the first operating scenario, we used BANDIT to identify novel ways to target microtubules. Anti-microtubule drugs make up one of the largest and most widely used classes of cancer chemotherapeutics, with tubulin being one of the most validated anticancer targets to date [117-120]. Interestingly, and unlike most classes of cancer chemotherapy drugs or targeted-therapies in oncology, microtubule inhibitors are further sub-categorized as microtubule-stabilizing (e.g. taxanes) and microtubule-depolymerizing drugs (e.g. vinca alkaloids). Each class shifts the cellular equilibrium that normally exists between soluble tubulin dimers and microtubule polymers, towards microtubules (taxanes) or soluble tubulin (vinca alkaloids). Despite the clinical success of the entire class of microtubule inhibitors, the development of drug resistance – which is the number one cause of cancer mortality in metastatic patients – along with the presence of toxic side effects limits their clinical applicability [121]. Hence, the discovery of novel microtubule-targeting small molecules could significantly improve cancer therapy by identifying compounds with activity on refractory tumors or compounds with less toxic side effects. To this aim, we further focused our list of high confidence orphan-target predictions to small molecules predicted to target microtubules. To see how our novel predictions related to known microtubule-targeting therapeutics, we created a network of all known and predicted anti-microtubule small molecules with edges representing a predicted shared target interaction (**Figure 3.5**). Interestingly we found that the 14 known microtubule-targeting agents tended to cluster together based on their distinct mechanism of action. For instance, we observe Paclitaxel clustering with Cabazitaxel and Docetaxel – all known microtubule-stabilizing drugs – while Colchicine clustered with other known microtubule-destabilizing drugs such as Podophyllotoxin. This is especially exciting since it demonstrates the potential for BANDIT to be used not only to identify a specific target for an orphan molecule but also to differentiate between different modes of action on the same target.

From our list of top anti-microtubule drug predictions we obtained a set of 24 compounds with varying structures for experimental testing. We chose the human breast cancer MDA-MB-231 cells for the validation experiments as microtubule-inhibitors (both stabilizing and destabilizing) are commonly used in the treatment of breast cancer patients. Cells were treated for 6 hours with 1 and 10 µM of each small molecule, and the integrity of the microtubule cytoskeleton (assessed by confocal microscopy following tubulin immunofluorescence), was used as the bio-assay endpoint. Our results showed that 16 of the 24 orphan small molecules exhibited significant effects on microtubules (**Figure 3.6A-F**), a much higher success rate (67%) than one would expect by chance ($p < 2e-16$). To more accurately quantify the extent of drug-target engagement, we employed a second biochemical assay quantifying the effect that each small molecule exerted on the equilibrium between microtubule polymers and soluble tubulin, following 6 hours of treatment. Our results confirmed and corroborated the microscopy results, further revealing that while several small molecules had maximal microtubule-inhibitory activity

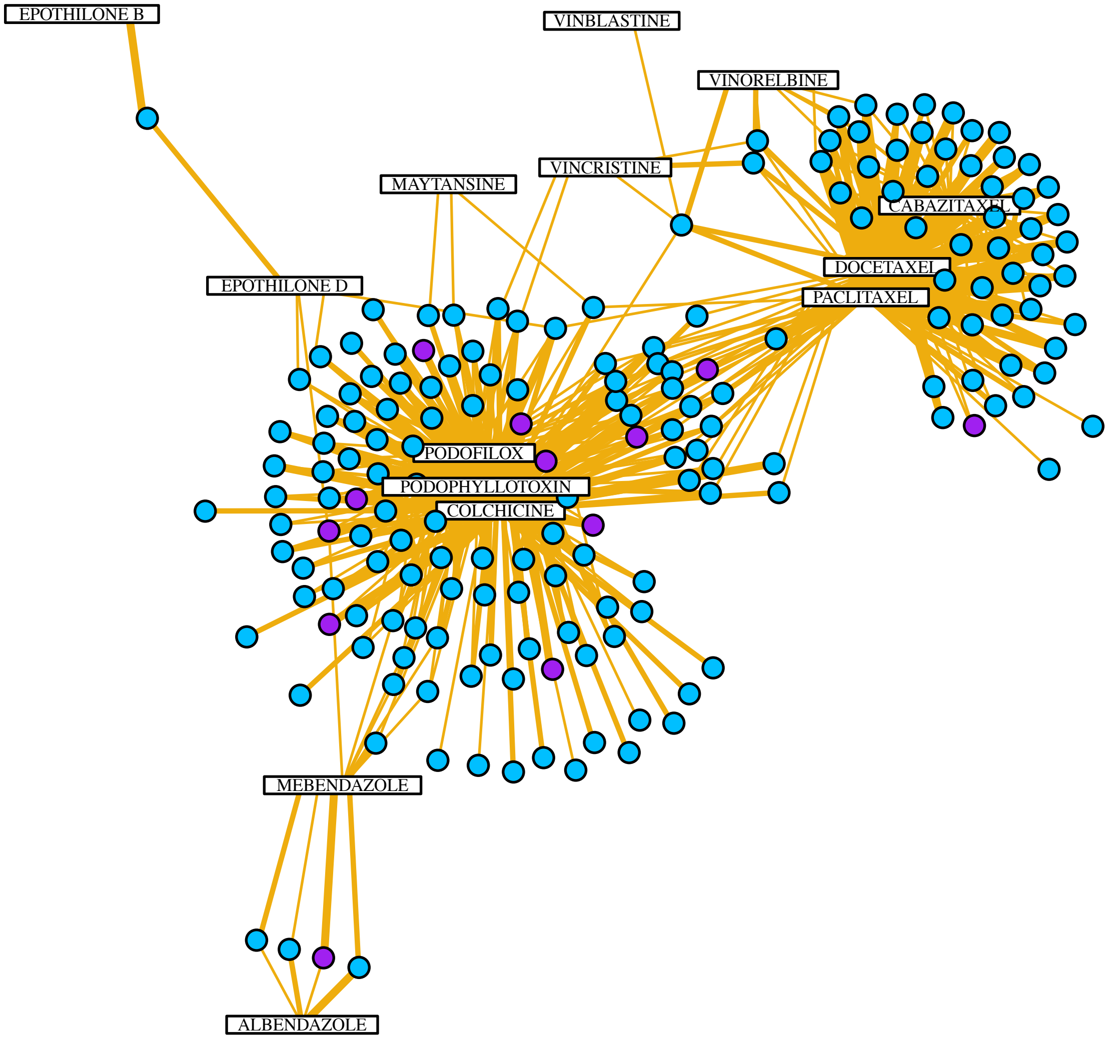

Figure 3.5 – Using BANDIT known microtubule inhibitors cluster based on mechanism of action – Network of known microtubules inhibitors and orphan molecules predicted to target molecules. Named boxes represent known inhibitors, blue circles represent predicted inhibitors, and purple circles represent predicted inhibitors that were validated experimentally.

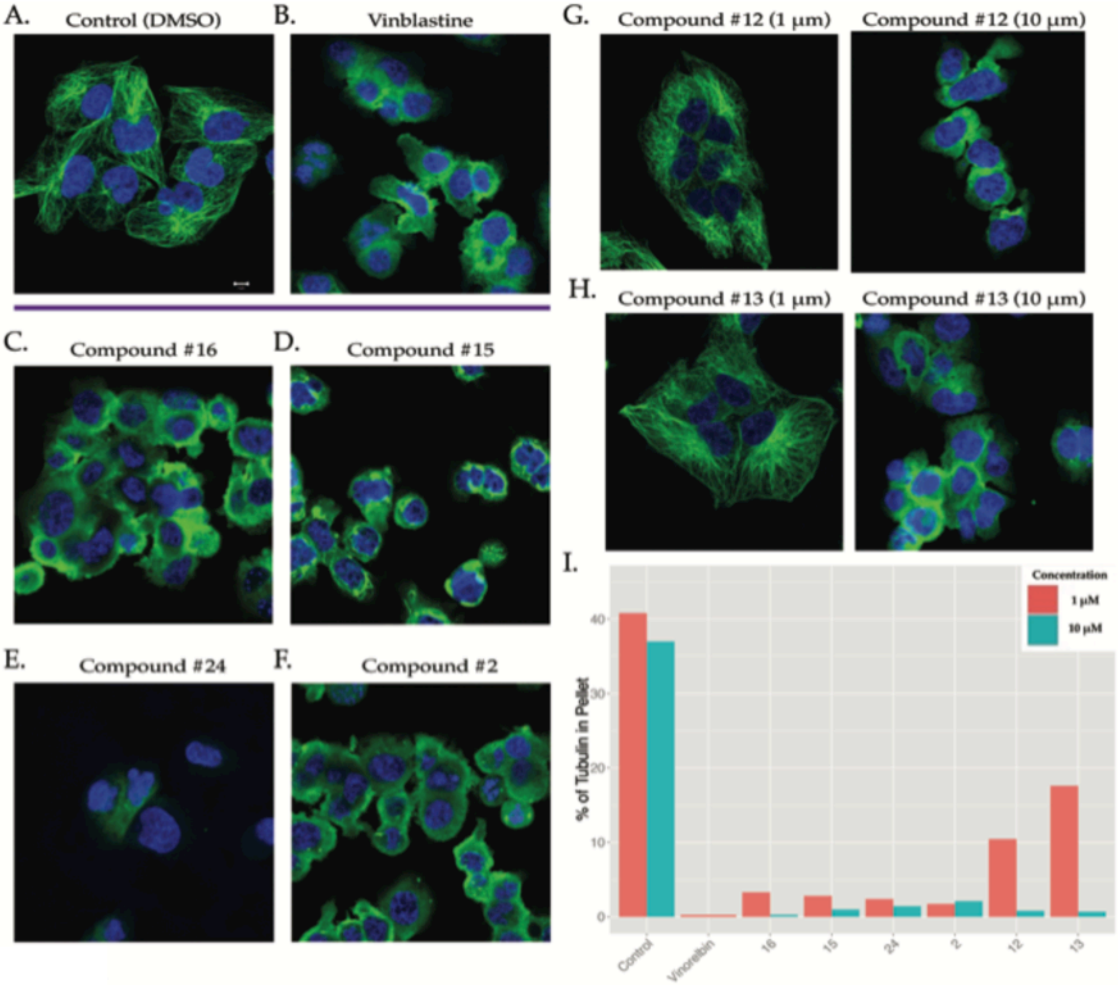

Figure 3.6 – Microtubules are a correct target of the newly identified small molecules – Effect of various compounds (1µM) on the microtubule integrity of MDA-MB-231 cells after 6 hours of treatment. A) Control with DMSO (Scale bar: 5 µm), B) Vinblastine as a positive control, C) Compound #16, D) Compound #15, E) Compound #24 F) Compound #2. G) Dose dependent effect of Compound #12 and H) Compound #13. I) Bar graph showing the % tubulin in the pellet compared to the supernatant (averaged over three independent replicates) for depolymerizing drugs at 1 and 10 µM.

at the lowest dose (1μM) (**Figure 3.6C-F**), others exhibited a dose-dependent effect on microtubule depolymerization (e.g. compounds #12, #13), further establishing microtubules as their bona-fide target (**Figure 3.6G-I**). Taken together, these experiments confirmed the predicted targets and mechanism of action for the majority of the newly identified microtubule inhibitors. While further testing will be needed before these small molecules can be used clinically, these results do demonstrate BANDIT's target prediction accuracy and how it can be used on compound libraries to identify small molecules acting with a specific mode of action on specific targets, for further investigation.

To inform future clinical development for these newly identified microtubule inhibitors, we next tested their activity against drug resistant models. Drug resistance remains one of the most challenging areas in clinical oncology, affecting both broad chemotherapy drugs and targeted-therapies. In the case of microtubule inhibitors, overcoming drug resistance is even more challenging as the mechanisms are often multifactorial. As previously demonstrated, BANDIT can accurately identify a set of structurally diverse small molecules that all bind a common target (in this case microtubules), therefore we next investigated whether any of our newly identified microtubule-depolymerizing small molecules could successfully act on tumors resistant to other known anti-microtubule drugs. Using the 1A9 human ovarian carcinoma cell line – which has previously been used successfully in selecting microtubule-inhibitor resistant clones and for high throughput small molecule screening, [122-126] – we created clones resistant to Eribulin mesylate, a microtubule depolymerizing drug that is FDA approved for the treatment of docetaxel-refractory breast cancer patients [127,128] (**Figure 3.7A)**. Interestingly, recent clinical data demonstrated that fewer than 50% of breast cancer patients showed any detectable response after treatment with Eribulin, further highlighting the importance of finding new molecules that share the same validated target but are active against the large population of refractory patients [129]. Our results, using 72-hr cytotoxicity assays showed that the Eribulin-resistant 1A9 cells (1A9-ERB) were more than 7,000 –fold more resistant to Eribulin than the parental cells and exhibited cross-resistance to all classes of clinically used microtubule-depolymerizing drugs (**Table 3.1**). To test whether the drug-resistance phenotype was due to impaired drug-target engagement, we treated parental and resistant cells for 6 hr only with 1uM of Eribulin or each of the FDA-approved depolymerizing drugs. Consistent with their drug resistance phenotypes, our results showed lack of drug-induced microtubule depolymerization in 1A9-ERB cells in contrast to the complete depolymerization observed in the microtubule network of drug-sensitive 1A9 parental cells (**Figure 3.7B-C**). These on-target drug efficacy results are in agreement with the lack of antitumor activity revealed by the cytotoxicity data further highlighting the importance of discovering novel small molecules that could act on these refractory tumors. We tested the top 4 performing small molecules (#15, 16, 24, and 2) on the 1A9-ERB cells and found that 3 out of 4 compounds tested, were active against the 1A9-ERB cells and effectively depolymerized microtubules, as evidenced by the diffuse soluble tubulin staining following drug treatment (**Figures 2.7E-F**), in contrast to the fine and intricate microtubule network observed in untreated cells (**Figures 2.7A**).

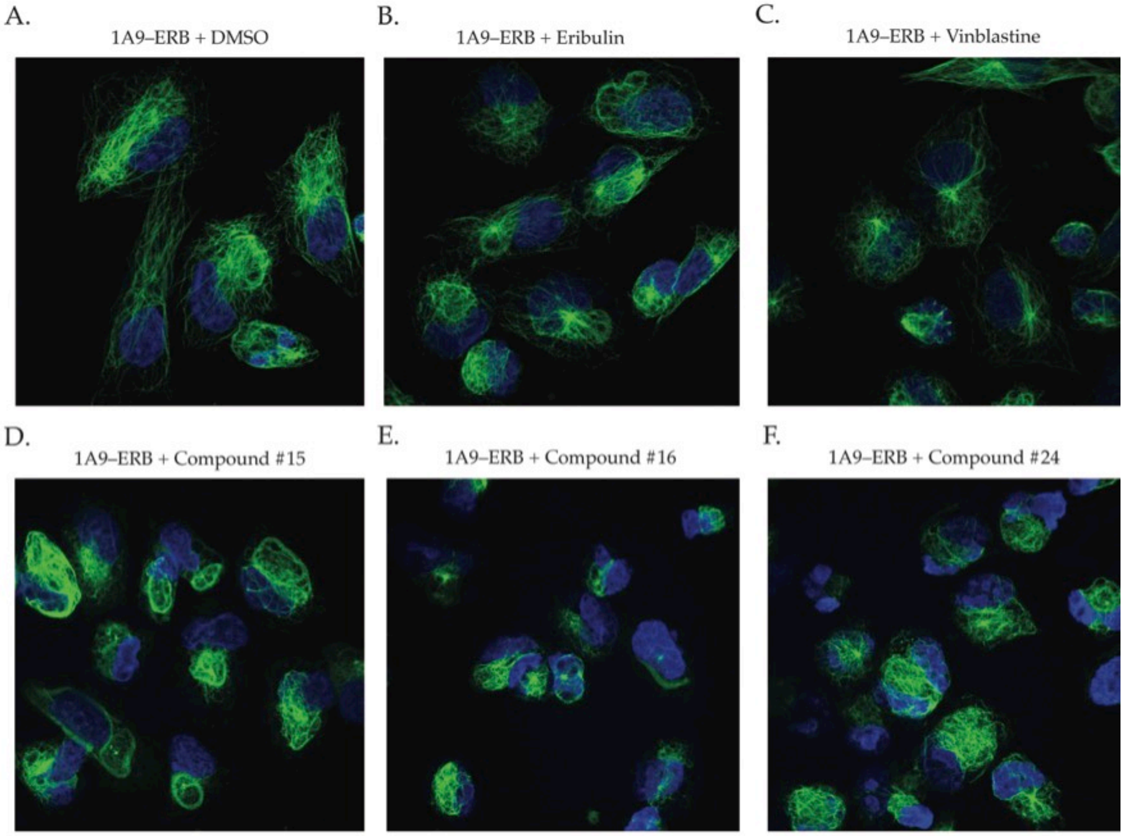

Figure 3.7 – A set of the BANDIT predicted small molecules can act on cells resistant to Eribulin and other microtubule depolymerizing drugs – Effect of various compounds on the microtubule integrity of 1A9-ERB cells after 6 hours of treatment: A) Control with DMSO (Scale bar: 5 µm), 100nM of B) Eribulin and C) Vinblastine, and 1µM of D) Compound #15, E) Compound #16 and F) Compound #24.

Table 3.1 – Newly identified compound can reverse resistance in cytotoxic assays – Cytotoxic activity of drugs tested in a 72-hr anti-proliferative assay against parental 1A9 and eribulin resistant (1A9-ERB) cells. IC50 (nM) values for each drug indicate the concentration that kills 50% of the cells in 72 hr

Drug	1A9 (nM)	1A9-ERB (nM)	Fold Resistance[a]
Eribulin	0.34	2397	7050
Vinblastin	0.08	208	2600
Colchicine	11	560	51
Drug No. 15	20	86	4.3

a=Ratio of IC50 values of resistant/parental cells (1A9-ERB/1A9)

Compound No 15, which was the most active of the 4 compounds, was tested using cytotoxicity assays and was found to almost completely reverse drug-resistance from 7050-fold observed with Eribulin down to 4-fold (**Table 3.1**). While further *in vitro* and *in vivo* studies are required for the clinical development of these compounds, these results clearly demonstrate BANDIT's utility in identifying lead small molecules with potential activity against drug resistance tumor models without the labor-and cost-intensive physical screening of thousands of small molecules. Even though BANDIT is "trained" using a database of drugs with known targets and mechanisms, our results show that it can accurately identify small molecules with distinct modes of action from any known drugs in the training set. This also highlights how BANDIT can pinpoint small molecules from large compound libraries with unique mechanisms that could potentially act on drug resistant cells. Compounds such as these could represent the next generation of clinically developed drugs reducing the need for extensive medicinal chemistry and structure-activity studies, therefore, expediting drug development.

BANDIT Uncovers Selective Antagonism of DRD2 by Anti-Cancer Small Molecule ONC201

Given BANDIT's demonstrated capability to accurately identify specific targets for orphan small molecules, we next investigated how we could integrate BANDIT directly into the drug development pipeline and test its ability to predict targets for small molecules with promising clinical activity but without a specific target. Therefore we applied BANDIT to ONC201– a small molecule discovered in a phenotypic screen for p53-independent inducers of TRAIL-mediated apoptosis – currently in multiple phase II clinical trials for select advanced cancers. Despite its promising preclinical and early clinical anticancer activity and its reported effects on a few signaling pathways, including Akt/ERK pathway [130-132], a bona-fide target for this compound remains elusive.

To identify direct binding targets for ONC201, we used BANDIT to compute likelihood ratios between ONC201 and all drugs with known targets in BANDIT's database. BANDIT's top shared target prediction were between ONC201 and Oxiperomide and Thioridazine, both a dopaminergic antagonists previously used the treatment of dyskinesias and schizophrenia respectively [133-136]. Interestingly, our voting analysis indicated that the most likely targets of ONC201 were dopamine receptors – specifically DRD2 – and adrenergic receptor alpha (**Figure 3.8A**), both of which are members of the G-protein coupled receptor (GPCR) superfamily.

To test these predicted targets we performed in vitro profiling of GPCR activity using a hetereologous reporter assay for arrestin recruitment, which is a hallmark of GPCR activation[137]. Our results indicated that ONC201 selectively antagonized the D2-like (DRD2/3/4L), but not D1-like (DRD1/5L), subfamily of dopamine receptors (**Figure 3.8B, Figure 3.9A**), with no observed antagonism of other GPCRs under the evaluated conditions. Among the DRD2 family, ONC201 antagonized both short and long isoforms of DRD2 and DRD3, with weaker potency for DRD4. Further characterization of ONC201-mediated antagonism of arrestin recruitment to DRD2L was assessed by a Gaddam/Schild EC50 shift analysis, which determined a dissociation constant of 2.9 uM for ONC201 that is equivalent to its effective dose in many human cancer cells (**Figure 3.8C**).

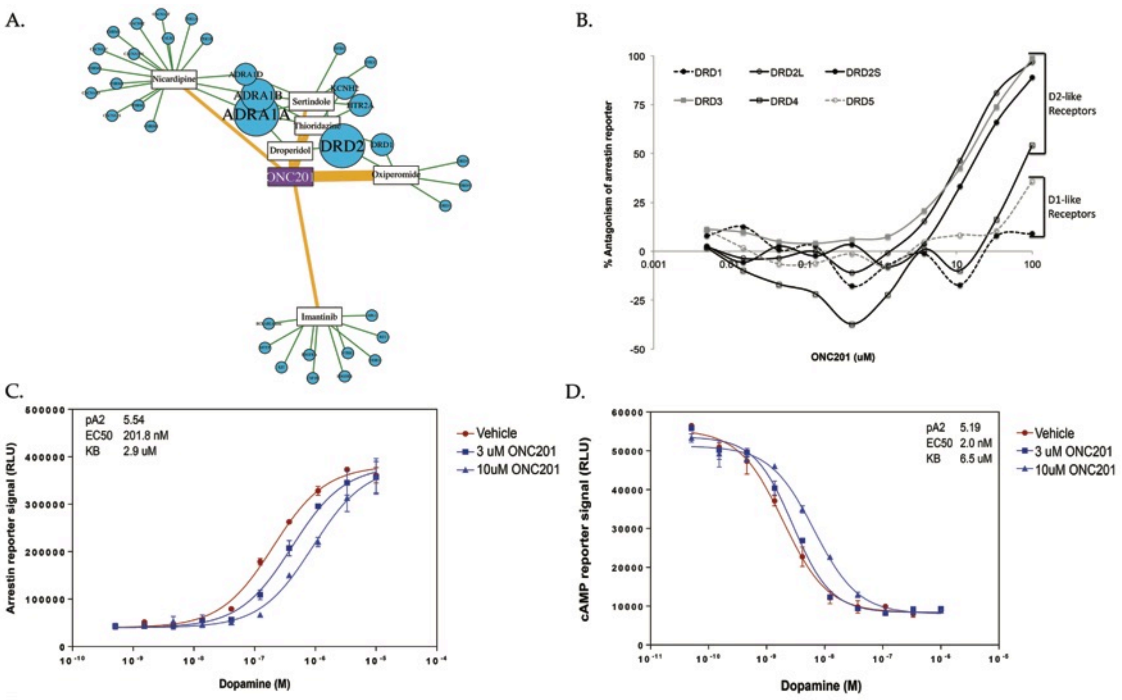

Figure 3.8 – ONC201 is a selective DRD2 antagonist – (A) BANDIT target predictions for ONC201. Connections between ONC201 and known drugs are weighted based on the likelihood ratio and predicted targets are sized based on the prediction strength. (B) Antagonism of ligand-stimulated dopamine receptors by ONC201. C) Schild analysis of DRD2L antagonism by ONC201 using arrestin recruitment or (D) cAMP modulation reporters.

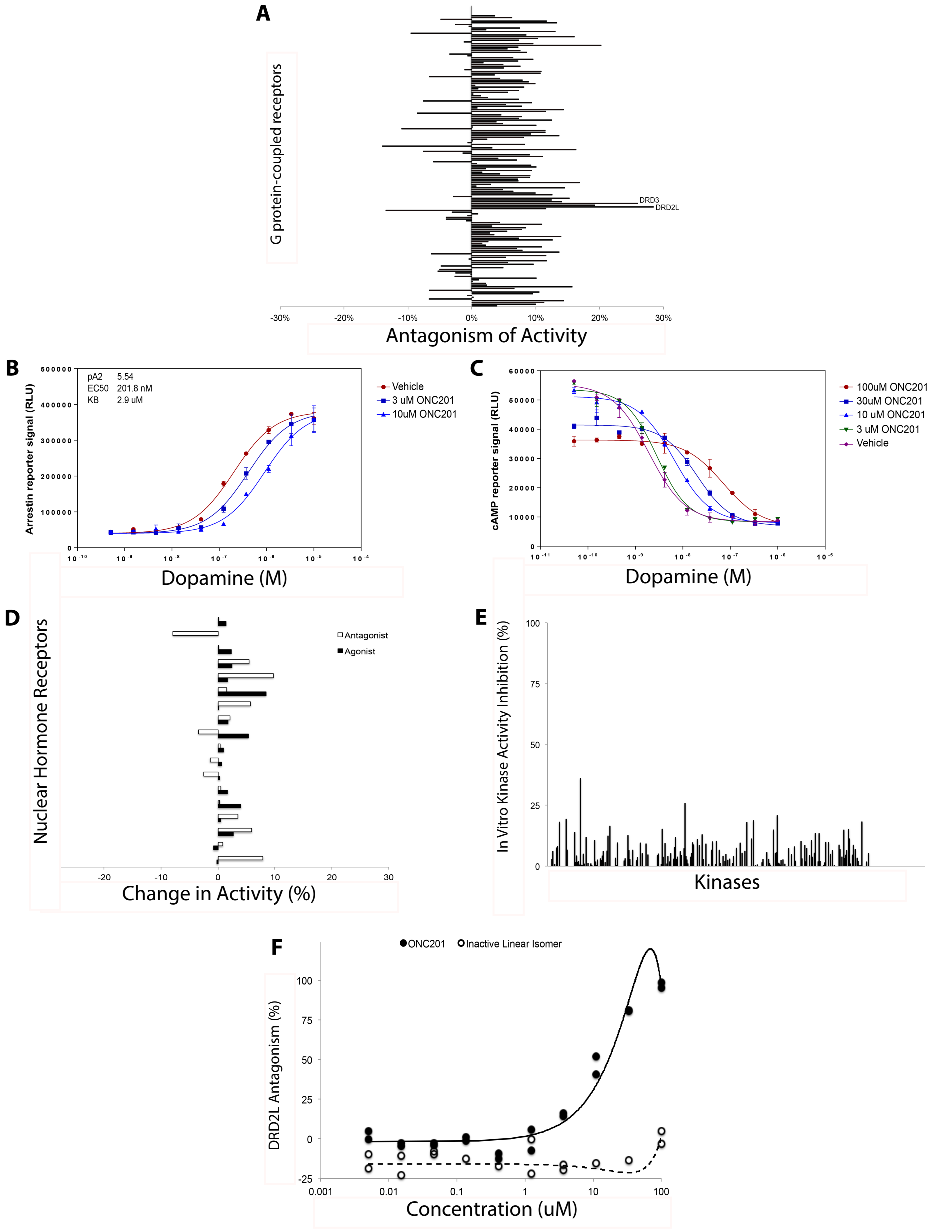

Figure 3.9 – The antagonism of DRD2 by ONC201 is highly specific across GPCRs and other cancer drug targets – A) Antagonism of GPCRs using an arrestin recruitment reporter assay (10μM). B) Competition of ONC201-mediated antagonism of DRD2L by dopamine in arrestin recruitment or (C) cAMP modulation reporters. D) Antagonism or agonism of nuclear hormone receptors by ONC201 (2 or 20uM) using a nuclear translocation reporter assay. (E) Inhibition of in vitro kinase enzymatic activity by ONC201 (1 uM). F) DRD2L antagonistic activity of ONC201 or a linear constitutional isomer of ONC201 that has no biological activity using an arrestin recruitment reporter assay.

Confirmatory results were obtained for cAMP modulation in response to ONC201, which is another measure of DRD2L activation (**Figure 3.8D**). The ability of dopamine to completely reverse the dose-dependent antagonism of up to 100uM ONC201 suggests direct, competitive antagonism of DRD2L (**Figure 3.9B-C**). In agreement with the specificity of ONC201 for the target predicted by BANDIT, no significant interactions were identified between ONC201 and nuclear hormone receptors, the kinome, or other drug targets of FDA-approved cancer therapies (**Figure 3.9D-E**). Interestingly, a biologically inactive constitutional isomer of ONC201 [138]) did not inhibit DRD2L, suggesting that antagonism of this receptor could be linked to its biological activity (**Figure 3.9F**). In summary, these studies establish that ONC201 selectively antagonizes the D2-like subfamily of dopamine receptors, which is an "unconventional" target for oncology drugs and further demonstrate BANDIT's ability to act as a tool to advance drug development.

This unexpected discovery on DRD2L being a direct-binding target for ONC201, has also led to the design and launch of a clinical trial of ONC201 in pheochromocytomas, owing to high levels of DRD2L expression in this rare tumor type. Taken together, these results demonstrate the potential of BANDIT to expedite drug development by using drug-target engagement predictions in combination with gene expression to enable the identification of select patient and indications groups more likely to benefit from a particular drug treatment.

BANDIT can determine drug mechanisms and can help understand the drug "universe"

Following validation that BANDIT could accurately determine the specific targets for small molecules, we then examined how it could also be used to understand the target binding mechanism, otherwise known as its mechanism of action (MoA). First we used BANDIT to test all known microtubule-targeting drugs, and created a hierarchical cluster based on their TLR outputs. We observed a clean separation between drugs known to destabilize microtubule depolymerizing and polymerizing agents (**Figure 3.10A**). A similar MoA-based clustering was observed when we tested all known protein kinase inhibitors, which showed a clear separation between receptor tyrosine kinase inhibitors, serine/threonine kinase inhibitors, and nucleoside analogs (**Figure 3.10B**). Overall these results demonstrate that BANDIT can be used to differentiate small molecules based on their specific MoA without additional model training. Combined with the earlier voting algorithm, this demonstrates an efficient pipeline for small molecule target and mechanism identification: first using BANDIT to predict targets for an orphan small molecule, followed by clustering with other drugs known to act on the same target to discern MoA.

We next used BANDIT to get an overview of how different classes of drugs, spanning the entire clinical landscape, may be related to one another. Based on the TLR between each drug pair, we constructed a network representative of the drug "universe," or all known drugs with at least one predicted shared target interaction (**Figure 3.10C**). Each drug was classified according to its 1st order Anatomical Therapeutic Chemical (ATC) classification – characteristic of the type and intended use of each drug. As expected, drugs of a similar ATC code cluster together, however we also observed many "unexpected" clusters indicative of drug mechanisms or effect. Interestingly, among all classes of cancer chemotherapeutics, microtubule inhibitors clustered together with

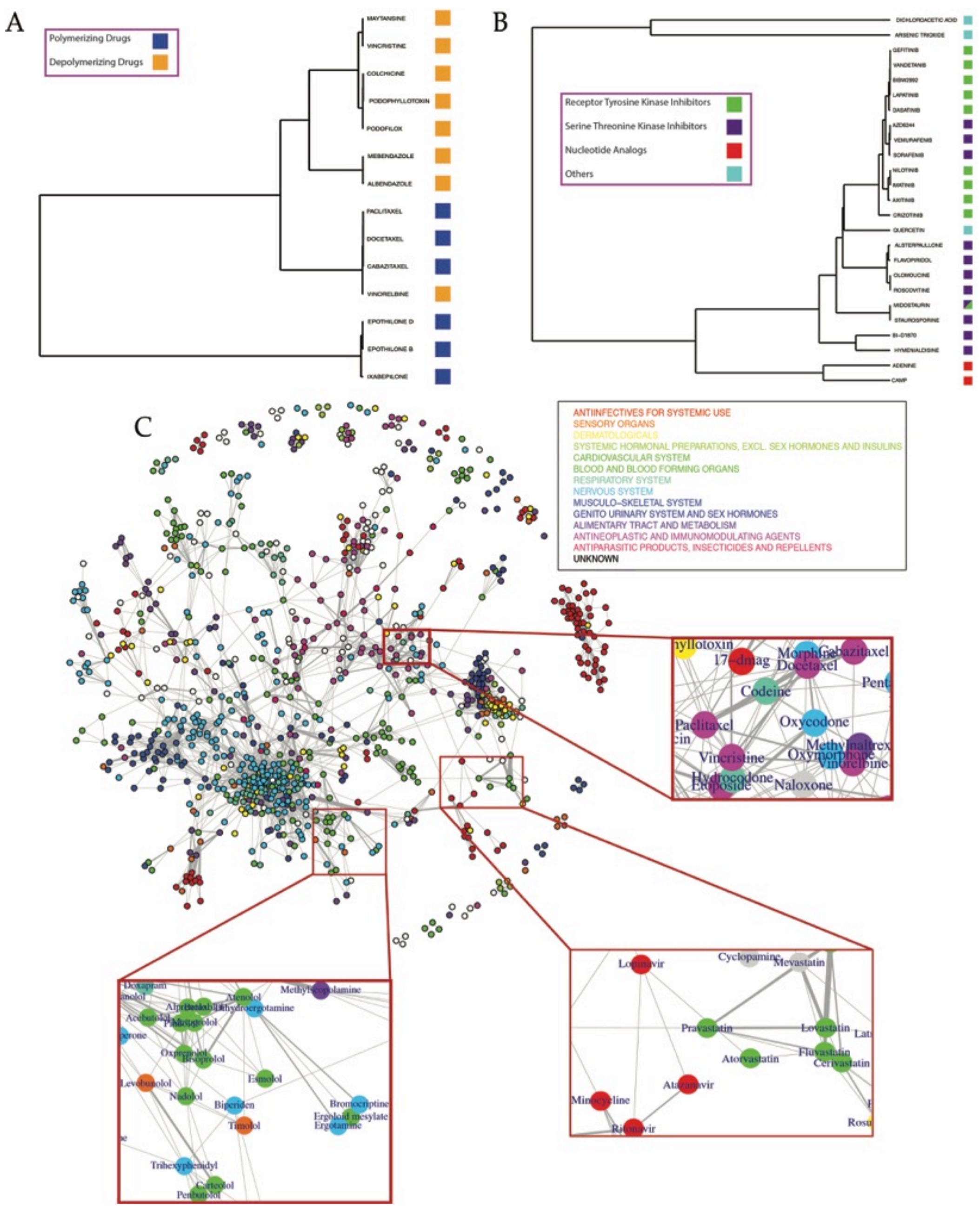

Figure 3.10 – BANDIT can predict specific mechanisms of action and connections between drug classes – A) Hierarchical clustering of drugs known to target microtubules and B) drugs known to target protein kinases. C) Network of drugs based on shared target interactions. Drugs are colored based on their most prevalent ATC code. Three specific clusters corresponding to beta-blockers and Parkinson's medications, anti-retrovirals and statins, and opioids and anti-microtubule drugs are highlighted.

camptothecin analogues, for which a dual role as topoisomerase I and tubulin polymerization inhibitors has been previously reported [139], but which is not widely acknowledged in clinical oncology. Conversely, we unexpectedly found opioids closely interconnected with microtubule targeting agents; this unanticipated observation is in line with previous reports showing how exposure to microtubule targeting drugs can increase the levels of the opioid receptor in rat cerebellums and that treatment of cardiac myocytes with opioids induces microtubule alterations [140,141]. This unexploited finding could reveal novel biology linking the opioid receptor-signaling pathway with the microtubule cytoskeleton, as well as potentially represent an example of drug repurposing, suggesting novel clinical indications for drugs already FDA-approved. As further proof of the clinical value of the broad universe clustering information revealed by BANDIT, we detected close clustering of known beta-blockers with many Parkinson's medications, which was especially interesting given that one of the most controversial clinical applications of beta-blockers was to reduce tremors in Parkinson's patients [142]. Drug clustering was also strongly indicative of potential side effects, as suggested by the link between antiretroviral medications, which often cause metabolic side effects like hypercholesterolemia, and statins, FDA-approved cholesterol lowering drugs [143]. Overall we believe this broad universe clustering approach could greatly advance future drug development by "indicating" novel synergistic drug combinations, cumulative side effects, and by assisting in drug repositioning.

Discussion

One of the strengths of the Bayesian framework is that it can easily accommodate new features, and, as we have observed, we expect that the addition of new data to only improve the overall performance. In addition, as more information becomes available there are many aspects of the current implementation that can be improved. For instance, we can better understand the dependencies between distinct data types and model those within our Bayesian network, and as more information on binding kinetics becomes available, BANDIT could be adapted to better predict on versus off-target effects. As drug development often stops in early clinical studies due to "unanticipated" toxic side effects, BANDIT could help overcome these roadblocks by identifying side effects due to unknown off-target bindings.

In summary, we have developed BANDIT, an integrative Big-Data approach that combines a set of individually weak features into a single reliable and robust predictor of shared-target drug relationships. Not dependent on complex 3D models or large known target cohorts, BANDIT can be used to predict shared target drugs and mechanisms of action for any drug or small molecule (over 50,000 in our database) which differentiates it from other target prediction approaches. By using the top shared-target predictions we can further predict with high accuracy specific targets for a given small molecule and demonstrate how BANDIT can be used to both efficiently discover new drugs with novel mechanisms for specific targets and identify targets for small molecules in the development pipeline – all without tedious, labor-intense and inaccurate drug screening approaches.

Our BANDIT predictions replicated shared-target relationships, individual drug-target relationships, and known mechanisms of action within our test set and replicated results of large-scale experimental screens. Moreover, we experimentally confirmed several of our novel predictions using different bioassays and model systems and demonstrated BANDIT's capability to efficiently discover novel small molecules, which could be used in refractory tumors. As the development of drug resistance is inevitable in oncology and applicable to both chemotherapy and targeted therapies, BANDIT has the potential to quickly and accurately identify drugs that can potentially overcome resistance and improve patient outcomes. Finally, BANDIT can be used on a broader scale to discern mechanisms of approved drugs, characterize the global drug universe landscape, and explain existing, yet puzzling, clinical phenotypes. That function alone holds tremendous potential for drug repurposing, identification of novel drug combinations, and side effect predictions.

We show herein the potential of BANDIT in expediting drug development, as it spans the entire space ranging from new target-identification and validation to clinical drug development and beyond, by informing repurposing efforts. We expect that BANDIT will help reduce failure rates in the clinic and shorten the time required for drug approval by identifying the right patient population most likely to benefit from a given therapeutic. By allowing researchers to quickly obtain target predictions it could streamline all subsequent drug development efforts and save both time and resources. Furthermore BANDIT could be used to rapidly screen a large database of compounds and efficiently identify any promising therapeutics that could be further evaluated. Overall our results demonstrate that BANDIT is a novel and effective screening and target-prediction platform for drug development and is poised to positively impact current efforts.

CHAPTER FOUR

GENOMIC AND MACHINE LEARNING APPROACHES TO ADVANCING THE DEVELOPMENT OF IMIPRIDONE FAMILY COMPOUNDS[*]

PREAMBLE

This chapter consists of analyses that are included in a number of different papers that are either published, submitted, or in preparation. I performed all computational and genomic analyses. Selected experiments were designed in partnerships with Drs. Allen and El-Deiry. The El-Deiry lab and Oncoceutcs Inc performed all experimental validation and clinical profiling.

INTRODUCTION

ONC201 is the founding member of the imipridone family of compounds, first identified as an anti-cancer candidate in a screen for p53-independent inducers of TNF-related apoptosis-inducing ligand (TRAIL) gene transcription in (TRAIL-resistant) bax- null HCT116 human colorectal cancer (CRC) cells [144,145]. Based on our finding that DRD2 was the specific binding target of ONC201 (**See Chapter 2**) we investigated how this information could be combined with genomic and other computational analyses to advance development of this unique compound family. Traditional analog selection and lead optimization processes are time–consuming and require precise chemical

[*] Allen JE, Kline CLB, Prabhu VV, et al. Discovery and clinical introduction of first-in-class imipridone ONC201. *Oncotarget*. 2016;7(45):74380-74392. doi:10.18632/oncotarget.11814.

[*] Prabhu, VV, Madhukar NS, Kline LB, et al. Dopamine receptor dysregulation in cancer and its role in tumor response to the anti-cancer DRD2 antagonist, ONC201 (In Preparation)

experiments, large-scale screenings, and often resource-intensive computational simulations [146]. Furthermore, in this age of “precision-medicine” it has become crucial to design clinical trials and approval strategies for the indications and patients most likely to see a large improvement [147,148]. In this chapter I will discuss how, starting from a validated target, we have been able to use BANDIT and pathway-based genomic analyses to accelerate analog optimization, identify responsive cancer types, and select genomic biomarkers predictive of efficacy. Together these results the clinical applicability of imipridone compounds and ultimately improve chances of FDA approval at a later stage.

RESULTS

Computational Selection of Imipridone Analogs Based on DRD2 Selectivity

To rank analogs of ONC201 based on their ability to selectively bind DRD2 vs. other dopamine receptors we calculated the structural similarity between each of the 9 ONC201 analogs and all drugs with known targets in BANDIT’s database. Since the only available data type for all analogs was the chemical structure, we used BANDIT to compute total likelihood value for each analog-known drug pair based on the calculated structural similarity score. For each analog we calculated a BANDIT-DRD2 rank based on whether D2 receptors (DRD2, DRD3) were predicted as targets and not D1 receptors (DRD1, DRD4, DRD5). We then tested each analog against HCT116 cancer cells and observed a significant correlation between each analog’s effect on cell viability and its calculated BANDIT-DRD2 rank (**Figure 4.1**). This result highlights how

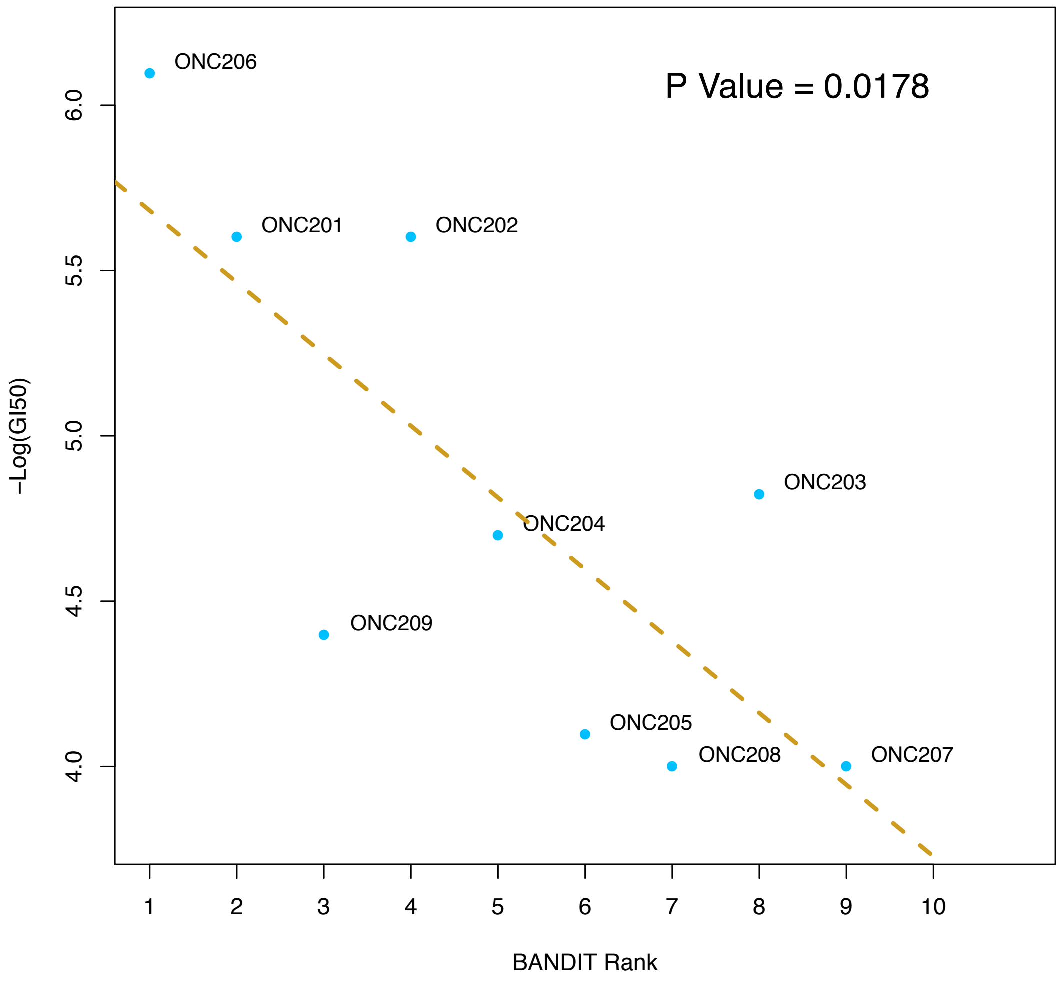

Figure 4.1 – Imipridone analog ranking correlates to measured efficacy. Ranking of each analog based on predicted DRD2-specificity by BANDIT against growth inhibition efficacy measured in HCT116 lines.

BANDIT along with focused target information can be used to expedite analog selection and optimization.

We next focused in on two particular analogs – ONC206 and ONC212 – orphan compounds that had both previously shown efficacies in multiple cancer types. We ran both compounds through BANDIT and predicted ONC206 to be a more potent DRD2 binder than ONC201, whereas ONC212 was not predicted to bind to any previously targeted GPCRs. To further investigate the potential for these analogs to bind to DRD2, we calculated the total likelihood value for each shared target prediction between ONC206/ONC212 and drugs known to target DRD2. Looking at the top scoring pairs as well as all known DRD2 binders, BANDIT consistently predicted ONC206 to be stronger binder to DRD2 than ONC212 (**Figure 4.2A-B**). Testing both against a panel of GPCRs we confirmed this results, revealing that ONC206 was a strong and selective binder to DRD2 (in fact stronger than ONC201) whereas ONC212 actually bound to a GPR132 – an orphan GPCR (**Figure 4.3A-B**). This result further highlighted the potential of BANDIT and other computational approaches to advance analog selection efforts.

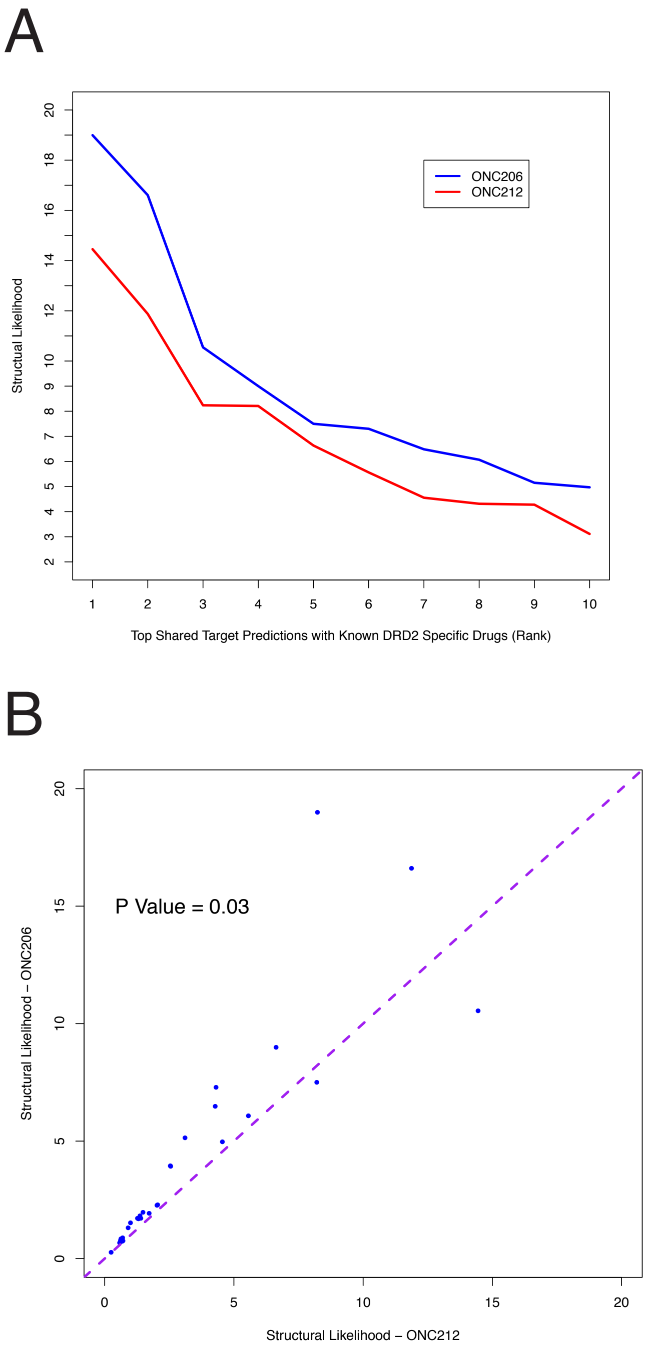

Figure 4.2 – Ranking of ONC206 and ONC212. A) Structural likelihood of each of the top 10 predictions in BANDIT–DRD2 for ONC206 and ONC212. We observe that ONC206 consistently has a higher likelihood value for each prediction. B) Structural likelihood values for ONC206 and ONC212 for shared target predictions with known drugs that only target DRD2 and no other dopamine receptors. P value was calculated using a paired t-test to measure whether ONC206 values were significantly higher than ONC212's.

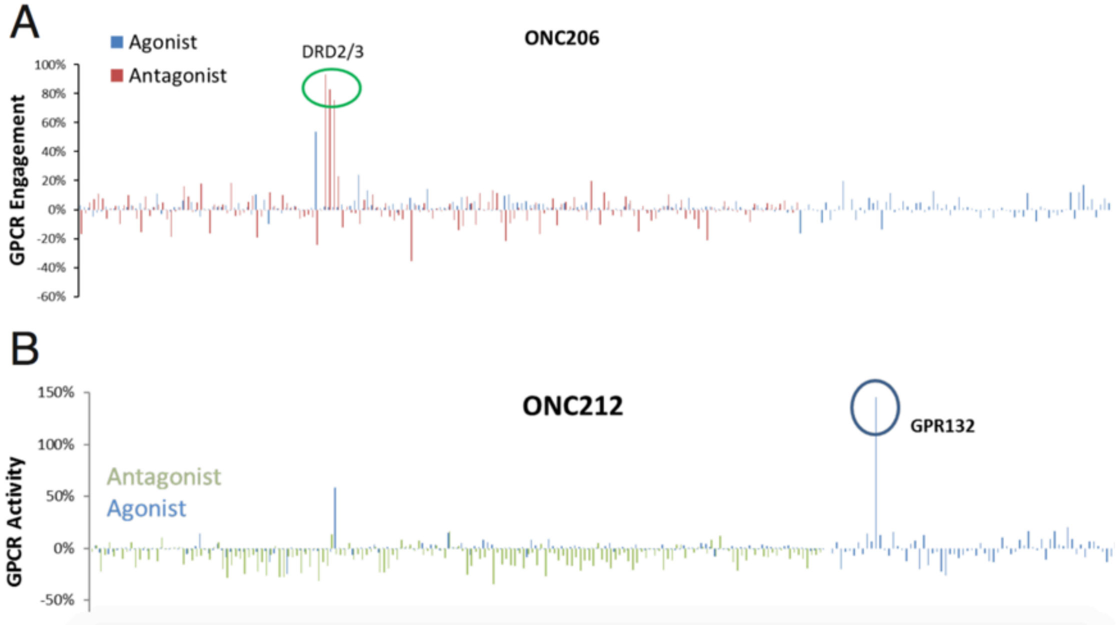

Figure 4.3 – GPCR Profiling of ONC201 Analogs. A) GPCR Engagement Assays for ONC206 and B) ONC212 highlighting binding to DRD2/3 and GPR132 respectively

Target Based Indication Selection for ONC201

Examining pan-cancer expression data in The Cancer Genome Atlas (TCGA) we found that DRD2 is broadly expressed in many different cancer types compared to the respective normal tissues (**Figure 4.4A**). This identified pheochromacytomas/paragangliomas (PCPG) and glioblastomas (GBM) as the cancer types with the highest expression of DRD2 and these results were confirmed on the protein level (**Figure 4.4B**). Based on these findings, we hypothesized that ONC201 would be effective at treating these cancer types and these results were confirmed with in vitro cell line studies (**Figure 4.5A-B**). To further evaluate efficacy in these indications two clinical trials for glioblastoma (NCT02525692) and pheochromacytoma (NCT03034200) have been started. To further evaluate imipridone efficacy we tested ONC206 (a more potent binder of DRD2) on GBM and PCPG cell lines. We observed a better overall efficacy for ONC206 compared to ONC201, providing additional evidence that ONC201/ONC206's anti-cancer efficacy is due to DRD2 antagonism (**Figure 4.5C-D**).

DRD5 as a Potential Biomarker of ONC201 Response and Resistance

To gain more insight into the mechanism of ONC201 susceptibility, we generated RKO cells with acquired and stable resistance to ONC201 from the ONC201-sensitive parental cells. Analysis of exome sequencing of resistant and parental cells revealed a heterozygous Q366R mutation in the DRD5 gene only in both resistant clones. DRD5 is a member of the D1 class of dopamine receptors whose activation opposes the activity of D2 class of dopamine receptors (which includes DRD2) and is known to dimerize with DRD2 via electrostatic interactions between intracellular residues [149,150].

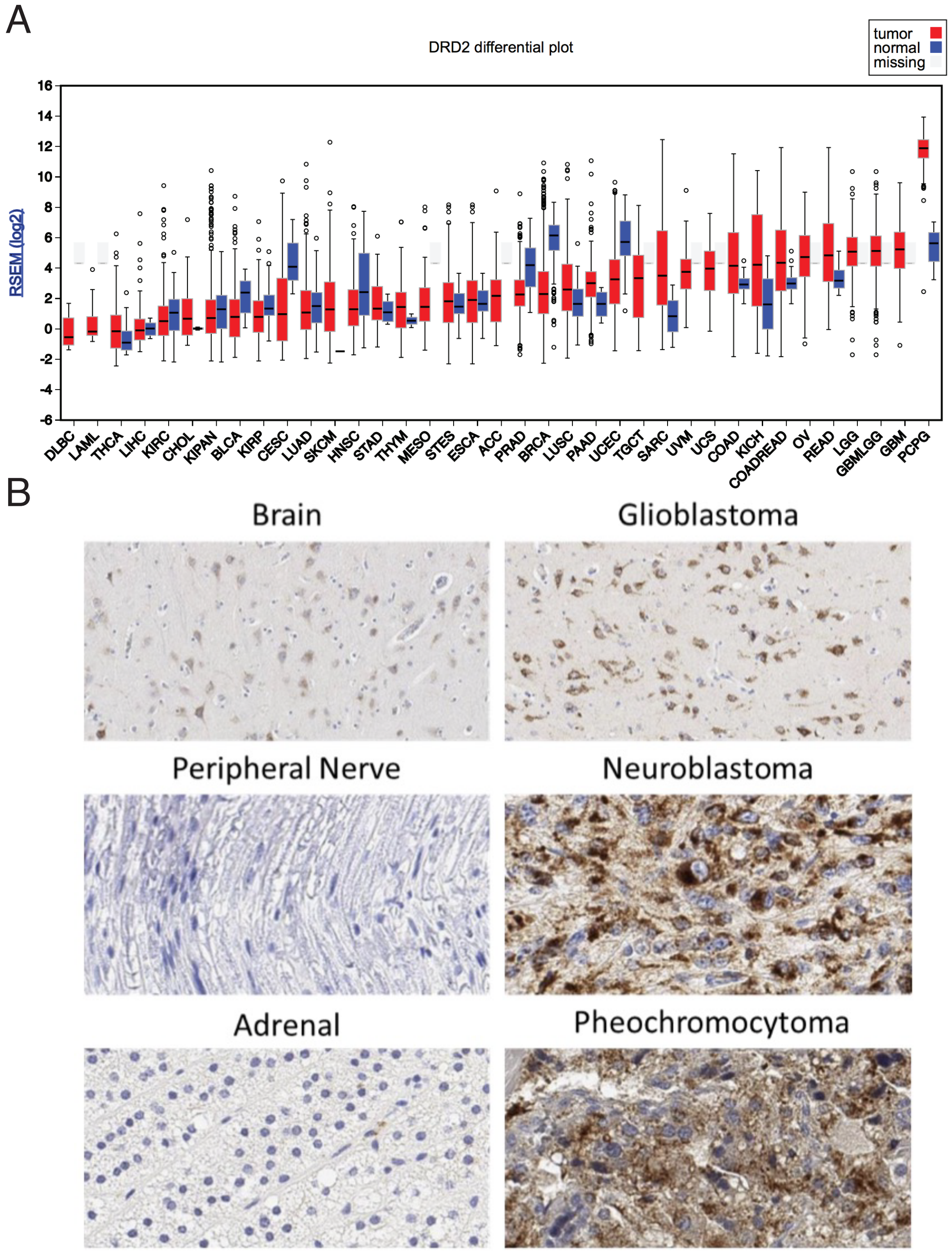

Figure 4.4 – Overexpression of DRD2 in select cancers. A) Comparison of tumor and normal tissue expression across various cancers according to TCGA data. B) Immunohistochemical staining for DRD2 in patient-derived tissue microarrays

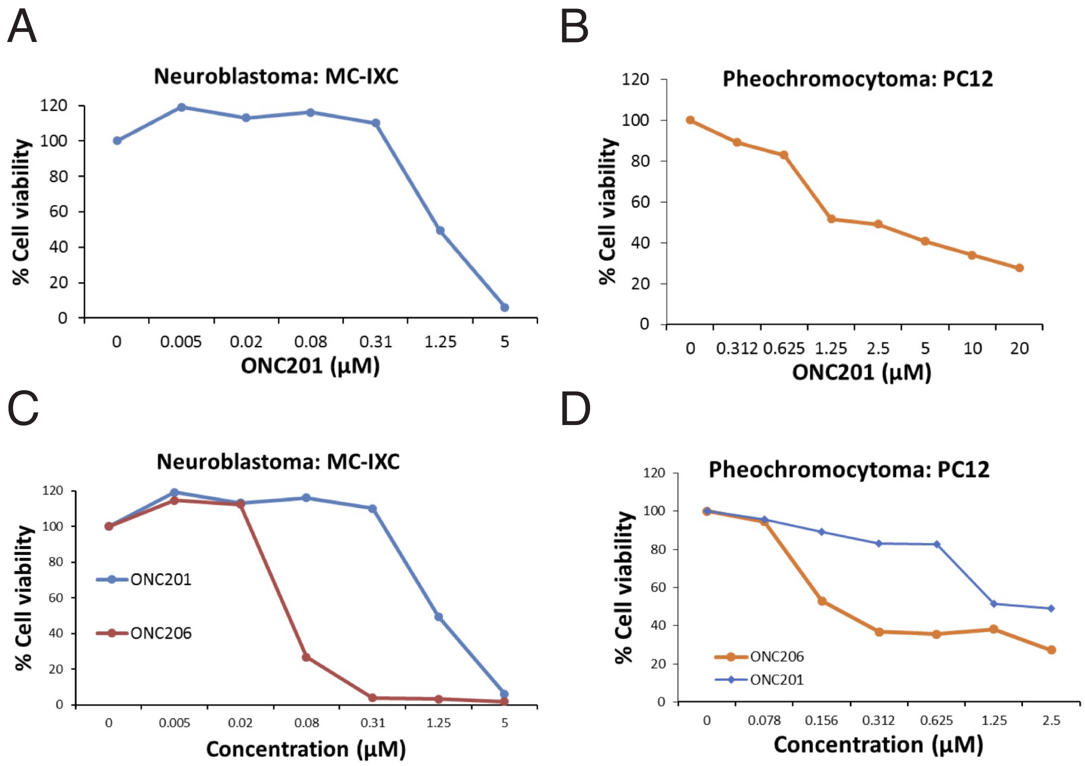

Figure 4.5 – ONC201 and ONC206 efficacy in select cancers. A) Efficacy testing for ONC201 in MC-IXC neuroblastoma and B) PC12 pheochromacytoma cell lines. C) Efficacy testing for ONC201 and ONC206 in MC-IXC neuroblastoma and D) PC12 pheochromacytoma cell lines.

Interestingly the Q366R mutation imparts an electrostatic change from a net neutral to a positive charge at an intracellular amino acid; thus we hypothesized that the Q366R mutation could enhance dimerization and antagonize downstream DRD2 signaling. Further supporting that the Q366R mutation could confer resistance to ONC201, we found that overexpression of the Q366R DRD5, but not the wild type, induced tumor cell death and could reconstitute resistance in parental ONC201-sensitive cells.

Based on these findings we investigated the role of DRD5 expression as a predictive biomarker for ONC201. Previous studies had reported on the efficacy of ONC201 on a set of well-characterized cancer cell lines [151,152]. We found that across all cell lines, high expression of DRD5 correlated with lower overall ONC201 efficacy (**Figure 4.6A**). This result was also confirmed with clinical results as all 3 GBM patients treated with ONC201 who had progression free survival scores of greater than 5 months did not have detectable levels of DRD5 expression (**Figure 4.6B**). Further examining how low levels of DRD5 could be used as a predictive biomarker, we overlaid expression of DRD2 and observed that ONC201 had the best overall efficacy when both DRD5 was lowly expressed and DRD2 was highly expressed (**Figure 4.6C**). This finding opens the door for using DRD2+/DRD5- as a biomarker for patient and indication selection.

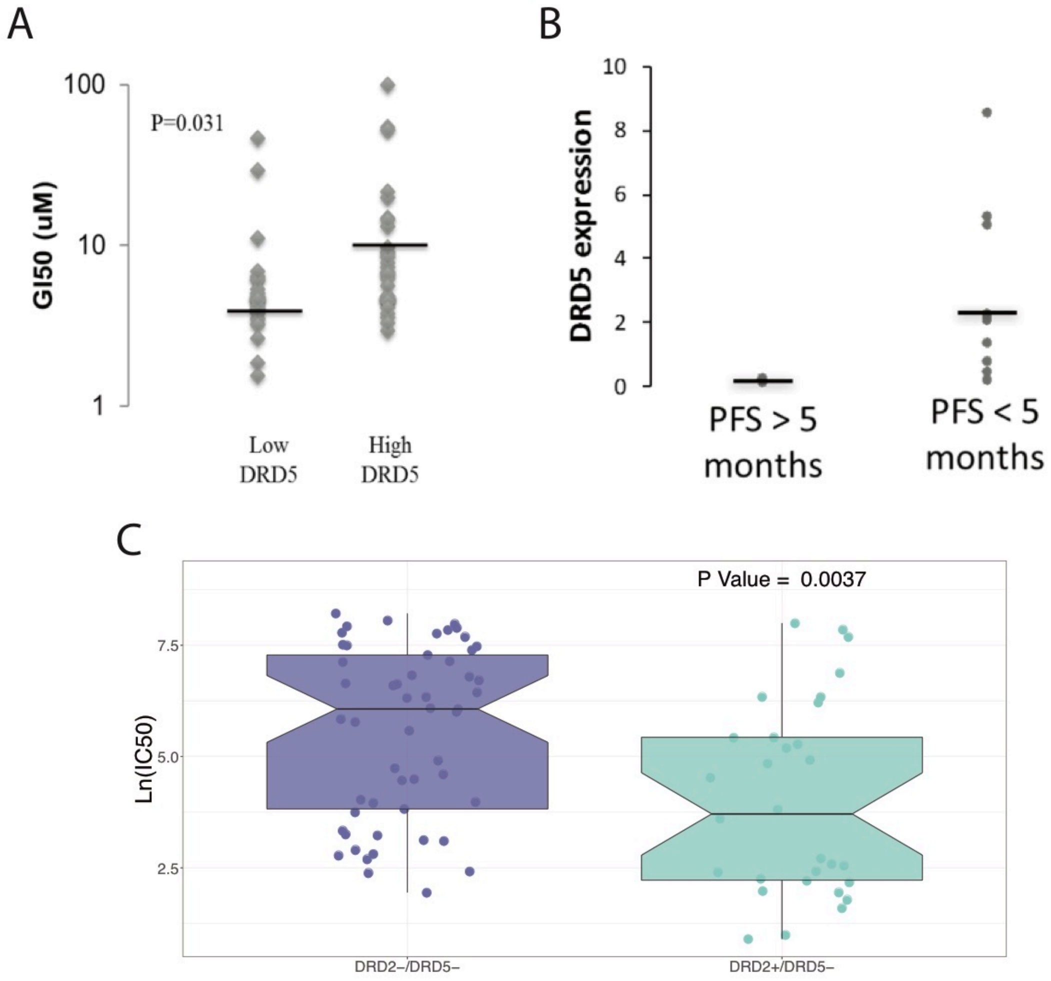

Figure 4.6 – DRD5 as a predictive biomarker. A) ONC201 GI50 of NCI60 cells categorized by DRD2 mRNA expression using z-score. P value was calculated using a KS test B) DRD5 expression in archival tumor samples categorized by PFS>5 months (n=3) or PFS<5 months (n=12). C) IC50 of ONC201 in GDSC cells based on expression of DRD2 and DRD5. P value calculated using a KS test.

CHAPTER FIVE

CANCER STEM CELL-RELATED GENE EXPRESSION AS A POTENTIAL BIOMARKER OF RESPONSE FOR FIRST-IN-CLASS IMIPRIDONE ONC201 IN SOLID TUMORS[*]

PREAMBLE

This chapter is derived from a paper that has been submitted and is currently under review at *PLoS One* (as of July 2017). Experiments were designed in partnership with VVP, ARL, MDB, JA, and WSED. I performed all computational and genomic analyses (determining expression changes and identifying potential biomarkers). In vitro validation experiments were done by the El-Deiry Lab (ARL, MDB, and WSED). Clinical testing and prior screening results were provided by Oncoceutics Inc (VVP and JA).

INTRODUCTION

Several clinical studies have demonstrated the relevance of cancer stem cells (CSCs) that clearly correlate with recurrence, metastasis and poor survival in solid tumors [153-155]. Recent objective responses observed in Phase I/II clinical trials of various CSC-targeted agents in a number of advanced refractory solid tumors have further established the importance of CSCs as a therapeutic target [156-158].

[*] Prabhu VV[a], Lulla AR[a], Madhukar NS[a], Baumeister MD[a], Zhao D, et al. “Cancer stem cell-related gene expression as a potential biomarker of response for first-in-class imipridone ONC201 in solid tumors.” 2017. (Submitted) ([a] = co first authors)

The first-in-class small molecule imipridone ONC201 is currently in Phase I/II clinical trials for advanced cancer [159]. The first-in-human Phase I study in advanced solid tumors demonstrated ONC201 to be safe, and exhibit predicted pharmacokinetics, sustained pharmacodynamics and tumor shrinkage [160]. The anti-CSC efficacy of ONC201 has been previously demonstrated *in vitro* and *in vivo* in colorectal cancer and acute myeloid leukemia (AML) [161,162]. ONC201-mediated depletion of chemotherapy-resistant colorectal CSCs involves dual inactivation of Akt and ERK signaling that results in transcription factor Foxo3 activation that leads to DR5/TRAIL-dependent inhibition of self-renewal [144,161]. In the current study, we evaluated whether the anti-CSC effects of ONC201 involve early changes in stem-cell related gene expression prior to tumor cell death. We examined if ONC201-mediated inhibition of CSCs extends to other solid tumors. Additionally, we tested whether CSC expression can serve as a potential biomarker of ONC201 response.

RESULTS

ONC201 modulates stem cell-related gene expression

A targeted network analysis of gene expression profiles of HCT116 p53-null human colon cancer cells treated with ONC201 (18 h and 48 h) revealed that several stem cell-related genes, transcription factors and signaling pathways are significantly modulated by the compound (**Figure 5.1A**). Specifically, mRNA levels of *ID1* (colon/glioblastoma CSC-regulation [163], 2.5-fold), *ID2* (glioma stem cell regulation [163], 3.2-fold), *ID3* (colon/glioma CSC-regulation [163], 2.9-fold), *ALDH7A1* (prostate CSC marker/metastasis [164], 2-fold) were significantly downregulated and *KLF9* (glioblastoma stem cell inhibitor

[165], 1.5-fold) was significantly upregulated in HCT116 p53-null cells upon 48 hour ONC201 treatment (**Table 5.1**), indicative of potential anti-CSC effects in these solid tumors. Also, mRNA levels of Wnt pathway-related genes such as ligand *WNT16* (hematopoietic stem cell [166]/prostate cancer resistance-related [167], 13.5-fold), receptors *FZD2* (regulator of epithelial-mesenchymal transition (EMT)/colon cancer metastasis [168], 2.98-fold), *FZD4*

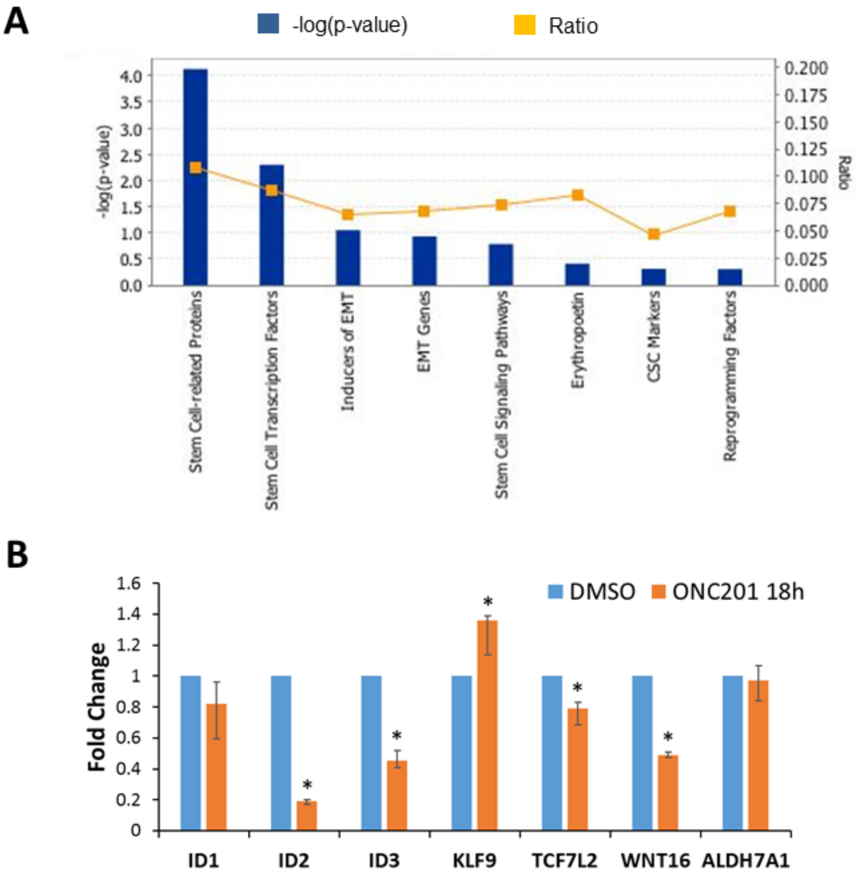

Figure 5.1 – ONC201 modulates stem cell-related gene expression. (A) Summary of targeted network analysis of stem cell-related changes in ONC201-treated (10 μM) HCT116 p53-null cells by Ingenuity Pathway Analysis. The –log(p-value) is indicated for each group of genes. Ratio indicates the relative number of genes that were significantly changed upon ONC201 treatment compared to total number of genes in the group. (B) qRT-PCR for indicated stem cell-related genes in DMSO/ONC201-treated (5 μM, 18 h/48 h, n = 3) HCT116 p53-null cells. * indicates $p < 0.02$ relative to DMSO.

Table 5.1 – ONC201-mediated CSC- and Wnt-pathway-related changes in gene expression. CSC- and Wnt pathway-related drug-induced changes identified with Ingenuity Pathway Analysis for gene expression profiles of HCT116 p53-null cells treated with ONC201 (10 μM) for 48 h. Fold change relative to DMSO treated cells.

Gene	**Fold change**	**mRNA Level**	**P value**	**CSC function**
ALDH7A1	2.003	down	3.16E-03	Prostate CSC marker
ID1	2.519	down	5.85E-04	Colorectal/glioblastoma CSC-related protein
ID2	3.236	down	8.49E-05	Glioma stem cell-related protein
ID3	2.884	down	1.05E-03	Colorectal/glioma CSC-related protein
KLF9	1.542	up	4.88E-03	Glioblastoma stem cell-related protein
WNT16	13.496	down	1.98E-03	Prostate Cancer Resistance, HSC regulation
Gene	**Fold change**	**mRNA Level**	**P value**	**Wnt pathway function**
WNT16	13.496	down	1.98E-03	Ligand
FZD2	2.990	down	8.61E-04	Receptor
FZD4	3.932	down	1.38E-03	Receptor
TCF7L2	3.550	down	5.11E-03	Transcription factor

(glioma stemness [169], 3.9-fold) and transcription factor *TCF7L2* (stem cell differentiation [170], 3.55-fold) were significantly downregulated (**Table 5.1**). Genes involved in Wnt signaling, Hedgehog signaling and stem cell pluripotency were significantly modulated as early as 18 h upon ONC201 treatment. Modulation of stem cell-related transcription was further confirmed in RKO colorectal cancer cells upon ONC201-treament (48 h). Validation with qRT-PCR indicated that *ID2*, *ID3*, *TCF7L2*, *WNT16* mRNA levels were significantly downregulated while *KLF9* mRNA was significantly upregulated in response to ONC201 treatment (18 h) in HCT116 p53-null cells (**Figure 5.1B**). Clearly, ONC201 specifically impacts stem cell-related transcription at time points (18 and 48 h) that precede cell death, which occur beyond 48 h in solid tumor cells [144]. These early effects on stem-cell related transcription are followed by inhibition of CSC markers and self-renewal by ONC201 at 48-72 h [161].

ONC201 targets cancer stem cells in prostate and glioblastoma tumors

Based on the relevance of the CSC-related genes modulated by ONC201 in prostate cancer and glioblastoma, we tested the effects of ONC201 on CSC-related gene expression and self-renewal in these tumor types. ONC201 was tested in CSC-enriched 3-dimensional neurosphere culture models of primary glioblastoma samples, including newly diagnosed (GBM8, GBM18) and recurrent (GBM67R and GBM152) samples. ONC201 potently inhibited *in vitro* cell proliferation of all 4 lines, with IC50 values of 433 nM (GBM18), 1.09 *μ*M (GBM8), 3.97 *μ*M (GBM67R) and 688 nM (GBM152) (**Figure 5.2A**). We have previously demonstrated that ONC201 downregulated CSC markers *CD133*,

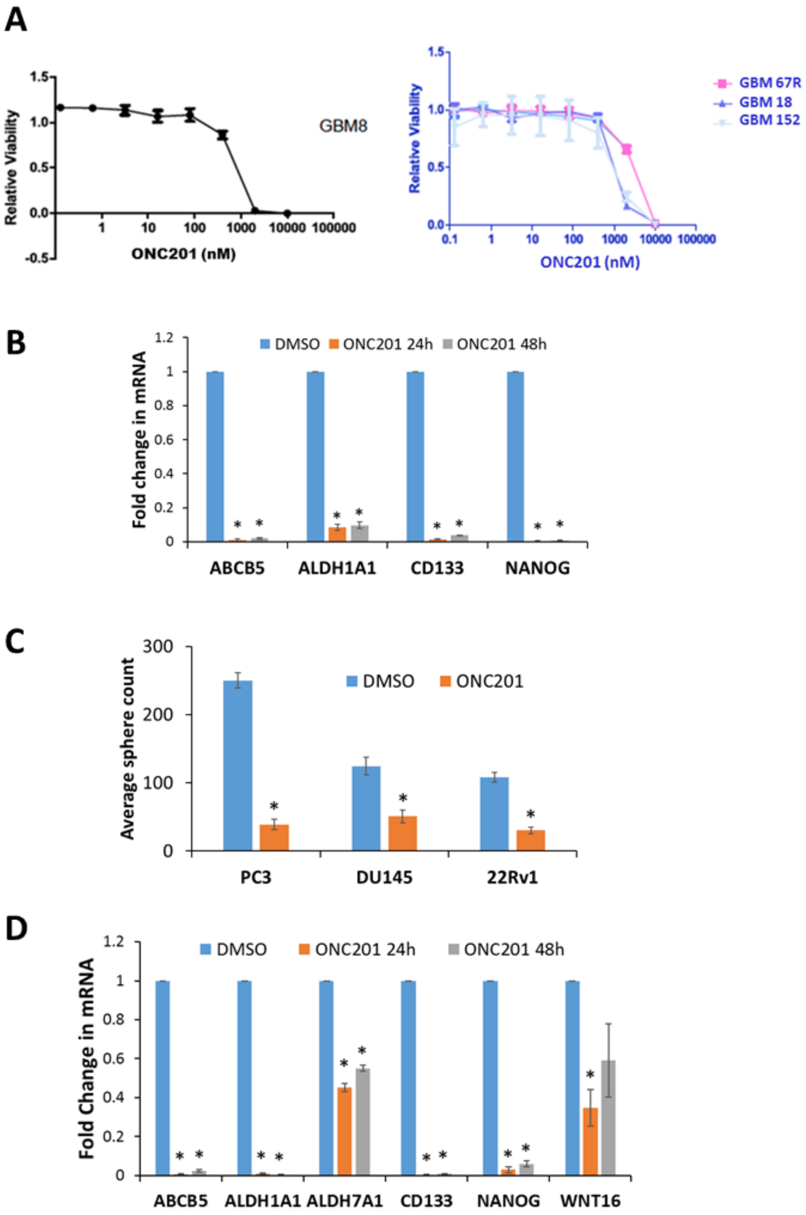

Figure 5.2 – ONC201 targets cancer stem cells in prostate and glioblastoma tumors. (A) Effect of indicated concentrations of ONC201 (72 h) on viability of newly diagnosed (GBM8, GBM18) and recurrent (GBM67R, GBM 152) glioblastoma cells in 3D neurosphere culture. (B) qRT-PCR for indicated stem cell-related genes in DMSO/ONC201-treated (5 μM, 24 h/48 h, n = 3) SNB19 cells. * indicates $p < 0.0002$ relative to DMSO. (C) Effect of DMSO/ONC201 (5 μM, 72 h, n = 3) on tumor sphere formation of indicated prostate cancer cell lines. * indicates $p < 0.025$ relative to DMSO. (D) qRT-PCR for indicated stem cell-related genes in DMSO/ONC201-treated (5 μM, 24 h/48 h, n = 3) DU145 cells. * indicates $p < 0.04$ relative to DMSO.

ALDH1A1 and *CD44* in colorectal cancer cells *in vitro* and *in vivo* [161]. Consistent with these findings, ONC201 significantly downregulated CSC-related genes *ABCB5*, ALDH1A1, *CD133* and *NANOG* in SNB19, T98G and U251 glioblastoma cells (**Figure 5.2B, 5.3A and 5.3B**). Western blotting showed that *CD133, ALDH1, NANOG, ID1* and *ID3* were downregulated in U251 and T98G glioblastoma cells upon ONC201 treatment at 72 h (**Figure 5.3C-D**). *ID1* protein is upregulated at 24 h, however, mRNA levels decrease at 48 h (**Table 5.1**) and protein levels decrease by 72 h post ONC201 treatment (S1C and S1D Figs). ONC201 significantly reduced tumorsphere formation of 22Rv1, DU145 and PC3 human prostate cancer cells (**Figure 5.2C**). ONC201 significantly downregulated CSC-related genes *ABCB5, ALDH1A1, ALDH7A1, WNT16, CD133* and *NANOG* in DU145 prostate cancer cells (**Figure 5.2D**). Western blotting revealed that *WNT16* was downregulated in LNCaP and 22Rv1 while CSC marker *CD44* was downregulated in 22Rv1 cells upon ONC201 treatment at 72 h (**Figure 5.3E-F**). Thus, changes in stem cell-related transcription and anti-CSC effects of ONC201 observed in colorectal cancer extend to prostate cancer and glioblastoma.

Inhibition of cancer stem cells does not occur in tumor cells with acquired resistance to ONC201

We explored the correlation of ONC201-mediated changes in stem cell-related gene transcription with anti-tumor efficacy. ONC201 inhibited sphere formation of parental RKO wild-type (wt) cells but not RKO cells with acquired resistance to ONC201 (**Figure 5.4A-B**). Accordingly, ONC201 significantly downregulated mRNA levels of the stem cell-related genes *ID1* (2.1-fold), *FZD4* (1.6-fold), *HES7* (2.5-fold), *CCNB1* (3.7-fold) and *TCF3* (1.8-fold) in

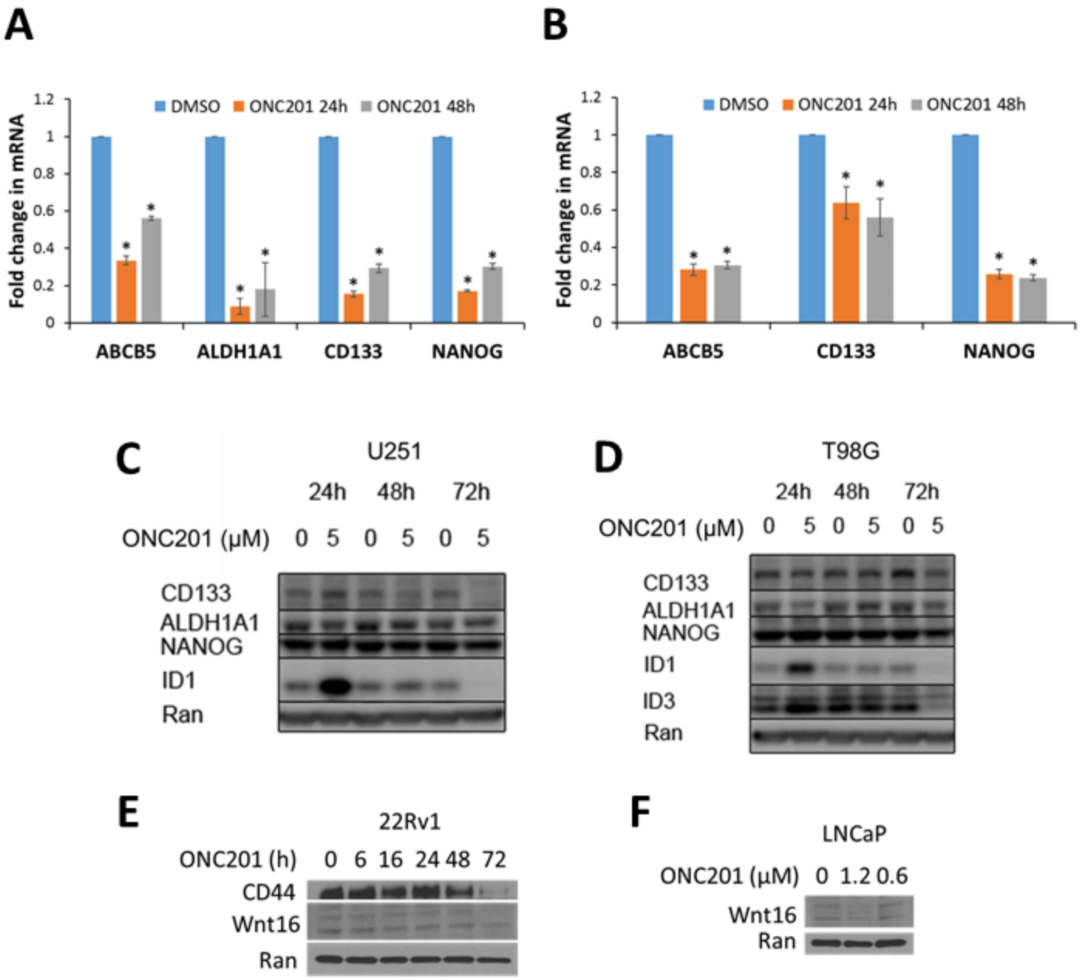

Figure 5.3 – ONC201 targets cancer stem cells in prostate and glioblastoma tumors. qRT-PCR for indicated stem cell-related genes in DMSO/ONC201-treated (5 μM, 24h/48h, n = 3) (A) T98G and (B) U251 cells. * indicates $p < 0.02$ relative to DMSO. (C) and (D) Western blot for indicated stem cell-related proteins in glioblastoma cells treated with indicated doses of DMSO/ONC201 for indicated time. (E) Western blot for indicated proteins in DMSO/5 μM ONC201-treated 22Rv1 cells for indicated time. (F) Western blot for indicated proteins in DMSO/ONC201-treated LNCaP cells for 72 h.

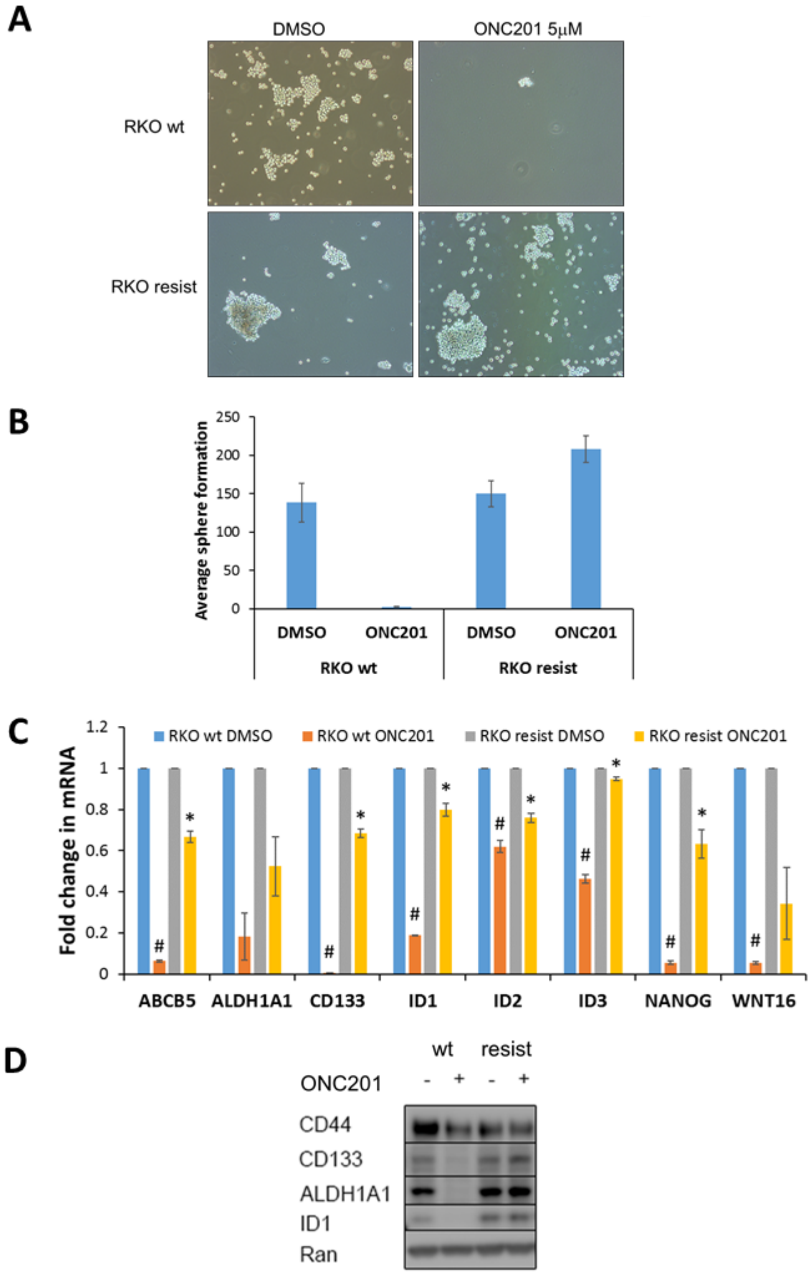

Figure 5.4. Inhibition of cancer stem cells does not occur in tumor cells with acquired resistance to ONC201. (A) Effect of DMSO/ONC201 (5 μM, 72 h, n = 3) on tumor sphere formation of RKO wild-type (wt) and ONC201-resistant (resist) cells. Representative image (10X magnification) of spheres (> 60 μm) (B) Quantification of spheres in (A). (C) qRT-PCR for indicated stem cell-related genes in DMSO/ONC201-treated (5 μM, 48 h, n = 3) RKO wild-type (wt) and ONC201-resistant (resist) cells. # indicates $p < 0.003$ relative to wt DMSO. * indicates $p < 0.05$ relative to wt ONC201. (D) Western blot for indicated stem cell-related proteins in DMSO/ONC201-treated (5 μM, 72 h) RKO wild-type (wt) and ONC201-resistant (resist) cells.

RKO wt cells but not in ONC201-resistant RKO cells, indicating that CSC-inhibition could serve as a biomarker of ONC201 response. Validation with qRT-PCR indicated that ONC201-mediated inhibition of CSC-related genes *ABCB5, CD133, ID1, ID2, ID3* and *NANOG* in RKO wt cells was significantly reduced in ONC201-resistant RKO cells (**Figure 5.4C**). Western blot confirmed that ONC201-mediated downregulation of *CD44, CD133, ALDH1* and *ID1* occurred in RKO wt cells, but not in ONC201-resistant RKO cells (**Figure 5.4D**). Thus, CSC depletion is a critical component of ONC201's anti-cancer efficacy and can serve as a potential pharmacodynamic biomarker of ONC201 response.

Cancer stem cell expression in solid tumors as a potential biomarker of response for ONC201

Finally, we used the GDSC panel of approximately 1,000 unique cancer cell lines [159] to determine whether ONC201 *in vitro* efficacy correlates with the expression of CSC-related genes in the treatment-naïve setting. All genes identified in the earlier studies were tested and a significant correlation with *ID1* (D stat = 0.18), *CD44* (D stat = 0.173), *TCF3* (D stat = 0.253) and *HES7* (D stat = 0.254) expression was observed. Interestingly we found that high expression of *TCF3* and *HES7* significantly predicted sensitivity to ONC201 (**Figure 5.5A-B**), suggesting that ONC201 may be efficacious in tumors with high basal Wnt signaling. Also, low expression of *ID1* and *CD44* significantly predicted sensitivity to ONC201. These data are consistent with the heterogeneity observed within CSC populations with various combinations of markers representing different cell populations [171]. Furthermore, when we tested ONC201 efficacy in cell lines that fulfilled at least two of the expression

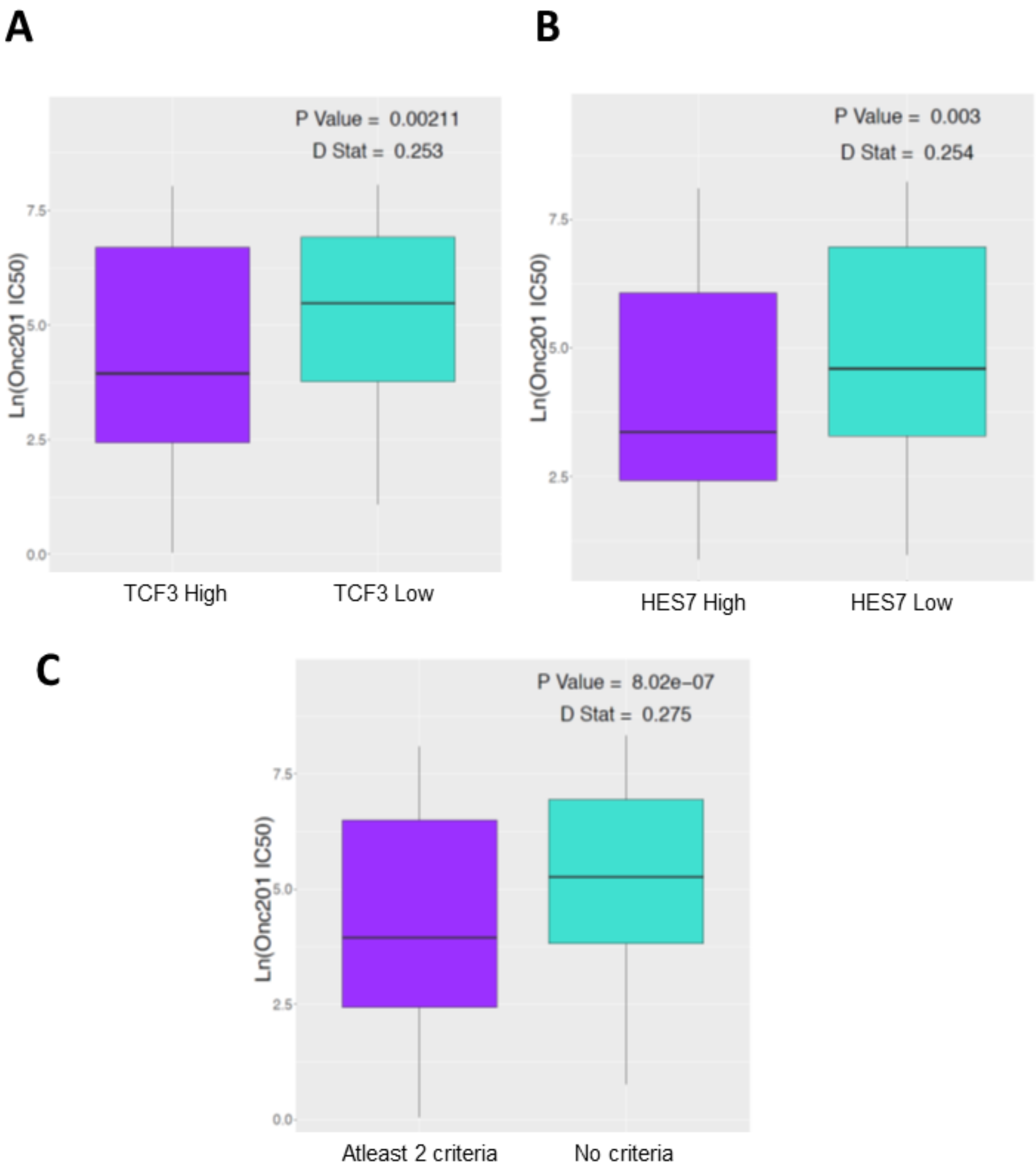

Figure 5.5 –Cancer stem cell expression in solid tumors as a potential biomarker of response for ONC201. Distribution of ONC201 efficacy (IC50) in >1000 GDSC cell lines based on basal RNA expression of (A) *TCF3*, (B) *HES7*. (C) Distribution of ONC201 efficacy (IC50) in >1000 GDSC cell lines based on fulfillment of at least two expression based criteria (low expression of *ID1/CD44* and high expression of *TCF3/HES7*) against cell lines that fulfilled none. P value and D statistic are indicated.

based criteria (low expression of *ID1/CD44* and high expression of *TCF3/HES7*) against cell lines that fulfilled none, there was a greater degree of separation (D stat = 0.2749, P-value = 8.02e-07) (**Figure 5.5C**). These results indicate that pre-treatment expression of certain CSC genes can serve as predictive biomarkers for ONC201 response and that combining the expression of multiple CSC genes results in a stronger overall prediction.

Discussion

We have previously demonstrated the anti-CSC efficacy of ONC201 *in vitro* using established CSC markers, sphere cultures and *in vivo* using limiting dilution studies in colorectal cancer [161]. Additionally, ONC201-mediated inhibition of leukemic stem cells has been confirmed *in vivo* [162]. Depletion of chemotherapy-resistant colorectal CSCs by ONC201 involves an Akt-ERK-Foxo3-DR5-TRAIL-dependent mechanism of inhibition of self-renewal and cell death induction [144,161]. However, it was unclear whether ONC201 depletion of CSCs is a consequence of cell death or involves specific effects on stem-cell related genes that precede inhibition of self-renewal and cell death. In this study, we show that ONC201 specifically impacts stem cell-related transcription at time points (18 and 48 h) that precede cell death which occurs 60-72h post treatment in solid tumor cells (11). These early effects on stem-cell related transcription are followed by inhibition of CSC markers and self-renewal by ONC201 at 60-72 h (9).

ONC201 attenuates diverse CSC markers such as *CD44, CD133, ABCB5, ALDH1A1, ALDH7A1, NANOG, ID1, ID2, ID3* [155,154] and self-renewal signaling pathways such as Wnt, Notch and Hedgehog [166,169,165] that drive tumor-initiation [163,171], therapy resistance [167] and metastasis

[168,164] across various tumor types providing an opportunity for broad-spectrum anti-CSC and anti-cancer effects. Gene expression profiles in colorectal cancer cells revealed ONC201 targets CSC genes involved in prostate cancer and glioblastoma. Accordingly, ONC201 mediated inhibition of self-renewal in solid tumors was confirmed in prostate cancer cell lines and glioblastoma patient derived cells. This study provides further evidence of the broad spectrum anti-cancer efficacy of ONC201 and serves as a rationale for the ongoing single agent Phase I/II trials of ONC201 in advanced refractory solid tumors including prostate cancer and glioblastoma [159]. Drugs targeting differentiated bulk tumor cells alone are typically associated with early clinical responses that may or may not be durable. In contrast, CSC-targeting agents are likely to achieve delayed but durable responses [172]. ONC201's ability to target CSCs provides an opportunity to potentially achieve durable responses in patients with advanced therapy resistant disease, especially in high unmet need indications such as recurrent glioblastoma. Additionally, approved chemotherapies or targeted agents with anti-proliferative effects that do not target CSCs could be combined with ONC201 to provide rapid de-bulking and durable clinical benefit. These results also indicate that ONC201 could be used in the adjuvant/preventative setting for cancer recurrence and metastasis prevention.

Our results also demonstrate that CSC-related gene expression can serve as a potential predictive and pharmacodynamic biomarker of ONC201 response. ONC201 mediated CSC inhibition occurs in sensitive but not in resistant cancer cells confirmed by sphere formation, gene expression and protein levels of established CSC markers. Interestingly, baseline expression of CSC-related genes predicted ONC201 anti-cancer efficacy in >1000 cancer cell lines. Thus, correlative studies testing CSC expression at the RNA and protein level using circulating tumor cells and biopsies from ongoing ONC201 clinical studies are warranted.

CHAPTER SIX

A MACHINE LEARNING APPROACH TO PREDICTING TISSUE SPECIFIC ADVERSE EVENTS[*]

PREAMBLE

This chapter consists of a paper that will soon be submitted for publication (as of July 2017). The method (MAESTER) was conceived in partnership with Drs. Kaitlyn Gayvert and Olivier Elemento. Method development and subsequent analyses were done together with KG. CG assisted with generation of some model features. I primarily wrote the manuscript with input from KG and OE.

INTRODUCTION

Adverse events are currently one of the main causes of failure in drug development and are one of the top 10 causes of death in the developed world[173,174]. Toxicity issues remain a leading cause for the rising clinical trial attrition rates[12,15]. Even after a drug has been approved, adverse drug reactions remain a large burden on the medical system with the costs amounting to as much as $30 billion dollars annually in the USA[175]. Furthermore the identification of the serious adverse events associated with drugs frequently does not occur until after FDA approval, with as many as 50% of adverse events going undetected during human trials[176] . Due to the prevalence and impact of this problem, the U.S. Food and Drug

[*] Madhukar NS[a], Gayvert K[a], Gilvary,C, and Elemento O. “A Machine Learning Approach to Predicting Tissue Specific Adverse Events.” 2017 (In preparation) ([a] = co first authors)

Administration (FDA) has established the US FDA Adverse Event Reporting System (FAERS).

Most adverse event detection experiments are carried out in pre-clinical phases based on animal results or during early clinical trials. However not all adverse events are detected, due to several factors including limited relevance of animal models to human physiology, limited sample sizes during trials, and patient populations that may not be representative of the overall population[175]. Further complications may include the low frequency or late onset of some adverse events[175]. As a result, retrospective studies are currently an important method for further characterization of the side effects associated with drugs. However this requires a large number of patients to be treated first and is dependent on voluntary reporting, which is especially problematic as only 10% of all adverse events are reported post-approval[177].

Ideally possible adverse events would be detected during the pre-clinical phases of drug development, even before animal studies. Cell lines and reporter assays may help detect unwanted side effects early. Computational screening methods are also critical components of current drug development pipelines for evaluating pre-clinical toxicity. In particular, drug-likeness measures, which use molecular features to estimate oral bioavailability as a proxy for drug toxicity, have been widely adopted. Examples of drug-likeness methods include Lipinski's Rule of Five[178] and the Quantitative Estimate for Drug Likeness[179]. More recently machine learning based methods have been proposed for predicting drug toxicity, including previous work from our group (PrOCTOR) which integrates established molecular properties with target-based features to directly predict broad clinical trial toxicity[17]. However these types of approaches have not be systematically applied to predicting specific adverse events, such as liver or heart toxicity. Better methods for predicting such toxicities could improve fast-fail procedures and facilitate better trial design. To address this problem, we introduce MAESTER, a new machine-learning platform for the prediction of tissue-specific adverse events. We show that for a set of 6 serious adverse events MAESTER achieves unprecedented accuracy while maintaining high specificity and sensitivity. Additionally we demonstrate how MAESTER could have identified drug adverse events that were missed by traditional screening methodologies.

RESULTS

Target based features connects drugs to specific adverse events

Previous work by our group (PrOCTOR) has demonstrated the importance of considering drug targets and the genomic heterogeneity of different tissue systems when predicting general toxicity. Given this association, we hypothesized that we could better predict specific adverse events (AEs) if we included information on drug targets in the most relevant tissue to the given AE. We first focused on a set of six tissues whose corresponding AEs are correlated with clinical trial failures: liver, kidney, blood, heart, lung, and pancreas. We used the SIDER database of drug side effects to identify subsets of drugs (~150 drugs) that are associated with these tissue-specific adverse events (TSAEs) (**Table 6.1**)[111]. For each tissue, we also established a “safe” set of drugs for comparisons by filtering out any drugs correlated with those TSAEs or other AEs highly correlated with fatalities in

Table 6.1 – Table of the 6 major adverse event categories. In addition to the given adverse event, certain synonymous adverse events were also included and any drugs with containing an adverse event in the “other removed terms” category were removed excluded from the safe set.

Adverse Event	Synonyms	Tissue	Other Removed Terms
DILI	Liver Disease, Liver Injury, Liver Damage	Liver	"Nephro"
Heart Attack	Myocardial Infarction	Heart	"Immun"
Renal Failure	Kidney Failure	Kidney	
Neutropenia	-	Blood	
Pleural Effusion	-	Lung	-
Pancreatitis	-	Pancreas	-

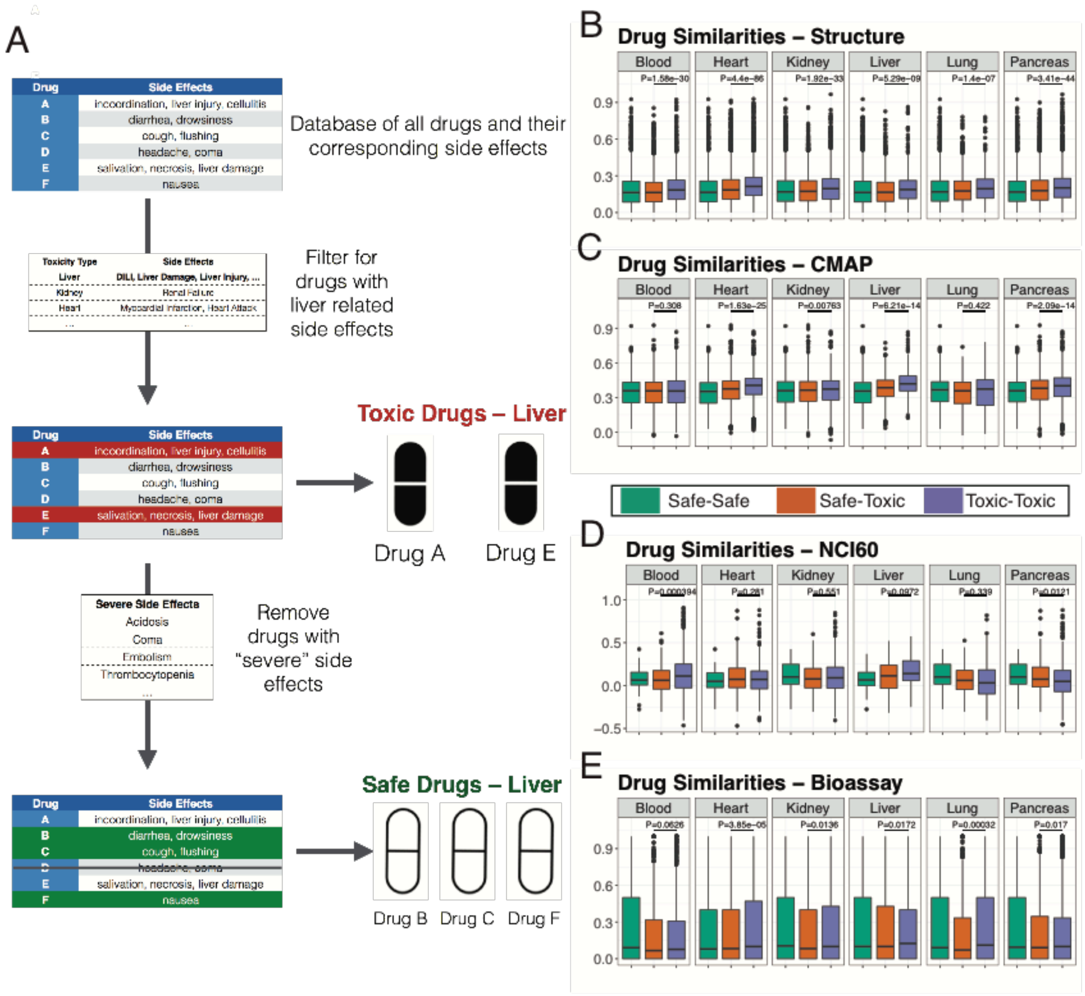

Figure 6.1 – A) Schematic describing the process by which we selected our toxic and safe drugs for each specific tissue. B) Similarities of across all toxic drugs pairs, safe drug pairs, and all combinations of toxic and safe drugs for drug structures, C) gene expression changes, D) growth efficacies, and E) bioassays. P values were calculated using a Wilcoxon Rank Sum test.

openFDA (https://open.fda.gov/) (**Figure 6.1A**). For each drug, we compiled structural representations in the format of SMILES from DrugBank, differential gene expression profiles from the Broad Institute's Connectivity Map (CMAP)[98], growth inhibition patterns across the NCI60 cell lines (NCI60) from the NCI's Developmental Therapeutics Program[108], and bioassay data from PubChem[110].

For each tissue we then investigated how these safe and toxic drugs compare to each other. For each pair of drugs, we calculated a similarity score for each of the considered data types. We found that in all tissues, tissue-specific toxic drugs were most structurally similar to each other (**Figure 6.1B**). Additionally, toxic drugs tended to also be most similar to other toxic drugs in terms of differential gene expression profiles (**Figure 6.1C**), growth inhibition screens (**Figure 6.1D**) and bioassays (**Figure 6.1E**). Interestingly we found distinct patterns across the different tissue types – for instance, growth inhibition was best able to separate out drugs with heart specific adverse events, whereas gene expression changes had the greatest utility in the liver. These patterns could be incredibly valuable for adverse event prediction as they highlight how we can model the diversity across drugs with a given side effect.

We next examined how expression of a drug's targets could be used to predict TSAEs. To test this we integrated tissue-specific expression data measured by the GTEX database. For each of the six tissues, we determined all related adverse events and identified sets of safe and toxic drugs for each tissue following the same procedure outlined in **Figure 6.1A**. For each toxic or safe drug in a given tissue set, we measured the expression of all of that drug's targets in the specific tissue (**Figure 6.2A-E**). Overall drugs with TSAEs tended to have higher target expression in the tissue corresponding to their toxicity than their safe drug counterparts. This information helps illustrate how its important to consider target based features and tissue-specific expression when predicting adverse events.

Distinct Patterns of Tissue-Specific Toxic and Safe Target Sets

Due to the significant relationship between drug target expression and related tissue adverse events, we next sought to define a set of tissue-specific "toxic targets"– proteins that are only targeted by drugs with known toxicity in that tissue – and "safe targets" – proteins only targeted by drugs with no related tissue toxicities. To do this, we begin by taking the safe and toxic drug sets described in **Figure 6.1A** and identifying any targets exclusive to each drug subset (**Figure 6.2F**). Interestingly we found that though there was a significant degree of overlap between the toxic and safe gene sets across multiple tissues, there were a number of proteins identified that were specifically associated with toxicity or non-toxicity in a single tissue (**Figure 6.2G-H)**. For instance, ABL1 was flagged as a toxic target in all six tissues, whereas KCNJ3 and KCNJ6 – proteins involved in voltage gated potassium channels and the regulation of heartbeats – were only marked as toxic targets in the heart.

To further investigate TSAEs, we expanded the procedure described in **Figure 6.2F** to generate toxic and safe targets for 30 different tissue types – including the 6 prior tested tissues. For each target, we then extracted a set of features to identify any patterns that were consistent across all tested tissues. For each gene, we computed a number of features, including tissue-

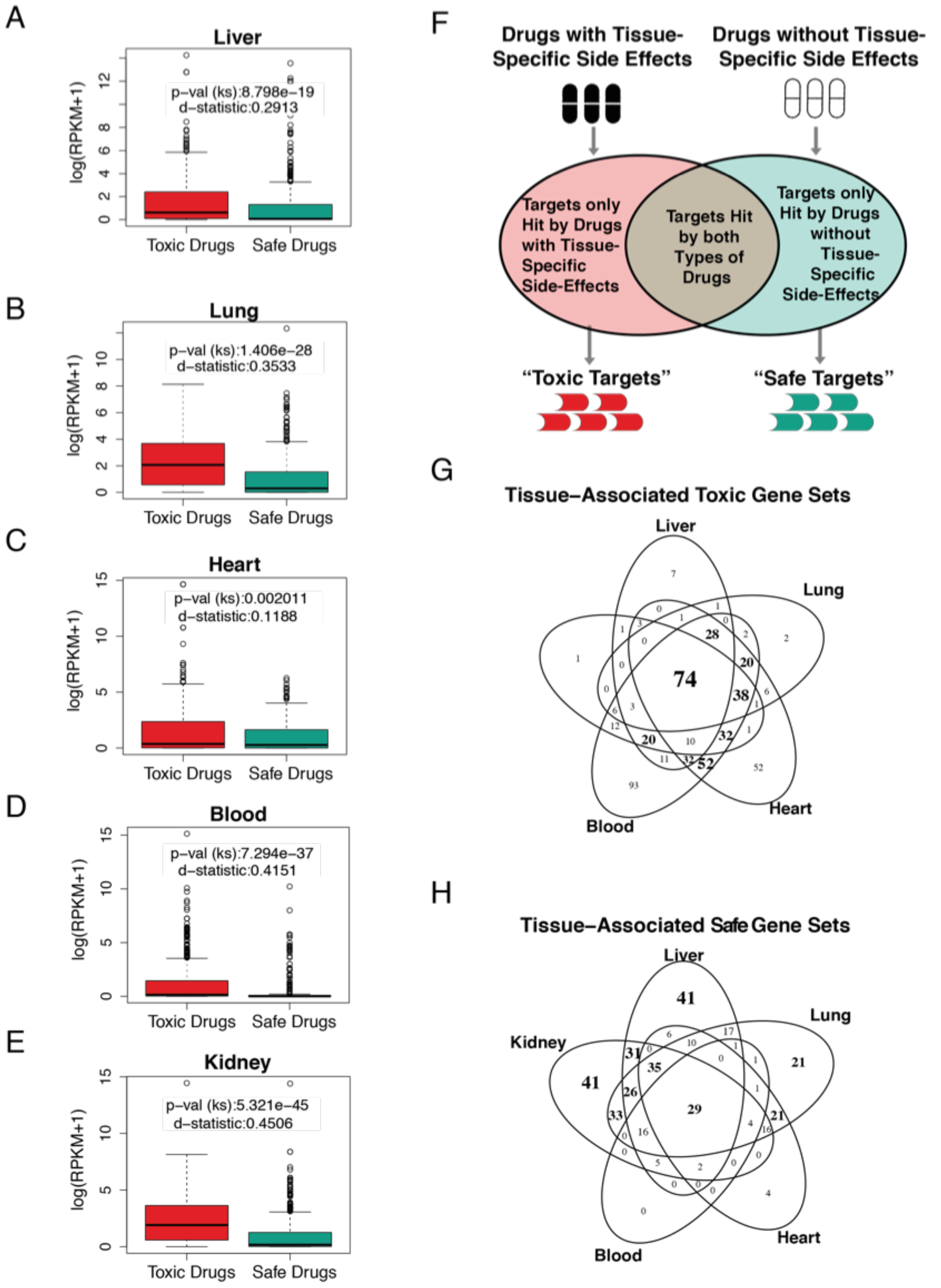

Figure 6.2 – A–E) Distribution of target expression in a specific tissue for drugs with and without any tissue specific adverse events (in that given tissue). F) Schematic for the selection of toxic and safe targets. G)Venn diagram across multiple tissues showing the overlap of toxic and H) safe targets.

specific expression, network properties (betweenness and degree), loss of function mutation frequency, and essentiality status. We found that toxic gene sets tend to be more connected in an aggregated gene-gene interaction network (**Figure 6.3A-B**), be more intolerant for LoF mutations (**Figure 6.3C**), and be enriched for essential genes (**Figure 6.3D**). Finally, we used the ConsensusPathDB framework[180] to measure for GO term enrichment and observed that for toxic gene sets the most commonly enriched terms to had to due with cell death, receptor signaling, and apoptotic processes (**Figure 6.3E**) – pathways one would expect to be related to toxicity – whereas safe targets did not appear to be related to any toxicity related processes (**Figure 6.3F**) – likely due to the diverse nature and function of safe targets.

Computational approach predicts likelihood of specific adverse events

To utilize these findings and more directly address the problem of adverse event prediction, we developed MAESTER (a **M**oneyball **A**pproach for **E**stimating **S**pecific **T**issue adverse **E**vents using **R**andom forests) to compute the probability of a compound presenting with a specific adverse event (**Figure 6.4A**). To do this, we expanded upon the framework of our previous work on predicting broad clinical trial toxicities, PrOCTOR, and narrowed down the classification task to a set of specific adverse events that are correlated with clinical toxicity and have high reported frequencies of fatality in openFDA: drug-induced liver injury (DILI), nephrotoxicity, neutropenia, heart attack, pleural effusion, and pancreatitis. We began by using the framework described in **Fig 6.1A** to define a training set of safe and toxic drugs for each adverse event and its corresponding tissue. For the

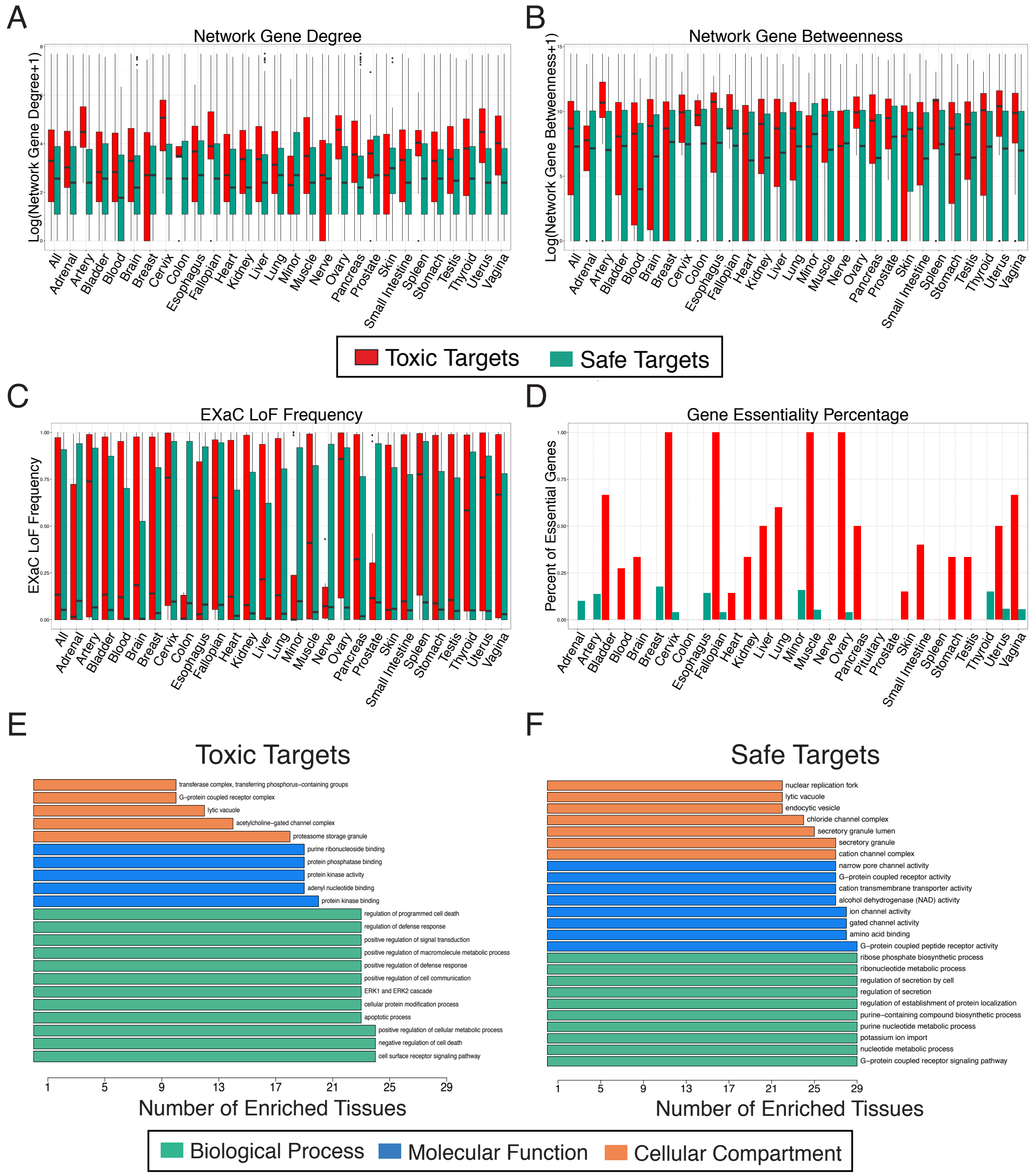

Figure 6.3 – A-D) Distribution of features across multiple tissues for their individual toxic and safe targets. E) Number of tissues whose respective toxic or F) safe targets are enriched for a specific Gene Ontology category.

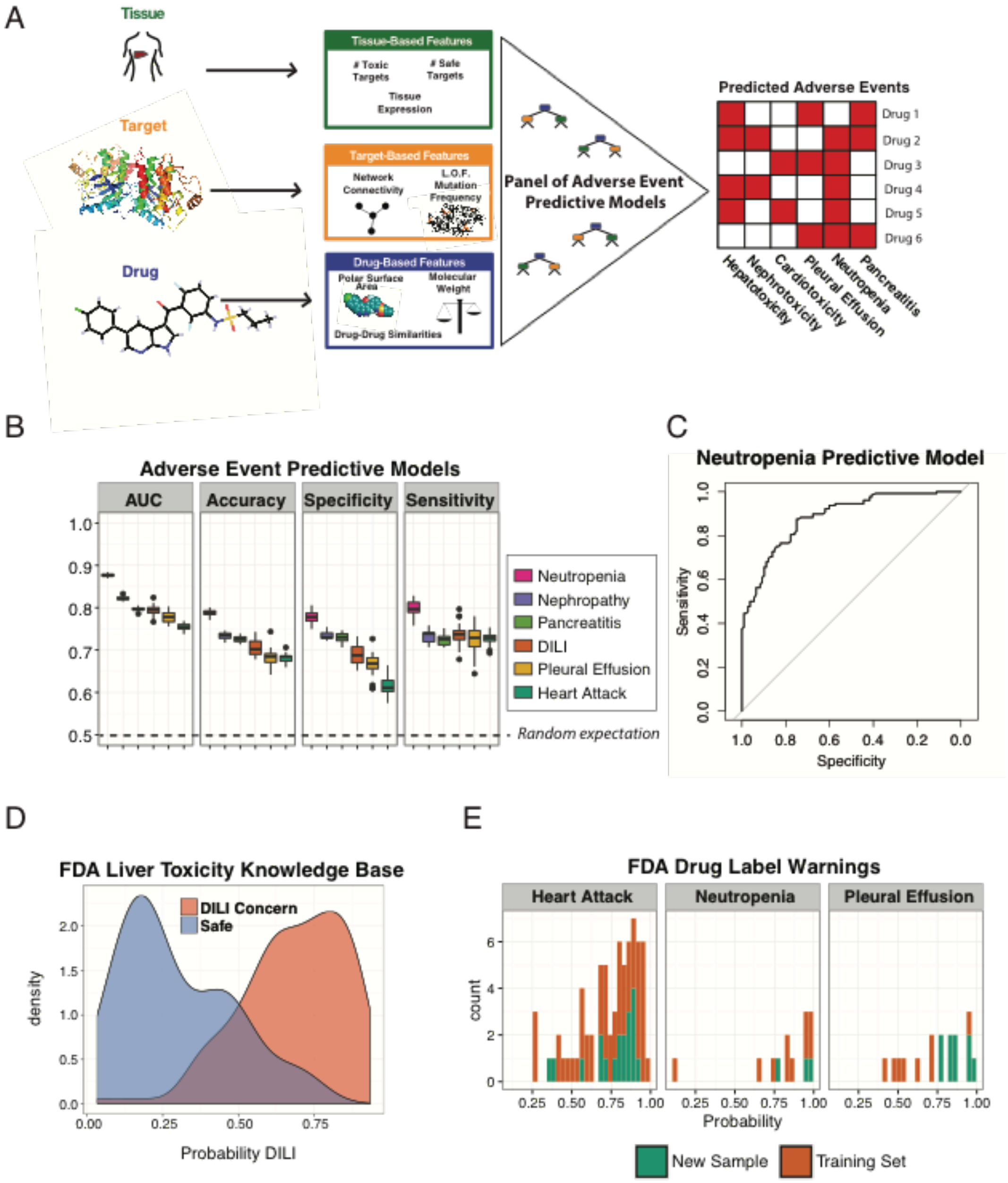

Figure 6.4 – A) Schematic of MAESTER's method of integrating multiple feature types to predict tissue specific adverse events. B) Performance metrics for multiple MAESTER prediction classes. C) Area under the receiver operating curve for MAESTER's Neutropenia model. D) Distribution of MAESTER DILI probabilities for drugs marked as "DILI Concern" or "Safe" by the FDA Liver Toxicity Knowledge base. E) MAESTER Predictions for drugs with FDA warning labels for heart attacks, neutropenia, or pleural effusion.

toxic drugs, we directly queried the database for drugs that are linked to each adverse event or its synonyms. We then took drugs that are not associated with any adverse event in the related tissue or any other severe adverse events (fatality frequency > 13%) to be the set of safe drugs. The set of keywords used to construct these training sets are fully described in **Table 6.1**.

Building upon the framework of PrOCTOR, MAESTER integrates 13 structural features, 35 target and tissue features, and 8 drug similarity properties to produce a suite of classifiers that are able to predict the likelihood of each adverse event (**Figure 6.4A**). Given the established validity of drug-likeness measures in capturing toxicity, we also included properties considered by the Lipinski[178], Veber[181], and Ghose[182] rules, and the Quantitative estimate for Drug-Likeness (Q.E.D.)[179] as well as the measures themselves. For tissue-based features, we considered the number of known drug targets that fall in the associated tissue-specific safe and toxic gene sets we created earlier. We also included the above described tissue expression features from GTEx[183], network properties (connectivity and degree), and loss of function mutation frequency[184]. Finally we integrated the different similarity scores (structural, CMAP, NCI60, and bioassay) through two different measures. The first similarity metric represents whether the drug is more similar to known safe or toxic molecules by using a signed Kolmogorov-Smirnov D-statistic. The second similarity metric is a count of the number of highly similar drugs with known TSAEs.

The classifiers were then evaluated using 10-fold cross validation. All adverse events achieved significant predictive performances with an average accuracy of 72% and area-under-the-receiver-operator curve (AUC) of .81 (**Fig 6.4B**). Focusing specifically on neutropenia – a major cause of clinical trial failure and mortality in cancer and immunocompromised patients[185]– MAESTER achieved an AUC, accuracy, specificity and sensitivity of 0.8843, 0.7839, 0.7778 and 0.7891 respectively – the highest reported results for the computational prediction of neutropenia(**Figure 6.4C)**.

We further assessed MAESTER's performance using an independent validation test set. For liver toxicity, the FDA has curated the Liver Toxicity Knowledge Base (LTKB) that classifies a number of compounds based on their risk of causing liver toxicity. We found that MAESTER can significantly distinguish drugs that are of DILI-concern from those classified as no concern using this independent database (**Figure 6.4D**) (p < 2.2e-16, Mann-Whitney U test). For heart attacks, pleural effusion, and neutropenia we turned to FDA drug label warnings as reported in openFDA (**Figure 6.4E**). We found that MAESTER correctly identified 76.3% of drugs with heart attack risk (p=0.04589, Binomial test), 75.0% with pleural effusion risk (p=0.01474, Binomial test), and 87.5% with neutropenia risk (p=0.0782, Binomial test) (**Figure 6.4E**). None of these tested compounds were in our original training set, further highlighting MAESTER's potential to predict adverse event on new compounds.

A feature importance analysis revealed that there is a subset of features that were consistently predictive across all of MAESTER's adverse event models. The toxic and safe gene sets, structural and bioassay similarity features, polar surface area, and expression of the drug target in mature B cells (centroblasts) are important in a majority of models. We also identified a subset of features that are uniquely predictive in specific models. For example, digestive organs (eg. colon, small intestine, stomach) were highly important in the prediction of DILI, immune-related features (centroblasts, T cells, spleen) were important for neutropenia prediction, and the network degree of the drug target was the most important feature in prediction of pleural effusion.

In order to test MAESTER's ability to detect adverse events that may have been missed by traditional approaches, we examined 7 drugs that had received FDA approval, but were later withdrawn due to previously unknown serious adverse events: Amineptine, Astemizole, Bromfenac, Chlormezanone, Cisapride, Dexfenfluramine, Lumiracoxib. Each of these drugs was run through MAESTER to determine if they were predicted to have the adverse event that eventually led to their withdrawal. We found that MAESTER accurately identified the specific adverse event for each drug (**Table 6.2**). Focusing in on two infamous cases of drug withdrawal – Vioxx and Avandia withdrawn for cardiac toxicity– we found that MAESTER scored each as highly likely to cause cardiac toxicity (**Figure 6.5A-B**). In fact, comparing Avandia to a less toxic analog (Pioglitazone) we observed that difference in reported toxicities corresponded to the difference in their MAESTER scores. Additionally compared to drugs of similar indications that were never withdrawn and were not known to have the reported adverse event, we found that MAESTER produced significantly higher toxicity scores for drugs pulled for cardio or hepatotoxicities (**Figure 6.5C-F**), highlighting its ability to specifically identify compounds with AEs that may be missed by traditional approaches.

Table 6.2 – List of withdrawn drugs, their reason for withdrawal, and the corresponding MAESTER score.

Drug	Reason For Withdrawal	Specific MAESTER Score[a]
Amineptine	Hepatotoxicity	0.862
Astemizole	Cardiac Toxicity	0.53
Bromfenac	Hepatotoxicity	0.526
Chlormezanone	Hepatotoxicity	0.534
Cisapride	Cardiac Toxicity	0.772
Dexfenfluramine	Cardiac Toxicity	0.854
Lumiracoxib	Hepatotoxicity	0.646

[a] = Score corresponds to either cardiac or hepatotoxicity model depending on the reason for withdrawal

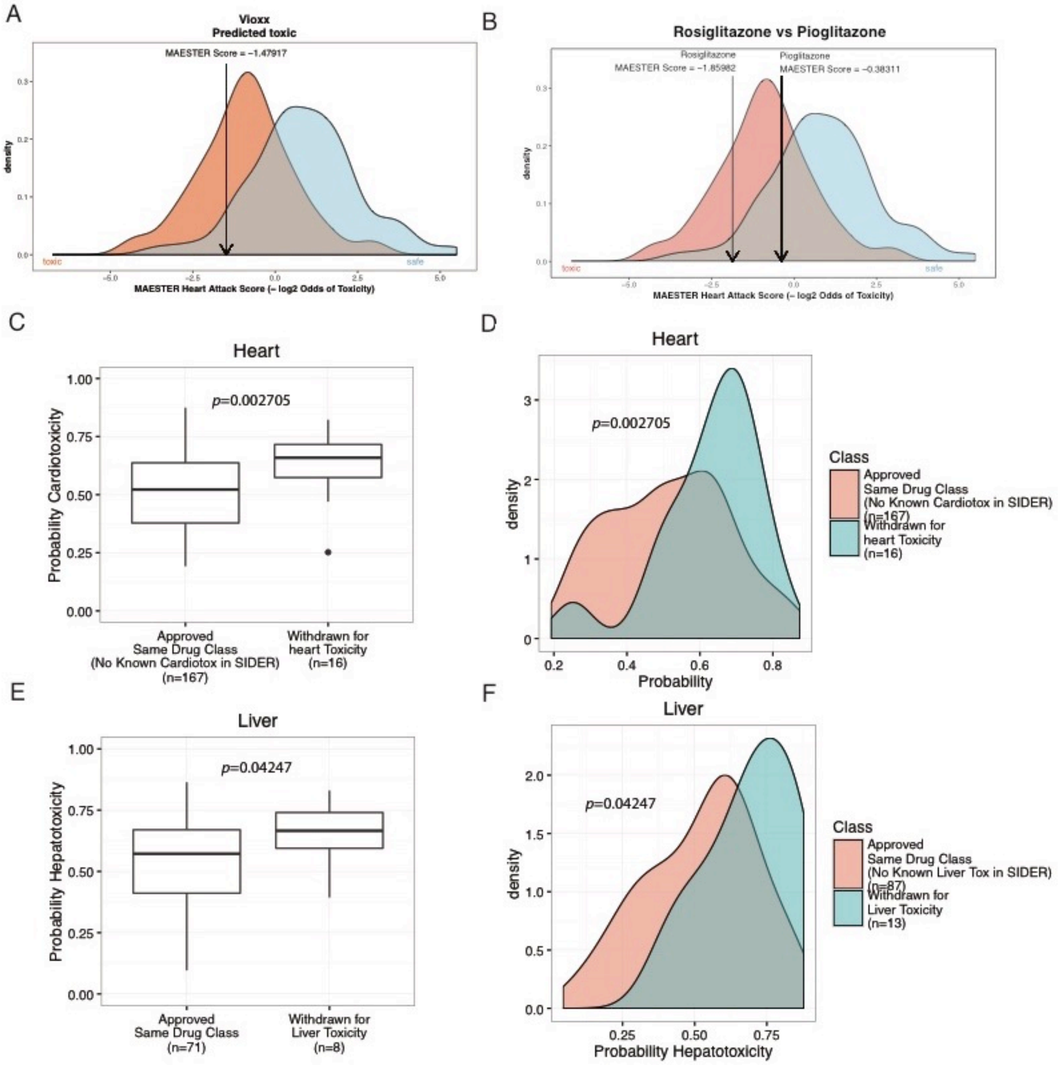

Figure 6.5 – A) Distributions of MAESTER scores for all drugs known to cause heart attacks and those considered safe. MAESTER scores for Vioxx, B) Rosiglitazone, and Pioglitazone are indicated with arrows. C-D) MAESTER scores for drugs withdrawn for cardiac toxicity compared to approved drugs of the same class with no known cardiac toxicities. E-F) MAESTER scores for drugs withdrawn for liver toxicity compared to approved drugs of the same class with no known liver toxicities.

DISCUSSION

Pre-clinical toxicity screening is one of the most important parts of drug development. However, prior computational methods have focused only on molecular properties and predicting broad clinical toxicities rather than specific adverse events. Additionally experimental methods are often cumbersome and often do not translate to clinical results. We have proposed MAESTER, a data-driven machine learning approach that integrates information on a compound's structure, targets, and downstream effects to predict the probability of a compound presenting with different adverse events. When trained on drugs with known adverse events, MAESTER performs at high accuracy, sensitivity, and specificity across six different prediction tasks. Additionally MAESTER performs with high accuracy on external FDA test sets and drug warning labels, and could accurately flag side effects for approved drugs that may have been missed during traditional analyses.

We have identified sets of toxic and safe drugs and genes that are associated with adverse events in specific tissues. We found that tissue-specific toxic drugs tend to be more similar to each other than known safe drugs and that their associated targets are more highly expressed in corresponding tissues. We found tissue-specific toxic targets tend to be enriched for growth related biological processes, more connected in protein-protein interaction networks, and are classified as more essential. Leveraging this data, we developed MAESTER to combine compound and target properties to predict the likelihood of specific adverse events. Because it is trained on drugs with known adverse events, MAESTER can directly predict clinical effects compared to cell or animal screening methods whose toxicity predictions may not translate to the clinic.

One of the strengths of our big data approach is that is able to consider a large number of features without prior bias. This will become especially powerful in the coming years as more large pharmacogenomics datasets become available to integrate. Analysis of these features can aid in future drug design by providing insight into what types of drugs are likely to be toxic and feeding this information back to the chemists. Additionally, while toxicity is often modeled as a broad feature, often times it is a patient specific effect. As more patient specific data becomes available MAESTER can be improved to predict patient specific adverse events. This could be used to guide clinical trial design by specifically selecting patients unlikely to present with toxic effects and radically change how people approach precision medicine.

APPENDIX

MATERIALS AND METHODS – BANDIT (CHAPTER 3)

Datasets:

1. Growth inhibition data: We used publicly available growth inhibition data from the National Cancer Institutes Development Therapeutics Program (NCI-DTP). Each of the NCI60 cell lines were treated with a small molecule and the concentration that caused a 50% decrease in cells was measured. When there were multiple high quality experiments done for the same compound, we averaged the values to obtain a single GI50 value for each small molecule – cell line pair. Contains data on 20,000+ unique compounds. Version 1.6.2 was downloaded from cellminer.com.

2. Gene expression data: All post-treatment gene expression data was downloaded from the Broad Connectivity Map (CMap) project. Fold change data across all cell lines were averaged to obtain a single gene expression signature for each compound. Contains data on 1309 different compounds. Build 02 was downloaded from the Broad CMap Portal.

3. Adverse effects: Side effects (mined from drug package inserts and public information) were downloaded from the SIDER database. Each side effect was classified using the MedDRA (version 16.1) dictionary.

4. Bioassays/Chemical structures: All bioassay results and chemical structures were downloaded from PubChem and organized based on each small molecule's PubChem Compound Identification (CID).

5. Known Drug Targets: All known drug targets were extracted from the DrugBank database (Version 4.1).

Calculating similarity scores:

1. Growth Inhibition Data: For each pair of drugs we calculated a pearson correlation value across the 60 data points (**Figure 3.1**).

2. Gene expression and Chemogenomic Fitness Scores: A pearson correlation was used to measure the degree of similarity for the profiles of two drugs
3. Bioassays: All bioassays were classified as either positive or negative based on the data available in Pubchem. A jaccard index was calculated based on the number of shared "positive" assays between two drugs. We required that each drug pair have been tested in at least one similar assay for a similarity score to be calculated.
4. Chemical Structures: For each drug we extracted the isomeric SMILES and used the atom-pair method [186] to calculate the structural similarity between two compounds (**Figure 3.1**).
5. Adverse Effects: Using the SIDER2 database [111] we extracted the "preferred term" side effects for each drug. A jaccard index was then calculated for the shared side effects for each drug pair.

Calculating correlations between similarity types:

For each pair of similarity scores we separated out drug pairs where both similarity types were measured and plotted the different similarity scores against one another (**Figure 3.2**). We computed the Pearson correlation coefficient (PCC) and the coefficient of determination (R^2) between each pair of similarity scores. Across all pairs, we observed a low correlation – measured by both the PCC and R^2. This finding demonstrated that high similarity of one type does not necessarily implied high similarity in another. Furthermore this indicated that each similarity score could be modeled as an independent variable.

Calculating the Total Likelihood Ratio:

For each data type BANDIT calculates a "likelihood ratio" $L(s_n)$ is defined as the fraction of drug pairs with a shared target (ST pairs) having a given similarity score $s_{n,}$ divided by the fraction of the non-ST pairs with the same similarity score:

Eq. 1:

$$L(s_i) = \frac{\Pr\,(s_i|ST)}{\Pr(s_i|\,non - ST)}$$

Our previous analysis highlighted the minimal correlation between the similarity types and how data types could be modeled independently under a Naïve Bayes framework. This assumption of independence implies that the joint probability of two drugs sharing a target given a set of similarity scores can be modeled as the product involving individual similarity scores.

Therefore the total likelihood ratio L(s) can be expressed as the product of the individual likelihood ratios:

Eq. 2:

$$TLR = L(s) = \prod_{n} L(s_{1-n}) = L(s_1)L(s_2)\dots L(s_n)$$

$$n = maximum\ \#\ of\ included\ datasets$$

The total likelihood ratio (TLR) is then proportional to the odds of two drugs sharing a given target *n* given sources of information

Overall we decided to use this Bayesian framework for multiple reasons, such as the readily interpretable nature of a likelihood ratio compared to other more complicated machine learning scores and the ability to easily add in new data types as they become available.

Testing Against Drugs with Known Targets:

Drug targets were extracted from DrugBank and drug pairs were classified as a “shared-target” pair if they had at least 1 target in common. We used 5-fold cross validation to split our set of drug pairs into a test and training set containing 20% and 80% of the drug pairs respectively. We sub-sampled the two classes (ST and non-ST drug pairs) and required the ratio of true positives (ST pairs) to true negatives (non-ST pairs) to remain the same as the total set. For each fold we computed TLRs for each drug pair in the test set based on the background probabilities within the training set. Each of the 5 test folds combined at the end to produce an ROC Curve and calculate the AUROC value. We also calculated the AUROC value for each individual likelihood ratio from a single data type (**Figure AX.1**).

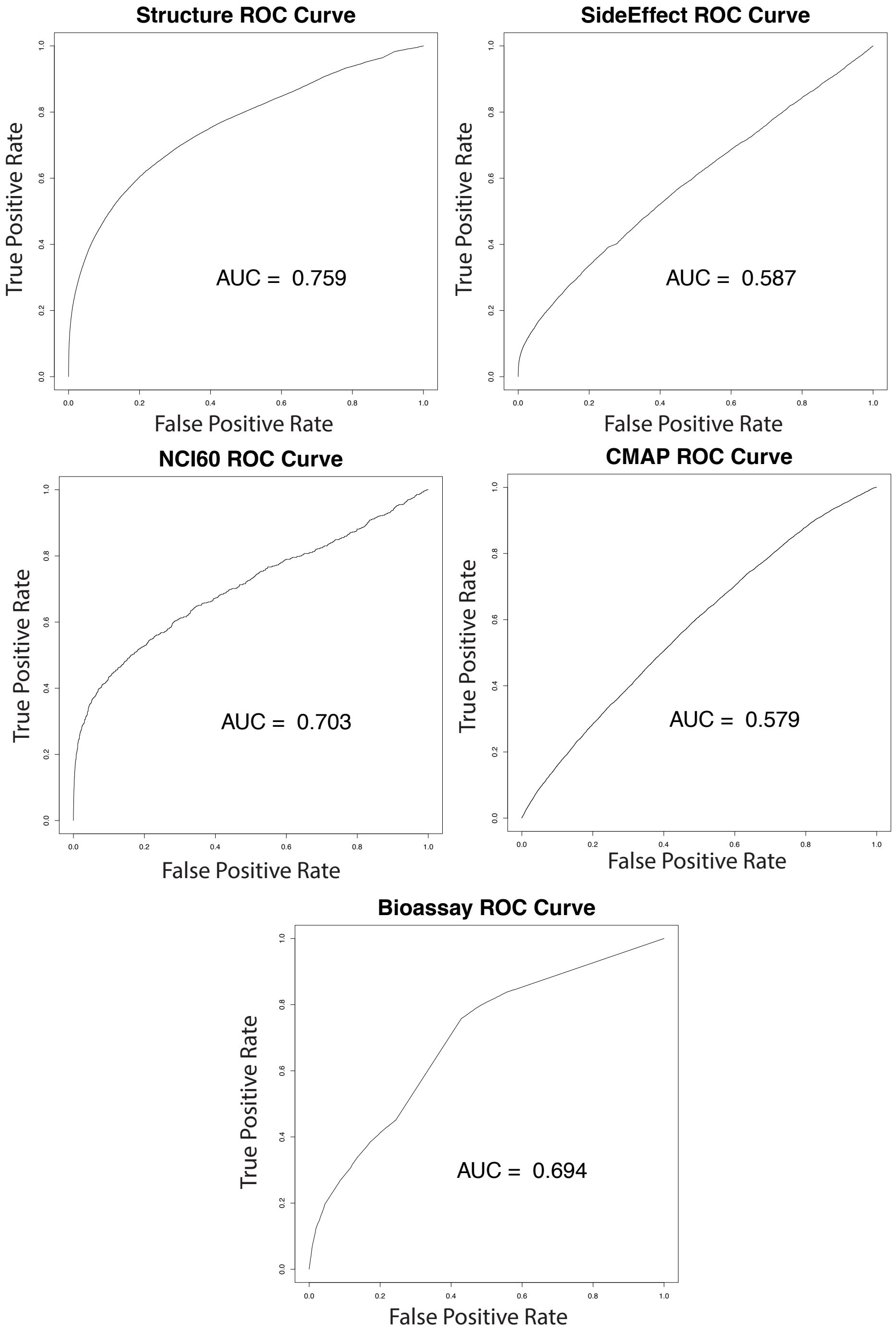

Figure AX.1 – Predictive power of individual data types– Area under the receiver-operating curve for different data type specific likelihood ratios.

We performed this analysis with the TLR output while varying the number of data types being considered and found a significant increase in the predictive power, measured by the AUROC, as we increased the number of included datasets (**Figure 3.3A**). We computed two sets of ROC curves – one where we required drugs have available data in each included data type (our preferred method) and another where we imputed the data type median for each missing data type. We varied the order in which datasets were added and observed a positive relationship between AUROC value and the number of included data types regardless of the addition order. Furthermore we used a KS test to measure how our TLR value could separate out ST and non-ST pairs and saw that in each case our TLR value outperformed any individual variable (**Figure AX.2**). We repeated this analysis increasing the minimum number of data types we required a pair of compounds to have and saw the separation steadily improve (D = .44 to .69).

Replicating Kinase Experimental Screen

We first separated out the kinases in the Peterson et al. database that were classified as BANDIT orphan small molecules – molecules that were in at least two of the considered BANDIT databases and had no known targets. For each orphan kinase inhibitor we used BANDIT to predict shared target drugs. Each known kinase target of the shared target drugs was classified as a potential kinase target of the orphan inhibitor. We then observed that the "percent remaining kinase activity" was significantly lower between the orphan kinase inhibitors and the BANDIT predicted kinases than between the orphan inhibitors and any non-predicted kinases (Wilcoxon Rank Sum Test P = 3.62e−06) (**Figure 3.4**).

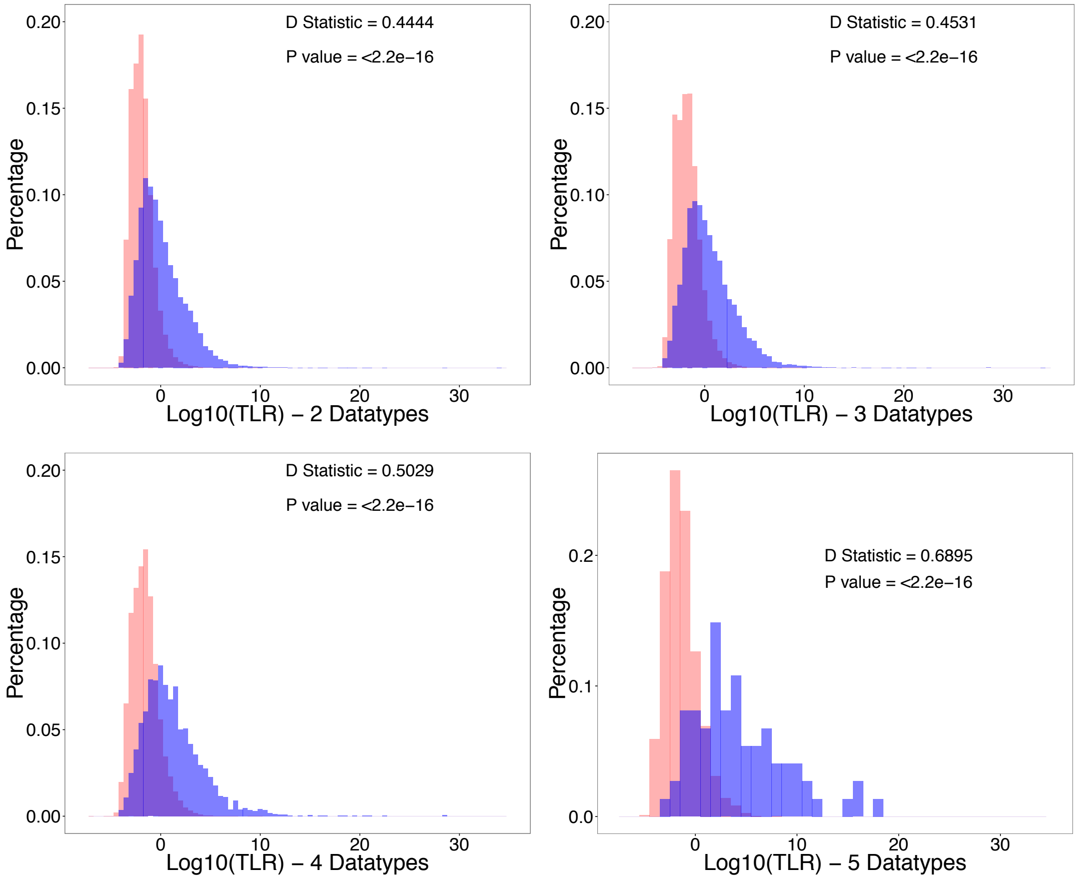

Figure AX.2 – BANDIT's TLR output accurately separates drug pairs with shared targets– Distributions of TLR scores across two sets – drug pairs known to share a target and those with no known shared targets – with increasing requirements on the number of overlapping data types. P values and D statistics were calculated using the Kolmogorov-Smirnov test. Blue = Shared target drug pairs; Pink = No shared target pairs.

Specific Target Voting

For each orphan small molecule we identified all shared target drug predictions, or any drugs with known targets that exceeded a given BANDIT likelihood ratio. For each shared target drug prediction, we compiled all known targets of that given drug and ranked specific protein targets based on how often it appeared as known target in shared drug target predictions. “Votes” for particular protein targets were weighted based on the likelihood ratio of the shared target prediction they originated from. The top voted target for each orphan small molecule that we tested was then predicted to be a novel specific target (**Figure 3.3E**).

To test the accuracy, we used leave-one-out cross validation on our test set of drugs with known targets. For each drug we used BANDIT to compare it to all other drugs with known targets and identify the top ranked target for the tested drug. This was repeated for every drug in our test set and we calculated how often the top ranked target was a known target of the drug being tested. We recomputed these accuracies while varying the likelihood ratio cutoff for a drug pair to be considered a shared-target prediction. As expected we observed a steady rise in accuracy as we increased the cutoff value, with the accuracy plateauing at an accuracy level of approximately 90% – revealing that BANDIT’s voting protocol could accurately identify specific targets (**Figure 3.3F**).

Identification of Novel Anti-Microtubule Small Molecules

For each orphan small molecule in BANDIT (defined as a molecule tested in any of the individual databases but without any known targets in DrugBank) we used the BANDIT voting protocol to predict specific protein targets. We required that each orphan small molecule be in at least 3 of BANDIT's databases, leaving us with a set of ~15,000 small molecules. To refine our initial list of predictions into a high confidence set, we required a TLR cutoff of 500, that each predicted target appear in the majority of shared target predictions, and that the highest ranked target appear in the top shared target prediction for each orphan molecule. From this list of high confidence predictions we identified a set of small molecules predicted to bind to microtubules.

For each predicted microtubule inhibitor (MTI) we examined how it related to known MTIs using a network approach (**Figure 3.5**). We required that each predicted MTI have a TLR greater than 500 with at least two known MTIs. Each edge in our network represents a predicted shared target interaction with the length and width of each corresponding to the strength of the prediction (measured by the TLR value). We used the Fruchterman Reingold projection within the R igraph package. We observed a distinct clustering of known MTIs based on their mechanism of action.

Most of the novel MTIs we predicted were not easily obtained, thus we specifically focused on the subset that we could obtain from the National Cancer Institutes Developmental Therapeutics Program.

Microtubule Imaging/Testing

Human breast MDA-MD-231 cells were cultured in DMEM (obtained from Corning Cellgro) with 10% fetal bovine serum and 1% penicillin and streptomycin. Cells were plated at the density of 90,000 Cells/ml onto 12mm round cover slips in 48 well plates for 24 hours and then treated for 6 hours with small molecules at the given concentrations. Small molecules (obtained from the NCI Drug Bank) were dissolved in DMSO and stored at -20^{o}C. Control experiments were done using DMSO and it was less than 0.5% of total media volume. After 6hrs drug treatment media was removed and cells were per-meabilized with 0.5% Triton X-100 and fixed with PHEMO Buffer (3.7% formaldehyde, 0.05% glutaraldehyde, 0.068M Pipes, 0.025M HEPES, 0.015M EGTANa$_2$, 0.003M MgCl$_2$6H2O and 10% DMSO and adjust pH=6.8) for 10minutes. Fixed cells were washed three times with PBS buffer. Cells were blocked with 10% goat serum at room temperature for 10 minutes. Cells were incubated with monoclonal α-tubulin antibody (clone YL 1/2, obtained from EMD Millipore), for 1hr and washed three times with PBS buffer before incubation with a secondary Alexa Fluor 488 goat anti-mouse antibody (obtained from Invitrogen). Cell chromatin was stained with DAPI for 5min and washed with water three times. Cover slips were mounted and photographed in a RSM 700 microscope for microtubule visualization. DNA was counterstained with DAPI. Images were acquired with Zeiss LSM 700 confocal microscope under a 63×/1.4NA objective (Zeiss, Germany).

A Fisher's exact test was used to determine whether the number of observed successes – defined as a predicted microtubule inhibitor showing an effect against microtubules in imaging – was greater than what would be expected by random chance. To determine the background probability we used the number of drugs with known targets in our database that were known to target microtubules (~ 1%).

Microtubule Effect Quantification

Following 6hrs treatment, cells (12 well plate) were washed once with warm phosphate-buffered saline. Each well was incubated with 150 µL either with low salts or high salt buffer at 37 ^{o}C for 10 minutes. Cell were then scraped and were either lysed in low salt buffer to test for the degree of tubulin polymerization (20 mM Tris–HCl pH 6.8, 1 mM $MgCl_2$, 2 mM EGTA, 0.5% NP-40, 1X protease inhibitor cocktail and 0.5% NP-40) or high salt buffer to test for the degree of tubulin depolymerization (0.1M Pipes, 1mM EGTA, 1mM MgSO4, 30% glycerol, 5% DMSO, 1mM DTT, 0.02% NAAzide, 0.125% NP-40, 1mM DTT and 1X protease inhibitor cocktail). Samples were spun at max speed in a tabletop centrifuge for 30 min at room temperature. The supernatant (S) was separated from the pellet (P). The pellet was resuspended in 150 µL 1 × Laemmli buffer and sonicated. Equal volumes of supernatant and pellet samples were loaded onto a 12% gel for a western blot. Tubulin bands were visualized with a DM1α monoclonal antibody (obtained from Sigma-Aldrich). % Tubulin in pellet levels were calculated as the densitometric value of the pellet band divided by the total densitometric value of the pellet and supernatant bands times 100. Three biological repeats were performed (**Figure AX.3**).

Imaging of Treatment Against Resistant Cell Lines

1A9-ERB is a clone of the 1A9 human ovarian carcinoma cell line resistant to the effects of Eribulin mesylate. It was prepared by exposing 1A9 cells to 1ng/ml Eribulin (obtained from Eisai pharmaceuticals) in the presence of 10ug/ml verapamil (obtained from Acros Organics), a Pgp antagonist. The cells were maintained in the 0.5ng/ml eribulin and 10ug/ml verapamil

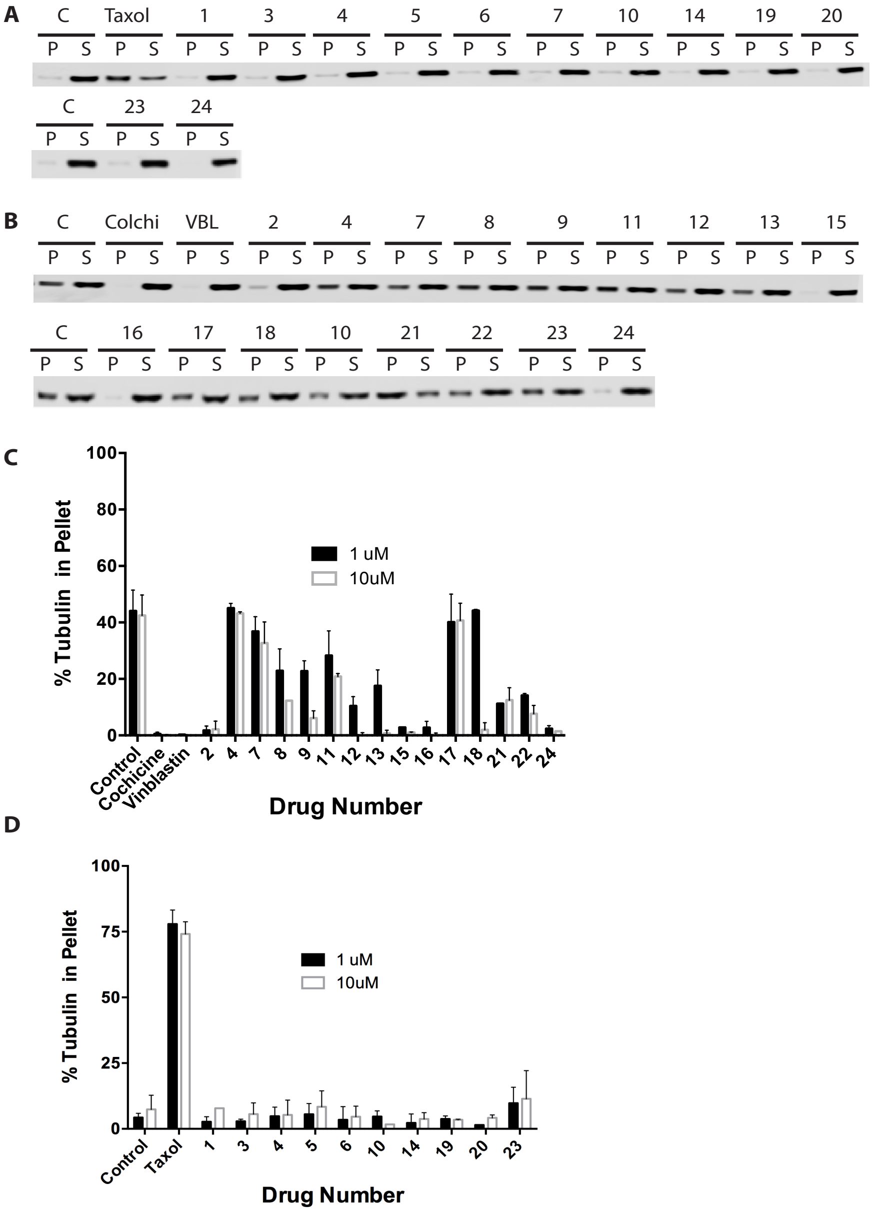

Figure AX.3 – Quantification experiments also backed up the activity of BANDIT predicted inhibitors on microtubules – Effect of drugs microtubule integrity of MDA-MB-231 cells after 6 hours of treatment. A) Western blots for supernatant (S) and sellet (P) fractions were examined by SDS-PAGE for MDA-MB-231 cells after 6 hours (1µM) of treatment for polymerizing drugs, B) Western blots for supernatant (S) and sellet (P) fractions were examined by SDS-PAGE for MDA-MB-231 cells after 6 hours (1µM) of treatment for depolymerizing drugs, C) Bar graph showing the % of tubulin in the pellet compared to the supernatant (averaged over three independent replicates) for depolymerizing drugs at 1 and 10 µM, and D) Bar graph showing the % of tubulin in the pellet compared to the supernatant (averaged over three independent replicates) for polymerizing drugs at 1 and 10 µM.

concentrations. Cells were removed from this drug solution 3 days prior to any future experimentation. Additional treatment and imaging was done using the same protocols as described earlier.

Characterization of ONC201-DRD2 Interaction

ONC201 dihydrochloride was obtained from Oncoceutics. Kinase inhibition assays for the kinome were performed as previously described [187]. GPCR arrestin recruitment and cAMP modulation reporter assays were performed as previously described [188]. PathHunter^{TM} (DiscoveRx) beta-arrestin cells expressing one of several GPCR targets were plated onto 384-well white solid bottom assay plates (Corning 3570) at 5000 cells per well in a 20 μL volume in the appropriate cell plating reagent. Cells were incubated at 37 °C, 5% CO_2 for 18-24 h. Samples were prepared in buffer containing 0.05% fatty-acid free BSA (Sigma). For agonist mode tests, samples (5 μL) were added to pre-plated cells and incubated for 90 minutes at 37 °C, 5% CO_2. For antagonist mode tests, samples (5 μL) were added to pre-plated cells and incubated for 30 minutes at 37 °C, 5% CO_2 followed by addition of EC80 agonist (5 μL) for 90 minutes at 37 °C, 5% CO_2. For Schild analysis, samples (5 μL) were added to pre-plated cells and incubated for 30 minutes at 37 °C, 5% CO_2 followed by addition of serially dliuted agonist (5 μL) for 90 minutes at 37 °C, 5% CO_2. Control wells defining the maximal and minimal response for each assay mode were tested in parallel. Arrestin recruitment was measured by addition of 15 μL PathHunter Detection reagent and incubated for 1-2 h at room temperature and read on a Perkin Elmer Envision Plate Reader. For agonist and antagonist tests, data was normalized for percent efficacy using the appropriate controls and fitted to a sigmoidal dose-response (variable slope), Y=Bottom + (Top-Bottom)/(1+10^((LogEC50-X)*HillSlope)), where X is the log concentration of compound.

For Schild analysis, data was normalized for percent efficacy using the appropriate controls and fitted to a Gaddum/Schild EC50 shift using global fitting, where Y=Bottom + (Top-Bottom)/(1+10^((LogEC-X)*HillSlope)), Antag=1+(B/(10^(-1*pA2)))^SchildSlope and LogEC=Log(EC50*Antag). EC50 / IC50 analysis was performed in CBIS data analysis suite (Cheminnovation) and Schild analysis performed in GraphPad Prism 6.0.5.

The kinase assay and nuclear hormone receptor profiling (S16) were performed as previously described by Reaction Biology Corp and DiscoverX respectively [189-191].

Drug Mechanism Clustering

For each drug pair we converted the TLR between them into a distance metric used to estimate "closeness" between any two drugs:

Eq. 3:

$$BANDIT\ Distance\ Score = \frac{1}{TLR}$$

We next separated all drugs know to target microtubules that were in at least 3 of BANDIT's dataset. With the BANDIT distance metric as an input we created a hierarchical cluster of all known MTIs using the hclust R method with an "average" based clustering method. Known MTIs were labeled based on whether they were known to polymerize or depolymerize microtubules, and we observed a distinct separation based on the mechanism of action (MoA). We repeated this clustering while removing drug structures from our likelihood calculations and continued to see a MoA-based separation. This revealed that BANDIT's clustering approach is not dependent on any single data type, and that observed results are due to BANDIT's integrative approach. This analysis was then repeated using similar conditions for known protein kinases.

Drug "Universe" Clustering

Using the same protocol as was used to create the MTI network, we created a network of all drugs with known targets with each edge representing a predicted shared target interaction and the edge weight corresponding to the strength of the interaction. Using the KEGG drug database[192] and DrugBank[113] we annotated each drug based on its most prevalent ATC code and colored each drug accordingly. We specifically isolated out 3 clusters representing: 1) beta-blockers with Parkinson's medications, 2) antiretrovirals and statins and 3) opioids and microtubule inhbitors.

To get a better understanding of how orphan small molecules fit into this drug "universe" we computed the distance between every pair of small molecules and used multi-dimensional scaling to visualize the overall structure. We used the same distance metric as described in the mechanism of action clustering section to create a distance matrix between all small molecules (known drugs and orphan) and used the R cmdscale package for the multi-dimensional scaling. We noticed a definite structure with known drugs tightly clustering around each other, while orphan molecules had a more diffuse organization. One explanation for this structure is that drugs with known targets are more likely to be used to treat patients and thus may have similar effects due to safety precautions, whereas orphan molecules which have not gone through clinical trials and FDA approval are more likely to have a wide variety of effects and characteristics.

MATERIALS AND METHODS – ONC201 CSC ANALYSIS (CHAPTER 5)

Cell culture and reagents

HCT116 p53-/- cells were kind gifts from Dr. Bert Vogelstein of Johns Hopkins University. ONC201 resistant RKO cells were generated previously in our lab in 2012-2013 [193]. All other cell lines were obtained from the American Type Culture Collection and cultured as previously described [193,144]. Cells were authenticated every month by growth and morphological observation. ONC201 was provided by Oncoceutics, Inc.

Tumorsphere culture

Tumorspheres were cultured as described previously [161] under non-adherent growth conditions in Ultra Low attachment plates (Corning) using the MammoCult™ Human Medium (STEMCELL Technologies) as per the manufacturer's protocol. Cells (1000-20,000 per well) were seeded medium containing DMSO or ONC201. Colonospheres of size > 60 μm were counted.

Patient-derived glioblastoma cells

Four lines were derived using neurosphere culture from untreated (GBM8, GBM18) and recurrent (GBM67R and GBM152) glioblastomas. Cell viability assays were performed using indicated concentrations of ONC201 and IC50 values were calculated.

Gene expression profiling and network analysis

Gene expression profiling of HCT116, RKO and ONC201-resistant RKO cells with DMSO or ONC201 treatment for indicated time points was performed in previous studies and data from these microarray studies are submitted to NCBI Gene Expression Omnibus [193,144]. For network analysis of stem cell-related transcriptional changes induced by ONC201, the dataset was analyzed with the Ingenuity Pathway Analysis software.

Quantitative RT-PCR (qRT-PCR)

Total RNA was isolated using the Quick-RNA™ MiniPrep kit (Zymo Research, Irvine, CA). 5μg of total RNA from each sample was subjected to cDNA synthesis using SuperScript® III Reverse Transcriptase kit (Life technologies, Grand Island, NY). The relative expression of the reported stem-cell markers was determined using real-time PCR performed on Applied Biosystems 7900HT Fast Real-Time PCR system. Each cDNA sample was amplified using Power SYBR Green (Applied Biosystems, CA). Briefly, the reaction conditions consisted of 0.4 μL of cDNA and 0.2 μM primers in a final volume of 10 μL of qPCR mix. Each cycle consisted of denaturation of 95^{o}C for 15 s, annealing at 60^{o}C for 15 s and extension at 72^{o}C for 1 min. Each cycle was followed by dissociation curves for every sample. GAPDH was used as an endogenous control to normalize each sample. At least two different independent experiments were performed for each result with triplicates per experiment.

Western blot

Western blotting was performed as described previously [144,193,161]. The following antibodies were used: CD44 (Cell Signaling), ALDH (BD), ID1 (Santa Cruz), ID2 (Santa Cruz), ID3 (Santa Cruz), CD133 (Santa Cruz Biotechnology), WNT16 (BD) and Ran (BD). Horseradish peroxidase labeled secondary antibodies were from Pierce.

Analysis of gene expression data from genomic of drug sensitivity in cancer (GDSC) cell line screening

Cell viability assays were performed with GDSC cell lines (1000 human cancer cell lines) at 72 hours post-ONC201 treatment to generate dose responses curves at concentrations from 78 nM up to 20 μM as described previously [159]. Gene expression data was downloaded from the COSMIC Cell Lines Project using an Affymetrix Human Genome U219 Array platform. GDSC cell lines were separated in low and high expression groups based on a Z-score cutoff of -1 and 1 respectively. Data were analyzed to generate IC50. A Kolmogorov–Smirnov test (using the ks.test method in the R statistical programming language) was used to test statistical significance with the accompanying D statistic used to measure the degree of separation between the two groups.

Other statistical analysis

Data are presented as the mean ± standard deviation or standard error of mean from at least three replicates. The Student's two-tailed t-test was used for pairwise analysis. Statistically significant changes (*) are indicated in the figures with p-values.

MATERIALS AND METHODS – MAESTER (CHAPTER 6)

Training Set:

We downloaded the Side Effect Resource (SIDER) database from sideeffects.embl.de. The *meddra_adverse_effects.txt* table was used to extract reported adverse events and the *MedDRA Preferred Term* descriptor to group similar side effects. For each of the 30 major tissue types reported in the Genotype-Tissue Expression (GTEx) project, we identified the drugs that are associated with any adverse events by searching for keywords associated with the tissue name. We used the openFDA resource (https://open.fda.gov/) to identify drugs in SIDER that are associated with high fatality rates, which we defined to be greater than 13% of reported cases. We further extracted the names of the drugs associated with specific adverse events, including synonyms listed in Table 6.1. To define a set of drugs not associated with a given adverse event, we took the remaining drugs in SIDER and removed those that are associated with any other toxicity in the tissue, according to terms listed in Table 6.1, or high fatality rates, as defined using the openFDA resource above.

Feature Derivation

1. **Chemical Features** – The structures (sdf format) were downloaded for all of the drugs in DrugBank. The molecular weight, polar surface area, hydrogen bond donor and acceptor counts, formal charge and number of rotatable bounds were extracted from the sdf file for each of these compounds. When that information was missing, it was filled in using PubChem or by computationally estimating these values using

ChemmineR in R. Drug-likeness rule outcomes for the Lipinski, Veber, and Ghose rules were derived using these features. The QED values were computed using the author-released script.

2. **Network features** – We constructed the aggregated biological network by taking the union across multiple databases of gene-gene interactions. [194-196]. The network degree of a gene was calculated as the number of gene neighbors that a particular gene has. The network betweenness for a particular gene (i.e. vertex) is defined as the number of shortest paths that travel through the vertex. For drug, we considered the maximum network degree and betweenness of its target genes. These measures were calculated using R's igraph package[197].
3. **Tissue features** – The Genotype-Tissue Expression (GTEx) project[183] dataset of 2921 RPKM RNA-Seq samples were downloaded from http://www.gtexportal.org/home/.

   For each tissue, the median RPKM was calculated for each gene. For each drug, the maximum RPKM of its targets was considered.
4. **Target Loss Frequency** – The Exome Aggregation Consortium database [184] was downloaded from www.exac.broadinstitute.org. The loss frequency was calculated to be percentage of deleterious mutations for each gene.

Drug Similarities (See Chapter 3)

1. Growth Inhibition Data: For each pair of drugs we calculated a pearson correlation value across the 60 data points (Figure 2.1).

2. Gene expression and Chemogenomic Fitness Scores: A pearson correlation was used to measure the degree of similarity for the profiles of two drugs.

3. Bioassays: All bioassays were classified as either positive or negative based on the data available in Pubchem. A jaccard index was calculated based on the number of shared “positive” assays between two drugs. We required that each drug pair have been tested in at least one similar assay for a similarity score to be calculated.

4. Chemical Structures: For each drug we extracted the isomeric SMILES and used the atom-pair method [186] to calculate the structural similarity between two compounds (Figure S1).

The MAESTER Approach

For each adverse event listed in Table 6.1, we trained a model using the training set and features described above. It was trained using the random forest model, an ensemble decision tree based approach, which constructs 50 bootstrapped decision trees. A sub-sampling approach was used to account for any imbalance in the ratio of toxic drugs to safe drugs, by randomly down-sampling the larger class of samples. To reduce the odds of poor representatives being sampled, this was repeated 30 times. The labels were assigned by taking the consensus across the set of bootstrapped trees and replicates. This approach also yields a probability for each test sample. This probability was used to calculate an odds score =

$$\log_2\left(\frac{P(\text{approval})}{P(\text{failure})}\right).$$

Independent Datasets

The FDA annotated drug-induced liver toxicity (DILI) dataset was downloaded from the FDA website at:

http://www.fda.gov/ScienceResearch/BioinformaticsTools/LiverToxicityKnowledgeBase/ucm226811.htm.

The openFDA resource (https://open.fda.gov) was used to identify drugs that have warning labels for the relevant toxicity events. We extracted the set of drugs that have been withdrawn for known liver or cardiotoxicity reasons from DrugBank descriptions for withdrawn drugs and supplemented with information curated from:

https://en.wikipedia.org/wiki/List_of_withdrawn_drugs.

These were compared to MAESTER predictions for all drugs of the same class that have not been previously annotated in SIDER for the given adverse event.

Statistical Analyses

We used area under the receiver operating characteristic (ROC) curve and 10-fold cross validation to evaluate the predictive power of our approach. For the analysis of predictions for the FDA drug warning dataset, we tested for enrichment of predictions using the binomial test. We tested for differences in predictions in the FDA DILI dataset and between classes for the withdrawn drug datasets using the unpaired Student's t-test.

References

1. Panahiazar M, Taslimitehrani V, Pereira N, Pathak J (2015) Using EHRs and Machine Learning for Heart Failure Survival Analysis. Stud Health Technol Inform 216:40-44

2. Gertych A, Ing N, Ma Z, Fuchs TJ, Salman S, Mohanty S, Bhele S, Velasquez-Vacca A, Amin MB, Knudsen BS (2015) Machine learning approaches to analyze histological images of tissues from radical prostatectomies. Comput Med Imaging Graph 46 Pt 2:197-208. doi:10.1016/j.compmedimag.2015.08.002

3. Shimura H, Masuda S, Kimura H (2014) Research and development productivity map: visualization of industry status. J Clin Pharm Ther 39 (2):175-180. doi:10.1111/jcpt.12126

4. Drug development costs jump to $2.6 billion (2015). Cancer Discov 5 (2):OF2. doi:10.1158/2159-8290.CD-NB2014-188

5. Drews J (2000) Drug discovery: a historical perspective. Science 287 (5460):1960-1964

6. Swinney DC, Anthony J (2011) How were new medicines discovered? Nat Rev Drug Discov 10 (7):507-519. doi:10.1038/nrd3480

7. Lee J, Bogyo M (2013) Target deconvolution techniques in modern phenotypic profiling. Curr Opin Chem Biol 17 (1):118-126. doi:10.1016/j.cbpa.2012.12.022

8. Mestres J, Gregori-Puigjane E, Valverde S, Sole RV (2009) The topology of drug-target interaction networks: implicit dependence on drug properties and target families. Molecular bioSystems 5 (9):1051-1057. doi:10.1039/b905821b

9. Butina D, Segall MD, Frankcombe K (2002) Predicting ADME properties in silico: methods and models. Drug Discov Today 7 (11):S83-88

10. Li HL, Gao ZT, Kang L, Zhang HL, Yang K, Yu KQ, Luo XM, Zhu WL, Chen KX, Shen JH, Wang XC, Jiang HL (2006) TarFisDock: a web server for identifying drug targets with docking approach. Nucleic acids research 34:W219-W224. doi:10.1093/nar/gkl114

11. Rarey M, Kramer B, Lengauer T, Klebe G (1996) A fast flexible docking method using an incremental construction algorithm. Journal of molecular biology 261 (3):470-489. doi:10.1006/jmbi.1996.0477

12. Hay M, Thomas DW, Craighead JL, Economides C, Rosenthal J (2014) Clinical development success rates for investigational drugs. Nat Biotechnol 32 (1):40-51. doi:10.1038/nbt.2786

13. Brennan CW, Verhaak RGW, McKenna A, Campos B, Noushmehr H, Salama SR, Zheng SY, Chakravarty D, Sanborn JZ, Berman SH, Beroukhim R, Bernard B, Wu CJ, Genovese G, Shmulevich I, Barnholtz-Sloan J, Zou LH, Vegesna R, Shukla SA, Ciriello G, Yung WK, Zhang W, Sougnez C, Mikkelsen T, Aldape K, Bigner DD, Van Meir EG, Prados M, Sloan A, Black KL, Eschbacher J, Finocchiaro G, Friedman W, Andrews DW, Guha A, Iacocca M, O'Neill BP, Foltz G, Myers J, Weisenberger DJ, Penny R, Kucherlapati R, Perou CM, Hayes DN, Gibbs R, Marra M, Mills GB, Lander E, Spellman P, Wilson R, Sander C, Weinstein J, Meyerson M, Gabriel S, Laird PW, Haussler D, Getz G, Chin L, Network TR (2013) The Somatic Genomic Landscape of Glioblastoma. Cell 155 (2):462-477. doi:10.1016/j.cell.2013.09.034

14. Cancer Genome Atlas Research N, Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM (2013) The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet 45 (10):1113-1120. doi:10.1038/ng.2764

15. Ledford H (2011) Translational research: 4 ways to fix the clinical trial. Nature 477 (7366):526-528. doi:10.1038/477526a

16. Leeson PD, Springthorpe B (2007) The influence of drug-like concepts on decision-making in medicinal chemistry. Nat Rev Drug Discov 6 (11):881-890. doi:10.1038/nrd2445

17. Gayvert KM, Madhukar NS, Elemento O (2016) A Data-Driven Approach to Predicting Successes and Failures of Clinical Trials. Cell Chem Biol 23 (10):1294-1301. doi:10.1016/j.chembiol.2016.07.023

18. Shanks N, Greek R, Greek J (2009) Are animal models predictive for humans? Philos Ethics Humanit Med 4:2. doi:10.1186/1747-5341-4-2

19. Fry RC, Svensson JP, Valiathan C, Wang E, Hogan BJ, Bhattacharya S, Bugni JM, Whittaker CA, Samson LD (2008) Genomic predictors of interindividual differences in response to DNA damaging agents. Gene Dev 22 (19):2621-2626. doi:10.1101/gad.1688508

20. Rice SD, Heinzman JM, Brower SL, Ervin PR, Song N, Shen K, Wang DK (2010) Analysis of Chemotherapeutic Response Heterogeneity and Drug Clustering Based on Mechanism of Action Using an In Vitro Assay. Anticancer Res 30 (7):2805-2811

21. Bosquet JG, Marchion DC, Chon H, Lancaster JM, Chanock S (2014) Analysis of chemotherapeutic response in ovarian cancers using publicly available high-throughput data. Cancer research 74 (14):3902-3912. doi:10.1158/0008-5472.CAN-14-0186

22. Rice SD, Heinzman JM, Brower SL, Ervin PR, Song N, Shen K, Wang D (2010) Analysis of chemotherapeutic response heterogeneity and drug clustering based on mechanism of action using an in vitro assay. Anticancer Res 30 (7):2805-2811

23. Sboner A, Elemento O (2016) A primer on precision medicine informatics. Brief Bioinform 17 (1):145-153. doi:10.1093/bib/bbv032

24. Shoemaker RH (2006) The NCI60 human tumour cell line anticancer drug screen. Nature Reviews Cancer 6 (10):813-823. doi:10.1038/nrc1951

25. Abaan OD, Polley EC, Davis SR, Zhu YJ, Bilke S, Walker RL, Pineda M, Gindin Y, Jiang Y, Reinhold WC, Holbeck SL, Simon RM, Doroshow JH, Pommier Y, Meltzer PS (2013) The Exomes of the NCI-60 Panel: A Genomic Resource for Cancer Biology and Systems Pharmacology. Cancer research 73 (14):4372-4382. doi:10.1158/0008-5472.Can-12-3342

26. Reinhold WC, Varma S, Sousa F, Sunshine M, Abaan OD, Davis SR, Reinhold SW, Kohn KW, Morris J, Meltzer PS, Doroshow JH, Pommier Y (2014) NCI-60 Whole Exome Sequencing and Pharmacological CellMiner Analyses. PloS one 9 (7). doi:ARTN e101670

10.1371/journal.pone.0101670

27. Scherf U, Ross DT, Waltham M, Smith LH, Lee JK, Tanabe L, Kohn KW, Reinhold WC, Myers TG, Andrews DT, Scudiero DA, Eisen MB, Sausville EA, Pommier Y, Botstein D, Brown PO, Weinstein JN (2000) A gene expression database for the molecular pharmacology of cancer. Nature Genetics 24 (3):236-244. doi:Doi 10.1038/73439

28. Gholami AM, Hahne H, Wu ZX, Auer FJ, Meng C, Wilhelm M, Kuster B (2013) Global Proteome Analysis of the NCI-60 Cell Line Panel. Cell Reports 4 (3):609-620. doi:10.1016/j.celrep.2013.07.018

29. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, Wilson CJ, Lehar J, Kryukov GV, Sonkin D, Reddy A, Liu M, Murray L, Berger MF, Monahan JE, Morais P, Meltzer J, Korejwa A, Jane-Valbuena J, Mapa FA, Thibault J, Bric-Furlong E, Raman P, Shipway A, Engels IH, Cheng J, Yu GK, Yu JJ, Aspesi P, de Silva M, Jagtap K, Jones MD, Wang L, Hatton C, Palescandolo E, Gupta S, Mahan S, Sougnez C, Onofrio RC, Liefeld T, MacConaill L, Winckler W, Reich M, Li NX, Mesirov JP, Gabriel SB, Getz G, Ardlie K, Chan V, Myer VE, Weber BL, Porter J, Warmuth M, Finan P, Harris JL, Meyerson M, Golub TR, Morrissey MP, Sellers WR, Schlegel R, Garraway LA (2012) The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity (vol 483, pg 603, 2012). Nature 492 (7428):290-290. doi:10.1038/nature11735

30. Stransky N, Ghandi M, Kryukov GV, Garraway LA, Lehar J, Liu M, Sonkin D, Kauffmann A, Venkatesan K, Edelman EJ, Riester M, Barretina J, Caponigro G, Schlegel R, Sellers WR, Stegmeier F, Morrissey M, Amzallag A, Pruteanu-Malinici I, Haber DA, Ramaswamy S, Benes CH, Menden MP, Iorio F, Stratton MR, McDermott U, Garnett MJ, Saez-Rodriguez J, Canc DS, Line CC, Inst B, Res NIB, Sensitivity GD, Hosp MG, Lab EMB, Inst EB, Inst WTS (2015) Pharmacogenomic agreement between two cancer cell line data sets. Nature 528 (7580):84-+. doi:10.1038/nature15736

31. Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, Lerner J, Brunet JP, Subramanian A, Ross KN, Reich M, Hieronymus H, Wei G, Armstrong SA, Haggarty SJ, Clemons PA, Wei R, Carr SA, Lander ES, Golub TR (2006) The connectivity map: Using gene-expression signatures to connect small molecules, genes, and disease. Science 313 (5795):1929-1935. doi:10.1126/science.1132939

32. Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, Armour CD, Bennett HA, Coffey E, Dai HY, He YDD, Kidd MJ, King AM, Meyer MR, Slade D, Lum PY, Stepaniants SB, Shoemaker DD, Gachotte D, Chakraburtty K, Simon J, Bard M, Friend SH (2000) Functional discovery via a compendium of expression profiles. Cell 102 (1):109-126. doi:Doi 10.1016/S0092-8674(00)00015-5

33. Gayvert KM, Dardenne E, Cheung C, Boland MR, Lorberbaum T, Wanjala J, Chen Y, Rubin MA, Tatonetti NP, Rickman DS, Elemento O (2016) A Computational Drug Repositioning Approach for Targeting Oncogenic Transcription Factors. Cell Rep 15 (11):2348-2356. doi:10.1016/j.celrep.2016.05.037

34. Dudley JT, Deshpande T, Butte AJ (2011) Exploiting drug-disease relationships for computational drug repositioning. Brief Bioinform 12 (4):303-311. doi:10.1093/bib/bbr013

35. Low SK, Takahashi A, Mushiroda T, Kubo M (2014) Genome-Wide Association Study: A Useful Tool to Identify Common Genetic Variants Associated with Drug Toxicity and Efficacy in Cancer Pharmacogenomics. Clinical Cancer Research 20 (10):2541-2552. doi:10.1158/1078-0432.Ccr-13-2755

36. Zhou KX, Pearson ER (2013) Insights from Genome-Wide Association Studies of Drug Response. Annu Rev Pharmacol 53:299-310. doi:10.1146/annurev-pharmtox-011112-140237

37. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JPA, Hirschhorn JN (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nature Reviews Genetics 9 (5):356-369. doi:10.1038/nrg2344

38. Xu ZL, Taylor JA (2009) SNPinfo: integrating GWAS and candidate gene information into functional SNP selection for genetic association studies. Nucleic acids research 37:W600-W605. doi:10.1093/nar/gkp290

39. McKinney BA, Pajewski NM (2011) Six Degrees of Epistasis: Statistical Network Models for GWAS. Frontiers in genetics 2:109. doi:10.3389/fgene.2011.00109

40. Howie B, Marchini J, Stephens M (2011) Genotype imputation with thousands of genomes. G3 (Bethesda) 1 (6):457-470. doi:10.1534/g3.111.001198

41. Byun E, Caillier SJ, Montalban X, Villoslada P, Fernandez O, Brassat D, Comabella M, Wang J, Barcellos LF, Baranzini SE, Oksenberg JR (2008) Genome-wide pharmacogenomic analysis of the response to interferon beta therapy in multiple sclerosis. Arch Neurol-Chicago 65 (3):337-E332. doi:DOI 10.1001/archneurol.2008.47

42. Liu CY, Batliwalla F, Li WT, Lee A, Roubenoff R, Beckman E, Khalili H, Damle A, Kern M, Furie R, Dupuis J, Plenge RM, Coenen MJH, Behrens TW, Carulli JP, Gregersen PK (2008) Genome-wide association scan identifies candidate polymorphisms associated with differential response to anti-TNF treatment in rheumatoid arthritis. Mol Med 14 (9-10):575-581. doi:10.2119/2008-00056.Liu

43. Link E, Parish S, Armitage J, Bowman L, Heath S, Matsuda F, Gut I, Lathrop M, Collins R, Grp SC (2008) SLCO1B1 variants and statin-induced myopathy - A genomewide study. New Engl J Med 359 (8):789-799

44. Bush WS, Moore JH (2012) Chapter 11: Genome-Wide Association Studies. PLoS computational biology 8 (12). doi:ARTN e1002822

10.1371/journal.pcbi.1002822

45. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC (2007) PLINK: A tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81 (3):559-575. doi:10.1086/519795

46. Burton PR, Clayton DG, Cardon LR, Craddock N, Deloukas P, Duncanson A, Kwiatkowski DP, McCarthy MI, Ouwehand WH, Samani NJ, Todd JA, Donnelly P, Barrett JC, Davison D, Easton D, Evans D, Leung HT, Marchini JL, Morris AP, Spencer CCA, Tobin MD, Attwood AP, Boorman JP, Cant B, Everson U, Hussey JM, Jolley JD, Knight AS, Koch K, Meech E, Nutland S, Prowse CV, Stevens HE, Taylor NC, Walters GR, Walker NM, Watkins NA, Winzer T, Jones RW, McArdle WL, Ring SM, Strachan DP, Pembrey M, Breen G, St Clair D, Caesar S, Gordon-Smith K, Jones L, Fraser C, Green EK, Grozeva D, Hamshere ML, Holmans PA, Jones IR, Kirov G, Moskvina V, Nikolov I, O'Donovan MC, Owen MJ, Collier DA, Elkin A, Farmer A, Williamson R, McGuffin P, Young AH, Ferrier IN, Ball SG, Balmforth AJ, Barrett JH, Bishop DT, Iles MM, Maqbool A, Yuldasheva N, Hall AS, Braund PS, Dixon RJ, Mangino M, Stevens S, Thompson JR, Bredin F, Tremelling M, Parkes M, Drummond H, Lees CW, Nimmo ER, Satsangi J, Fisher SA, Forbes A, Lewis CM, Onnie CM, Prescott NJ, Sanderson J, Mathew CG, Barbour J, Mohiuddin MK, Todhunter CE, Mansfield JC, Ahmad T, Cummings FR, Jewell DP, Webster J, Brown MJ, Lathrop GM, Connell J, Dominiczak A, Marcano CAB, Burke B, Dobson R, Gungadoo J, Lee KL, Munroe PB, Newhouse SJ, Onipinla A, Wallace C, Xue MZ, Caulfield M, Farrall M, Barton A, Bruce IN, Donovan H, Eyre S, Gilbert PD, Hider SL, Hinks AM, John SL, Potter C, Silman AJ, Symmons DPM, Thomson W, Worthington J, Dunger DB, Widmer B, Frayling TM, Freathy RM, Lango H, Perry JRB, Shields BM, Weedon MN, Hattersley AT, Hitman GA, Walker

M, Elliott KS, Groves CJ, Lindgren CM, Rayner NW, Timpson NJ, Zeggini E, Newport M, Sirugo G, Lyons E, Vannberg F, Brown MA, Franklyn JA, Heward JM, Simmonds MJ, Hill AVS, Bradbury LA, Farrar C, Pointon JJ, Wordsmith P, Gough SCL, Seal S, Stratton MR, Rahman N, Ban M, Goris A, Sawcer SJ, Compston A, Conway D, Jallow M, Bumpstead SJ, Chaney A, Downes K, Ghori MJR, Gwilliam R, Inouye M, Keniry A, King E, McGinnis R, Potter S, Ravindrarajah R, Whittaker P, Withers D, Easton D, Pereira-Gale J, Hallgrimsdottir IB, Howie BN, Su Z, Teo YY, Vukcevic D, Bentley D, Caulfield M, Mathew CG, Worthington J, Consortium WTCC, Syndicate BRGGS, Collaborat BCS (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447 (7145):661-678. doi:10.1038/nature05911

47. Gamazon ER, Huang RS, Cox NJ, Dolan ME (2010) Chemotherapeutic drug susceptibility associated SNPs are enriched in expression quantitative trait loci. Proceedings of the National Academy of Sciences of the United States of America 107 (20):9287-9292. doi:10.1073/pnas.1001827107

48. Chung CM, Wang RY, Chen JW, Fann CSJ, Leu HB, Ho HY, Ting CT, Lin TH, Sheu SH, Tsai WC, Chen JH, Jong YS, Lin SJ, Chen YT, Pan WH (2010) A genome-wide association study identifies new loci for ACE activity: potential implications for response to ACE inhibitor. Pharmacogenomics Journal 10 (6):537-544. doi:10.1038/tpj.2009.70

49. Diboun I, Wernisch L, Orengo CA, Koltzenburg M (2006) Microarray analysis after RNA amplification can detect pronounced differences in gene expression using limma. Bmc Genomics 7. doi:Artn 252

10.1186/1471-2164-7-252

50. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic acids research 43 (7):e47. doi:10.1093/nar/gkv007

51. Anders S, Huber W (2010) Differential expression analysis for sequence count data. Genome biology 11 (10). doi:ARTN R106

10.1186/gb-2010-11-10-r106

52. Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26 (1):139-140. doi:10.1093/bioinformatics/btp616

53. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nature Protocols 7 (3):562-578. doi:10.1038/nprot.2012.016

54. Wright G, Tan B, Rosenwald A, Hurt EH, Wiestner A, Staudt LM (2003) A gene expression-based method to diagnose clinically distinct subgroups of diffuse large B cell lymphoma. Proceedings of the National Academy of Sciences

of the United States of America 100 (17):9991-9996. doi:10.1073/pnas.1732008100

55. Lam LT, Davis RE, Pierce J, Hepperle M, Xu Y, Hottelet M, Nong Y, Wen D, Adams J, Dang L, Staudt LM (2005) Small molecule inhibitors of IkappaB kinase are selectively toxic for subgroups of diffuse large B-cell lymphoma defined by gene expression profiling. Clin Cancer Res 11 (1):28-40

56. Briones J (2009) Emerging therapies for B-cell non-Hodgkin lymphoma. Expert Rev Anticanc 9 (9):1305-1316. doi:10.1586/Era.09.86

57. Jimeno A, Tan AC, Coffa J, Rajeshkumar NV, Kulesza P, Rubio-Viqueira B, Wheelhouse J, Diosdado B, Messersmith WA, Lacobuzio-Donahue C, Maitra A, Varella-Garcia M, Hirsch FR, Meijer GA, Hidalgo M (2008) Coordinated epidermal growth factor receptor pathway gene overexpression predicts epidermal growth factor receptor inhibitor sensitivity in pancreatic cancer. Cancer research 68 (8):2841-2849. doi:10.1158/0008-5472.Can-07-5200

58. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdottir H, Tamayo P, Mesirov JP (2011) Molecular signatures database (MSigDB) 3.0. Bioinformatics 27 (12):1739-1740. doi:10.1093/bioinformatics/btr260

59. Fabregat A, Sidiropoulos K, Garapati P, Gillespie M, Hausmann K, Haw R, Jassal B, Jupe S, Korninger F, McKay S, Matthews L, May B, Milacic M, Rothfels K, Shamovsky V, Webber M, Weiser J, Williams M, Wu G, Stein L, Hermjakob H, D'Eustachio P (2016) The Reactome pathway Knowledgebase. Nucleic acids research 44 (D1):D481-487. doi:10.1093/nar/gkv1351

60. Croft D, Mundo AF, Haw R, Milacic M, Weiser J, Wu G, Caudy M, Garapati P, Gillespie M, Kamdar MR, Jassal B, Jupe S, Matthews L, May B, Palatnik S, Rothfels K, Shamovsky V, Song H, Williams M, Birney E, Hermjakob H, Stein L, D'Eustachio P (2014) The Reactome pathway knowledgebase. Nucleic acids research 42 (Database issue):D472-477. doi:10.1093/nar/gkt1102

61. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M (1999) KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic acids research 27 (1):29-34. doi:DOI 10.1093/nar/27.1.29

62. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25 (1):25-29. doi:10.1038/75556

63. Breuer K, Foroushani AK, Laird MR, Chen C, Sribnaia A, Lo R, Winsor GL, Hancock RE, Brinkman FS, Lynn DJ (2013) InnateDB: systems biology of innate immunity and beyond--recent updates and continuing curation. Nucleic acids research 41 (Database issue):D1228-1233. doi:10.1093/nar/gks1147

64. Lynn DJ, Winsor GL, Chan C, Richard N, Laird MR, Barsky A, Gardy JL, Roche FM, Chan TH, Shah N, Lo R, Naseer M, Que J, Yau M, Acab M, Tulpan D, Whiteside MD, Chikatamarla A, Mah B, Munzner T, Hokamp K, Hancock RE, Brinkman FS (2008) InnateDB: facilitating systems-level analyses of the mammalian innate immune response. Molecular systems biology 4:218. doi:10.1038/msb.2008.55

65. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP (2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences of the United States of America 102 (43):15545-15550. doi:10.1073/pnas.0506580102

66. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences of the United States of America 102 (43):15545-15550. doi:10.1073/pnas.0506580102

67. Rarey M, Kramer B, Lengauer T, Klebe G (1996) A fast flexible docking method using an incremental construction algorithm. Journal of molecular biology 261 (3):470-489. doi:DOI 10.1006/jmbi.1996.0477

68. Butina D, Segall MD, Frankcombe K (2002) Predicting ADME properties in silico: methods and models. Drug Discov Today 7 (11):S83-S88. doi:Pii S1359-6446(02)02288-2

Doi 10.1016/S1359-6446(02)02288-2

69. Nantasenamat C, Isarankura-Na-Ayudhya C, Prachayasittikul V (2010) Advances in computational methods to predict the biological activity of compounds. Expert Opin Drug Dis 5 (7):633-654. doi:10.1517/17460441.2010.492827

70. Wang KJ, Sun JZ, Zhou SF, Wan CL, Qin SY, Li C, He L, Yang L (2013) Prediction of Drug-Target Interactions for Drug Repositioning Only Based on Genomic Expression Similarity. PLoS computational biology 9 (11). doi:ARTN e1003315

10.1371/journal.pcbi.1003315

71. Campillos M, Kuhn M, Gavin AC, Jensen LJ, Bork P (2008) Drug target identification using side-effect similarity. Science 321 (5886):263-266. doi:10.1126/science.1158140

72. Madhukar NS, Huang L, Khade P, Gayvert K, Giannakakou P, Elemento O (2015) Abstract B162: Small molecule target prediction and identification of novel anti-cancer compounds using a data-driven bayesian approach.

Molecular cancer therapeutics 14 (12 Supplement 2):B162. doi:10.1158/1535-7163.targ-15-b162

73. Li J, Wood WH, Becker KG, Weeraratna AT, Morin PJ (2007) Gene expression response to cisplatin treatment in drug-sensitive and drug-resistant ovarian cancer cells. Oncogene 26 (20):2860-2872. doi:10.1038/sj.onc.1210086

74. Lamb J (2007) The Connectivity Map: a new tool for biomedical research. Nature reviews Cancer 7 (1):54-60. doi:10.1038/nrc2044

75. Paez JG, Janne PA, Lee JC, Tracy S, Greulich H, Gabriel S, Herman P, Kaye FJ, Lindeman N, Boggon TJ, Naoki K, Sasaki H, Fujii Y, Eck MJ, Sellers WR, Johnson BE, Meyerson M (2004) EGFR mutations in lung cancer: Correlation with clinical response to gefitinib therapy. Science 304 (5676):1497-1500. doi:10.1126/science.1099314

76. Cappuzzo F, Bemis L, Varella-Garcia M (2006) HER2 mutation and response to trastuzumab therapy in non-small-cell lung cancer. N Engl J Med 354 (24):2619-2621. doi:10.1056/NEJMc060020

77. Romond EH, Perez EA, Bryant J, Suman VJ, Geyer CE, Jr., Davidson NE, Tan-Chiu E, Martino S, Paik S, Kaufman PA, Swain SM, Pisansky TM, Fehrenbacher L, Kutteh LA, Vogel VG, Visscher DW, Yothers G, Jenkins RB, Brown AM, Dakhil SR, Mamounas EP, Lingle WL, Klein PM, Ingle JN, Wolmark N (2005) Trastuzumab plus adjuvant chemotherapy for operable HER2-positive breast cancer. N Engl J Med 353 (16):1673-1684. doi:10.1056/NEJMoa052122

78. Young K, Minchom A, Larkin J (2012) BRIM-1, -2 and -3 trials: improved survival with vemurafenib in metastatic melanoma patients with a BRAF(V600E) mutation. Future Oncol 8 (5):499-507. doi:10.2217/fon.12.43

79. Chapman PB, Hauschild A, Robert C, Haanen JB, Ascierto P, Larkin J, Dummer R, Garbe C, Testori A, Maio M, Hogg D, Lorigan P, Lebbe C, Jouary T, Schadendorf D, Ribas A, O'Day SJ, Sosman JA, Kirkwood JM, Eggermont AM, Dreno B, Nolop K, Li J, Nelson B, Hou J, Lee RJ, Flaherty KT, McArthur GA, Group B-S (2011) Improved survival with vemurafenib in melanoma with BRAF V600E mutation. N Engl J Med 364 (26):2507-2516. doi:10.1056/NEJMoa1103782

80. Bollag G, Tsai J, Zhang J, Zhang C, Ibrahim P, Nolop K, Hirth P (2012) Vemurafenib: the first drug approved for BRAF-mutant cancer. Nat Rev Drug Discov 11 (11):873-886. doi:10.1038/nrd3847

81. Vogel CL, Cobleigh MA, Tripathy D, Gutheil JC, Harris LN, Fehrenbacher L, Slamon DJ, Murphy M, Novotny WF, Burchmore M, Shak S, Stewart SJ, Press M (2002) Efficacy and safety of trastuzumab as a single agent in first-line treatment of HER2-overexpressing metastatic breast cancer. J Clin Oncol 20 (3):719-726. doi:10.1200/JCO.2002.20.3.719

82. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, Hasz R, Walters G, Garcia F, Young N, Foster B, Moser M, Karasik E, Gillard B, Ramsey K, Sullivan S, Bridge J, Magazine H, Syron J, Fleming J, Siminoff L, Traino H, Mosavel M, Barker L, Jewell S, Rohrer D, Maxim D, Filkins D, Harbach P, Cortadillo E, Berghuis B, Turner L, Hudson E, Feenstra K, Sobin L, Robb J, Branton P, Korzeniewski G, Shive C, Tabor D, Qi LQ, Groch K, Nampally S, Buia S, Zimmerman A, Smith A, Burges R, Robinson K, Valentino K, Bradbury D, Cosentino M, Diaz-Mayoral N, Kennedy M, Engel T, Williams P, Erickson K, Ardlie K, Winckler W, Getz G, DeLuca D, MacArthur D, Kellis M, Thomson A, Young T, Gelfand E, Donovan M, Meng Y, Grant G, Mash D, Marcus Y, Basile M, Liu J, Zhu J, Tu ZD, Cox NJ, Nicolae DL, Gamazon ER, Im HK, Konkashbaev A, Pritchard J, Stevens M, Flutre T, Wen XQ, Dermitzakis ET, Lappalainen T, Guigo R, Monlong J, Sammeth M, Koller D, Battle A, Mostafavi S, McCarthy M, Rivas M, Maller J, Rusyn I, Nobel A, Wright F, Shabalin A, Feolo M, Sharopova N, Sturcke A, Paschal J, Anderson JM, Wilder EL, Derr LK, Green ED, Struewing JP, Temple G, Volpi S, Boyer JT, Thomson EJ, Guyer MS, Ng C, Abdallah A, Colantuoni D, Insel TR, Koester SE, Little AR, Bender PK, Lehner T, Yao Y, Compton CC, Vaught JB, Sawyer S, Lockhart NC, Demchok J, Moore HF (2013) The Genotype-Tissue Expression (GTEx) project. Nature Genetics 45 (6):580-585. doi:10.1038/ng.2653

83. Sachlos E, Risueno RM, Laronde S, Shapovalova Z, Lee JH, Russell J, Malig M, McNicol JD, Fiebig-Comyn A, Graham M, Levadoux-Martin M, Lee JB, Giacomelli AO, Hassell JA, Fischer-Russell D, Trus MR, Foley R, Leber B, Xenocostas A, Brown ED, Collins TJ, Bhatia M (2012) Identification of Drugs Including a Dopamine Receptor Antagonist that Selectively Target Cancer Stem Cells. Cell 149 (6):1284-1297. doi:10.1016/j.cell.2012.03.049

84. Madhukar NS, Elemento O, Pandey G (2015) Prediction of Genetic Interactions Using Machine Learning and Network Properties. Frontiers in Bioengineering and Biotechnology 3 (172). doi:10.3389/fbioe.2015.00172

85. Chan DA, Giaccia AJ (2011) Harnessing synthetic lethal interactions in anticancer drug discovery. Nat Rev Drug Discov 10 (5):351-364. doi:10.1038/nrd3374

86. Jerby-Arnon L, Pfetzer N, Waldman YY, McGarry L, James D, Shanks E, Seashore-Ludlow B, Weinstock A, Geiger T, Clemons PA, Gottlieb E, Ruppin E (2014) Predicting cancer-specific vulnerability via data-driven detection of synthetic lethality. Cell 158 (5):1199-1209. doi:10.1016/j.cell.2014.07.027

87. Garnett MJ, Edelman EJ, Heidorn SJ, Greenman CD, Dastur A, Lau KW, Greninger P, Thompson IR, Luo X, Soares J, Liu Q, Iorio F, Surdez D, Chen L, Milano RJ, Bignell GR, Tam AT, Davies H, Stevenson JA, Barthorpe S, Lutz SR, Kogera F, Lawrence K, McLaren-Douglas A, Mitropoulos X, Mironenko T, Thi H, Richardson L, Zhou W, Jewitt F, Zhang T, O'Brien P, Boisvert JL, Price S, Hur W, Yang W, Deng X, Butler A, Choi HG, Chang JW, Baselga J, Stamenkovic I,

Engelman JA, Sharma SV, Delattre O, Saez-Rodriguez J, Gray NS, Settleman J, Futreal PA, Haber DA, Stratton MR, Ramaswamy S, McDermott U, Benes CH (2012) Systematic identification of genomic markers of drug sensitivity in cancer cells. Nature 483 (7391):570-575. doi:10.1038/nature11005

88. Menden MP, Iorio F, Garnett M, McDermott U, Benes CH, Ballester PJ, Saez-Rodriguez J (2013) Machine Learning Prediction of Cancer Cell Sensitivity to Drugs Based on Genomic and Chemical Properties. PloS one 8 (4). doi:ARTN e61318

10.1371/journal.pone.0061318

89. Costello JC, Heiser LM, Georgii E, Gonen M, Menden MP, Wang NJ, Bansal M, Ammad-ud-din M, Hintsanen P, Khan SA, Mpindi JP, Kallioniemi O, Honkela A, Aittokallio T, Wennerberg K, Collins JJ, Gallahan D, Singer D, Saez-Rodriguez J, Kaski S, Gray JW, Stolovitzky G, Community ND (2014) A community effort to assess and improve drug sensitivity prediction algorithms. Nature Biotechnology 32 (12):1202-U1257. doi:10.1038/nbt.2877

90. Cuatrecasas P (2006) Drug discovery in jeopardy. The Journal of clinical investigation 116 (11):2837-2842. doi:10.1172/JCI29999

91. Chan JN, Nislow C, Emili A (2010) Recent advances and method development for drug target identification. Trends in pharmacological sciences 31 (2):82-88. doi:10.1016/j.tips.2009.11.002

92. Weigelt J (2009) The case for open-access chemical biology. A strategy for pre-competitive medicinal chemistry to promote drug discovery. EMBO reports 10 (9):941-945. doi:10.1038/embor.2009.193

93. Williams M (2003) Target validation. Current opinion in pharmacology 3 (5):571-577

94. Dearden JC (2003) In silico prediction of drug toxicity. Journal of computer-aided molecular design 17 (2-4):119-127

95. Nantasenamat C, Isarankura-Na-Ayudhya C, Prachayasittikul V (2010) Advances in computational methods to predict the biological activity of compounds. Expert Opin Drug Discov 5 (7):633-654. doi:10.1517/17460441.2010.492827

96. Li H, Gao Z, Kang L, Zhang H, Yang K, Yu K, Luo X, Zhu W, Chen K, Shen J, Wang X, Jiang H (2006) TarFisDock: a web server for identifying drug targets with docking approach. Nucleic acids research 34 (Web Server issue):W219-224. doi:10.1093/nar/gkl114

97. Wang K, Sun J, Zhou S, Wan C, Qin S, Li C, He L, Yang L (2013) Prediction of drug-target interactions for drug repositioning only based on genomic expression similarity. PLoS computational biology 9 (11):e1003315. doi:10.1371/journal.pcbi.1003315

98. Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, Lerner J, Brunet JP, Subramanian A, Ross KN, Reich M, Hieronymus H, Wei G, Armstrong SA, Haggarty SJ, Clemons PA, Wei R, Carr SA, Lander ES, Golub TR (2006) The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. Science 313 (5795):1929-1935. doi:10.1126/science.1132939

99. Campillos M, Kuhn M, Gavin AC, Jensen LJ, Bork P (2008) Drug target identification using side-effect similarity. Science 321 (5886):263-266. doi:10.1126/science.1158140

100. Grundmark B, Holmberg L, Garmo H, Zethelius B (2014) Reducing the noise in signal detection of adverse drug reactions by standardizing the background: a pilot study on analyses of proportional reporting ratios-by-therapeutic area. Eur J Clin Pharmacol 70 (5):627-635. doi:10.1007/s00228-014-1658-1

101. Shang N, Xu H, Rindflesch TC, Cohen T (2014) Identifying plausible adverse drug reactions using knowledge extracted from the literature. J Biomed Inform 52:293-310. doi:10.1016/j.jbi.2014.07.011

102. Fortney K, Griesman J, Kotlyar M, Pastrello C, Angeli M, Sound-Tsao M, Jurisica I (2015) Prioritizing therapeutics for lung cancer: an integrative meta-analysis of cancer gene signatures and chemogenomic data. PLoS computational biology 11 (3):e1004068. doi:10.1371/journal.pcbi.1004068

103. Ma'ayan A, Rouillard AD, Clark NR, Wang Z, Duan Q, Kou Y (2014) Lean Big Data integration in systems biology and systems pharmacology. Trends in pharmacological sciences 35 (9):450-460. doi:10.1016/j.tips.2014.07.001

104. Wang Z, Clark NR, Ma'ayan A (2016) Drug-induced adverse events prediction with the LINCS L1000 data. Bioinformatics 32 (15):2338-2345. doi:10.1093/bioinformatics/btw168

105. Perlman L, Gottlieb A, Atias N, Ruppin E, Sharan R (2011) Combining drug and gene similarity measures for drug-target elucidation. Journal of computational biology : a journal of computational molecular cell biology 18 (2):133-145. doi:10.1089/cmb.2010.0213

106. Fakhraei S, Huang B, Raschid L, Getoor L (2014) Network-Based Drug-Target Interaction Prediction with Probabilistic Soft Logic. IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM 11 (5):775-787. doi:10.1109/TCBB.2014.2325031

107. Chen X, Yan CC, Zhang X, Zhang X, Dai F, Yin J, Zhang Y (2016) Drug-target interaction prediction: databases, web servers and computational models. Brief Bioinform 17 (4):696-712. doi:10.1093/bib/bbv066

108. Shoemaker RH (2006) The NCI60 human tumour cell line anticancer drug screen. Nat Rev Cancer 6 (10):813-823. doi:10.1038/nrc1951

109. Li Q, Cheng T, Wang Y, Bryant SH (2010) PubChem as a public resource for drug discovery. Drug Discov Today 15 (23-24):1052-1057. doi:10.1016/j.drudis.2010.10.003

110. Chen B, Wild DJ (2010) PubChem BioAssays as a data source for predictive models. Journal of molecular graphics & modelling 28 (5):420-426. doi:10.1016/j.jmgm.2009.10.001

111. Kuhn M, Campillos M, Letunic I, Jensen LJ, Bork P (2010) A side effect resource to capture phenotypic effects of drugs. Molecular systems biology 6:343. doi:10.1038/msb.2009.98

112. Law V, Knox C, Djoumbou Y, Jewison T, Guo AC, Liu Y, Maciejewski A, Arndt D, Wilson M, Neveu V, Tang A, Gabriel G, Ly C, Adamjee S, Dame ZT, Han B, Zhou Y, Wishart DS (2014) DrugBank 4.0: shedding new light on drug metabolism. Nucleic Acids Res 42 (Database issue):D1091-1097. doi:10.1093/nar/gkt1068

113. Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S, Tzur D, Gautam B, Hassanali M (2008) DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic acids research 36 (Database issue):D901-906. doi:10.1093/nar/gkm958

114. Yamanishi Y, Kotera M, Kanehisa M, Goto S (2010) Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework. Bioinformatics 26 (12):i246-254. doi:10.1093/bioinformatics/btq176

115. Hizukuri Y, Sawada R, Yamanishi Y (2015) Predicting target proteins for drug candidate compounds based on drug-induced gene expression data in a chemical structure-independent manner. BMC Med Genomics 8:82. doi:10.1186/s12920-015-0158-1

116. Anastassiadis T, Deacon SW, Devarajan K, Ma H, Peterson JR (2011) Comprehensive assay of kinase catalytic activity reveals features of kinase inhibitor selectivity. Nat Biotechnol 29 (11):1039-1045. doi:10.1038/nbt.2017

117. Jordan MA, Wilson L (2004) Microtubules as a target for anticancer drugs. Nature reviews Cancer 4 (4):253-265

118. Jordan MA, Wilson L (1998) Microtubules and actin filaments: dynamic targets for cancer chemotherapy. Curr Opin Cell Biol 10 (1):123-130

119. Giannakakou P, Sackett D, Fojo T (2000) Tubulin/microtubules: still a promising target for new chemotherapeutic agents. J Natl Cancer Inst 92 (3):182-183

120. Jordan A, Hadfield JA, Lawrence NJ, McGown AT (1998) Tubulin as a target for anticancer drugs: Agents which interact with the mitotic spindle. Medicinal Research Reviews 18 (4):259-296. doi:10.1002/(SICI)1098-1128(199807)18:4<259::AID-MED3>3.0.CO;2-U

121. Mukhtar E, Adhami VM, Mukhtar H (2014) Targeting microtubules by natural agents for cancer therapy. Molecular cancer therapeutics 13 (2):275-284. doi:10.1158/1535-7163.MCT-13-0791

122. Giannakakou P, Sackett DL, Kang YK, Zhan Z, Buters JT, Fojo T, Poruchynsky MS (1997) Paclitaxel-resistant human ovarian cancer cells have mutant beta-tubulins that exhibit impaired paclitaxel-driven polymerization. The Journal of biological chemistry 272 (27):17118-17125

123. Nicolaou KC, Winssinger N, Pastor J, Ninkovic S, Sarabia F, He Y, Vourloumis D, Yang Z, Li T, Giannakakou P, Hamel E (1997) Synthesis of epothilones A and B in solid and solution phase. Nature 387 (6630):268-272. doi:10.1038/387268a0

124. Giannakakou P, Gussio R, Nogales E, Downing KH, Zaharevitz D, Bollbuck B, Poy G, Sackett D, Nicolaou KC, Fojo T (2000) A common pharmacophore for epothilone and taxanes: molecular basis for drug resistance conferred by tubulin mutations in human cancer cells. Proceedings of the National Academy of Sciences of the United States of America 97 (6):2904-2909. doi:10.1073/pnas.040546297

125. Nicolaou KC, Namoto K, Ritzen A, Ulven T, Shoji M, Li J, D'Amico G, Liotta D, French CT, Wartmann M, Altmann KH, Giannakakou P (2001) Chemical synthesis and biological evaluation of cis- and trans-12,13-cyclopropyl and 12,13-cyclobutyl epothilones and related pyridine side chain analogues. J Am Chem Soc 123 (38):9313-9323

126. Nicolaou KC, Sasmal PK, Rassias G, Reddy MV, Altmann KH, Wartmann M, O'Brate A, Giannakakou P (2003) Design, synthesis, and biological properties of highly potent epothilone B analogues. Angew Chem Int Ed Engl 42 (30):3515-3520. doi:10.1002/anie.200351819

127. O'Rourke B, Yang CP, Sharp D, Horwitz SB (2014) Eribulin disrupts EB1-microtubule plus-tip complex formation. Cell Cycle 13 (20):3218-3221. doi:10.4161/15384101.2014.950143

128. Dybdal-Hargreaves NF, Risinger AL, Mooberry SL (2015) Eribulin mesylate: mechanism of action of a unique microtubule-targeting agent. Clin Cancer Res 21 (11):2445-2452. doi:10.1158/1078-0432.CCR-14-3252

129. Gamucci T, Michelotti A, Pizzuti L, Mentuccia L, Landucci E, Sperduti I, Di Lauro L, Fabi A, Tonini G, Sini V, Salesi N, Ferrarini I, Vaccaro A, Pavese I, Veltri E, Moscetti L, Marchetti P, Vici P (2014) Eribulin mesylate in pretreated breast cancer patients: a multicenter retrospective observational study. J Cancer 5 (5):320-327. doi:10.7150/jca.8748

130. Allen JE, Krigsfeld G, Mayes PA, Patel L, Dicker DT, Patel AS, Dolloff NG, Messaris E, Scata KA, Wang W, Zhou J-Y, Wu GS, El-Deiry WS (2013) Dual Inactivation of Akt and ERK by TIC10 Signals Foxo3a Nuclear Translocation,

TRAIL Gene Induction, and Potent Antitumor Effects. Science translational medicine 5 (171):171ra117-171ra117. doi:10.1126/scitranslmed.3004828

131. J. Ishizawa, K Kojima, D. Chachad, P. Ruvolo, V. Ruvolo, R. O. Jacamo, G. Borthakur, H. Mu, Z. Zeng, Y. Tabe, J. E. Allen, Z. Wang, W. Ma, H. C. Lee, R. Orlowski, D. D. Sarbassov, S. S. Neelapu, T. McDonnell, R.N. Miranda, M. Wang, H. Kantarjian, M. Konopleva, R. E. Davis, Andreeff M (2015 (in press)) ONC201 Induces p53-independent Apoptosis in Hematological Malignancies and Leukemic Stem/Progenitor Cells through ER Stress Response. Science Signaling

132. Kline CL, Vanden Heuvel P, Allen JE, Prabhu VV, Dicker DT, WS E-D (2015 (in press)) Anti-cancer agent ONC201 activates early ATF4/DR5 upregulation, and cell death associated with XIAP inhibition. Science Signaling

133. Bedard P, Parkes JD, Marsden CD (1978) Effect of new dopamine-blocking agent (oxiperomide) on drug-induced dyskinesias in Parkinson's disease and spontaneous dyskinesias. Br Med J 1 (6118):954-956

134. Casey DE, Gerlach J (1980) Oxiperomide in tardive dyskinesia. J Neurol Neurosurg Psychiatry 43 (3):264-267

135. Casey DE, Gerlach J (1979) Sulpiride and oxiperomide in tardive dyskinesia. Trans Am Neurol Assoc 104:210-211

136. Meltzer HY, Sachar EJ, Frantz AG (1975) Dopamine antagonism by thioridazine in schizophrenia. Biol Psychiatry 10 (1):53-57

137. Zhang R, Xie X (2012) Tools for GPCR drug discovery. Acta Pharmacol Sin 33 (3):372-384. doi:10.1038/aps.2011.173

138. Wagner J, Kline CL, Pottorf RS, Nallaganchu BR, Olson GL, Dicker DT, Allen JE, El-Deiry WS (2014) The angular structure of ONC201, a TRAIL pathway-inducing compound, determines its potent anti-cancer activity. Oncotarget 5 (24):12728-12737

139. Chang JY, Hsieh HP, Pan WY, Liou JP, Bey SJ, Chen LT, Liu JF, Song JS (2003) Dual inhibition of topoisomerase I and tubulin polymerization by BPR0Y007, a novel cytotoxic agent. Biochem Pharmacol 65 (12):2009-2019

140. Devillard L, Vandroux D, Tissier C, Bopassa JC, Ferrera R, Rochette L, Athias P Opioid-induced protection of cardiac myocytes from ischemic injury: Involvement of microtubules. Journal of Molecular and Cellular Cardiology 42 (6):S193-S194. doi:10.1016/j.yjmcc.2007.03.588

141. Borsodi A, Toth G (1986) Microtubule disassembly increases the number of opioid receptor binding sites in rat cerebrum membranes. Neuropeptides 8 (1):51-54

142. Crosby NJ, Deane KH, Clarke CE (2003) Beta-blocker therapy for tremor in Parkinson's disease. Cochrane Database Syst Rev (1):CD003361. doi:10.1002/14651858.CD003361

143. Carr A, Cooper DA (2000) Adverse effects of antiretroviral therapy. Lancet 356 (9239):1423-1430. doi:10.1016/S0140-6736(00)02854-3

144. Allen JE, Krigsfeld G, Mayes PA, Patel L, Dicker DT, Patel AS, Dolloff NG, Messaris E, Scata KA, Wang W, Zhou JY, Wu GS, El-Deiry WS (2013) Dual inactivation of Akt and ERK by TIC10 signals Foxo3a nuclear translocation, TRAIL gene induction, and potent antitumor effects. Sci Transl Med 5 (171):171ra117. doi:10.1126/scitranslmed.3004828

145. Allen JE, Krigsfeld G, Patel L, Mayes PA, Dicker DT, Wu GS, El-Deiry WS (2015) Identification of TRAIL-inducing compounds highlights small molecule ONC201/TIC10 as a unique anti-cancer agent that activates the TRAIL pathway. Molecular cancer 14:99. doi:10.1186/s12943-015-0346-9

146. Hughes JP, Rees S, Kalindjian SB, Philpott KL (2011) Principles of early drug discovery. Br J Pharmacol 162 (6):1239-1249. doi:10.1111/j.1476-5381.2010.01127.x

147. Radovich M, Kiel PJ, Nance SM, Niland EE, Parsley ME, Ferguson ME, Jiang G, Ammakkanavar NR, Einhorn LH, Cheng L, Nassiri M, Davidson DD, Rushing DA, Loehrer PJ, Pili R, Hanna N, Callaghan JT, Skaar TC, Helft PR, Shahda S, O'Neil BH, Schneider BP (2016) Clinical benefit of a precision medicine based approach for guiding treatment of refractory cancers. Oncotarget 7 (35):56491-56500. doi:10.18632/oncotarget.10606

148. Dolsten M, Sogaard M (2012) Precision medicine: an approach to R&D for delivering superior medicines to patients. Clin Transl Med 1 (1):7. doi:10.1186/2001-1326-1-7

149. Neve KA, Seamans JK, Trantham-Davidson H (2004) Dopamine receptor signaling. J Recept Signal Transduct Res 24 (3):165-205

150. Hasbi A, O'Dowd BF, George SR (2010) Heteromerization of dopamine D2 receptors with dopamine D1 or D5 receptors generates intracellular calcium signaling by different mechanisms. Current opinion in pharmacology 10 (1):93-99. doi:10.1016/j.coph.2009.09.011

151. Tarapore R, Garnett M, McDermott U, Benes C, Allen J (2016) Abstract 1236: ONC201 sensitivity profiling indicates pronounced sensitivity in lymphoid, prostate, colon and brain tumors. Cancer research 76:1236-1236

152. Yang W, Soares J, Greninger P, Edelman EJ, Lightfoot H, Forbes S, Bindal N, Beare D, Smith JA, Thompson IR, Ramaswamy S, Futreal PA, Haber DA, Stratton MR, Benes C, McDermott U, Garnett MJ (2013) Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. Nucleic acids research 41 (Database issue):D955-961. doi:10.1093/nar/gks1111

153. Liu R, Wang X, Chen GY, Dalerba P, Gurney A, Hoey T, Sherlock G, Lewicki J, Shedden K, Clarke MF (2007) The prognostic role of a gene signature from

tumorigenic breast-cancer cells. N Engl J Med 356 (3):217-226. doi:10.1056/NEJMoa063994

154. Zeppernick F, Ahmadi R, Campos B, Dictus C, Helmke BM, Becker N, Lichter P, Unterberg A, Radlwimmer B, Herold-Mende CC (2008) Stem cell marker CD133 affects clinical outcome in glioma patients. Clin Cancer Res 14 (1):123-129. doi:10.1158/1078-0432.CCR-07-0932

155. Chen J, Xia Q, Jiang B, Chang W, Yuan W, Ma Z, Liu Z, Shu X (2015) Prognostic Value of Cancer Stem Cell Marker ALDH1 Expression in Colorectal Cancer: A Systematic Review and Meta-Analysis. PLoS One 10 (12):e0145164. doi:10.1371/journal.pone.0145164

156. Smith DC, Eisenberg PD, Manikhas G, Chugh R, Gubens MA, Stagg RJ, Kapoun AM, Xu L, Dupont J, Sikic B (2014) A phase I dose escalation and expansion study of the anticancer stem cell agent demcizumab (anti-DLL4) in patients with previously treated solid tumors. Clin Cancer Res 20 (24):6295-6303. doi:10.1158/1078-0432.CCR-14-1373

157. Hitron M, Stephenson J, Chi KN, Edenfield WJ, Leggett D, Li Y, Li W, Gada K, Li C (2014 (suppl; abstr 2530)) A phase 1b study of the cancer stem cell inhibitor BBI608 administered with paclitaxel in patients with advanced malignancies. J Clin Oncol 32:5s

158. Rudin CM, Pietanza MC, Bauer TM, Spigel DR, Ready N, Morgensztern D, Glisson BS, Byers LA, Johnson ML, Burris HA, Robert F, Strickland DK, Zayed H, Govindan R, Dylla S, Peng SL (2016 (suppl; abstr LBA8505)) Safety and efficacy of single-agent rovalpituzumab tesirine (SC16LD6.5), a delta-like protein 3 (DLL3)-targeted antibody-drug conjugate (ADC) in recurrent or refractory small cell lung cancer (SCLC). J Clin Oncol 34

159. Allen JE, Kline CL, Prabhu VV, Wagner J, Ishizawa J, Madhukar N, Lev A, Baumeister M, Zhou L, Lulla A, Stogniew M, Schalop L, Benes C, Kaufman HL, Pottorf RS, Nallaganchu BR, Olson GL, Al-Mulla F, Duvic M, Wu GS, Dicker DT, Talekar MK, Lim B, Elemento O, Oster W, Bertino J, Flaherty K, Wang ML, Borthakur G, Andreeff M, Stein M, El-Deiry WS (2016) Discovery and clinical introduction of first-in-class imipridone ONC201. Oncotarget. doi:10.18632/oncotarget.11814

160. Stein MN, Bertino JR, Kaufman HL, Mayer T, Moss R, Silk A, Chan N, Malhotra J, Rodriguez-Rodriguez L, Aisner J, Aiken RD, Haffty BG, DiPaola RS, Saunders T, Zloza A, Damare S, Beckett Y, Yu B, Najmi S, Gabel C, Dickerson S, Zheng L, El-Deiry WS, Allen J, Stogniew M, Oster W, Mehnert JM (2017) First-in-human Clinical Trial of Oral ONC201 in Patients with Refractory Solid Tumors. Clin Cancer Res. doi:10.1158/1078-0432.CCR-16-2658

161. Prabhu VV, Allen JE, Dicker DT, El-Deiry WS (2015) Small-Molecule ONC201/TIC10 Targets Chemotherapy-Resistant Colorectal Cancer Stem-like

Cells in an Akt/Foxo3a/TRAIL-Dependent Manner. Cancer Res 75 (7):1423-1432. doi:10.1158/0008-5472.CAN-13-3451

162. Ishizawa J, Kojima K, Chachad D, Ruvolo P, Ruvolo V, Jacamo RO, Borthakur G, Mu H, Zeng Z, Tabe Y, Allen JE, Wang Z, Ma W, Lee HC, Orlowski R, Sarbassov dos D, Lorenzi PL, Huang X, Neelapu SS, McDonnell T, Miranda RN, Wang M, Kantarjian H, Konopleva M, Davis RE, Andreeff M (2016) ATF4 induction through an atypical integrated stress response to ONC201 triggers p53-independent apoptosis in hematological malignancies. Sci Signal 9 (415):ra17. doi:10.1126/scisignal.aac4380

163. Lasorella A, Benezra R, Iavarone A (2014) The ID proteins: master regulators of cancer stem cells and tumour aggressiveness. Nat Rev Cancer 14 (2):77-91. doi:10.1038/nrc3638

164. van den Hoogen C, van der Horst G, Cheung H, Buijs JT, Lippitt JM, Guzman-Ramirez N, Hamdy FC, Eaton CL, Thalmann GN, Cecchini MG, Pelger RC, van der Pluijm G (2010) High aldehyde dehydrogenase activity identifies tumor-initiating and metastasis-initiating cells in human prostate cancer. Cancer Res 70 (12):5163-5173. doi:10.1158/0008-5472.CAN-09-3806

165. Ying M, Tilghman J, Wei Y, Guerrero-Cazares H, Quinones-Hinojosa A, Ji H, Laterra J (2014) Kruppel-like factor-9 (KLF9) inhibits glioblastoma stemness through global transcription repression and integrin alpha6 inhibition. J Biol Chem 289 (47):32742-32756. doi:10.1074/jbc.M114.588988

166. Clements WK, Kim AD, Ong KG, Moore JC, Lawson ND, Traver D (2011) A somitic Wnt16/Notch pathway specifies haematopoietic stem cells. Nature 474 (7350):220-224. doi:10.1038/nature10107

167. Sun Y, Campisi J, Higano C, Beer TM, Porter P, Coleman I, True L, Nelson PS (2012) Treatment-induced damage to the tumor microenvironment promotes prostate cancer therapy resistance through WNT16B. Nat Med 18 (9):1359-1368. doi:10.1038/nm.2890

168. Gujral TS, Chan M, Peshkin L, Sorger PK, Kirschner MW, MacBeath G (2014) A noncanonical Frizzled2 pathway regulates epithelial-mesenchymal transition and metastasis. Cell 159 (4):844-856. doi:10.1016/j.cell.2014.10.032

169. Jin X, Jeon HY, Joo KM, Kim JK, Jin J, Kim SH, Kang BG, Beck S, Lee SJ, Kim JK, Park AK, Park WY, Choi YJ, Nam DH, Kim H (2011) Frizzled 4 regulates stemness and invasiveness of migrating glioma cells established by serial intracranial transplantation. Cancer Res 71 (8):3066-3075. doi:10.1158/0008-5472.CAN-10-1495

170. Zhang X, Chen Y, Ye Y, Wang J, Wang H, Yuan G, Lin Z, Wu Y, Zhang Y, Lin X (2016) Wnt signaling promotes hindgut fate commitment through regulating multi-lineage genes during hESC differentiation. Cell Signal 29:12-22. doi:10.1016/j.cellsig.2016.09.009

171. O'Brien CA, Kreso A, Ryan P, Hermans KG, Gibson L, Wang Y, Tsatsanis A, Gallinger S, Dick JE (2012) ID1 and ID3 regulate the self-renewal capacity of human colon cancer-initiating cells through p21. Cancer Cell 21 (6):777-792. doi:10.1016/j.ccr.2012.04.036

172. Jones RJ, Matsui WH, Smith BD (2004) Cancer stem cells: are we missing the target? J Natl Cancer Inst 96 (8):583-585

173. Lazarou J, Pomeranz BH, Corey PN (1998) Incidence of adverse drug reactions in hospitalized patients: a meta-analysis of prospective studies. JAMA 279 (15):1200-1205

174. White TJ, Arakelian A, Rho JP (1999) Counting the costs of drug-related adverse events. Pharmacoeconomics 15 (5):445-458

175. Sultana J, Cutroneo P, Trifiro G (2013) Clinical and economic burden of adverse drug reactions. J Pharmacol Pharmacother 4 (Suppl 1):S73-77. doi:10.4103/0976-500X.120957

176. Pierce CE, Bouri K, Pamer C, Proestel S, Rodriguez HW, Van Le H, Freifeld CC, Brownstein JS, Walderhaug M, Edwards IR, Dasgupta N (2017) Evaluation of Facebook and Twitter Monitoring to Detect Safety Signals for Medical Products: An Analysis of Recent FDA Safety Alerts. Drug Saf. doi:10.1007/s40264-016-0491-0

177. Heinrich J (2000) Adverse drug events: Substantial problem but magnitude uncertain. Committee on Health, Education, Labor, and Pensions. United States General Accounting Office (US GAO),

178. Lipinski CA (2004) Lead- and drug-like compounds: the rule-of-five revolution. Drug Discov Today Technol 1 (4):337-341. doi:10.1016/j.ddtec.2004.11.007

179. Bickerton GR, Paolini GV, Besnard J, Muresan S, Hopkins AL (2012) Quantifying the chemical beauty of drugs. Nature chemistry 4 (2):90-98. doi:10.1038/nchem.1243

180. Kamburov A, Wierling C, Lehrach H, Herwig R (2009) ConsensusPathDB--a database for integrating human functional interaction networks. Nucleic acids research 37 (Database issue):D623-628. doi:10.1093/nar/gkn698

181. Veber DF, Johnson SR, Cheng HY, Smith BR, Ward KW, Kopple KD (2002) Molecular properties that influence the oral bioavailability of drug candidates. Journal of medicinal chemistry 45 (12):2615-2623

182. Ghose AK, Viswanadhan VN, Wendoloski JJ (1999) A knowledge-based approach in designing combinatorial or medicinal chemistry libraries for drug discovery. 1. A qualitative and quantitative characterization of known drug databases. Journal of combinatorial chemistry 1 (1):55-68

183. Consortium GT (2015) Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science 348 (6235):648-660. doi:10.1126/science.1262110

184. http://exac.broadinstitute.org. Accessed February 2015

185. Vandyk AD, Harrison MB, Macartney G, Ross-White A, Stacey D (2012) Emergency department visits for symptoms experienced by oncology patients: a systematic review. Support Care Cancer 20 (8):1589-1599. doi:10.1007/s00520-012-1459-y

186. Carhart RE, Smith DH, Venkataraghavan R (1985) Atom pairs as molecular features in structure-activity studies: definition and applications. Journal of Chemical Information and Computer Sciences 25 (2):64-73. doi:10.1021/ci00046a002

187. Anastassiadis T, Deacon SW, Devarajan K, Ma H, Peterson JR (2011) Comprehensive assay of kinase catalytic activity reveals features of kinase inhibitor selectivity. Nat Biotech 29 (11):1039-1045. doi:http://www.nature.com/nbt/journal/v29/n11/abs/nbt.2017.html - supplementary-information

188. McGuinness D, Malikzay A, Visconti R, Lin K, Bayne M, Monsma F, Lunn CA (2009) Characterizing Cannabinoid CB 2 Receptor Ligands Using DiscoveRx PathHunter™ β-Arrestin Assay. Journal of Biomolecular Screening 14 (1):49-58. doi:10.1177/1087057108327329

189. Patel A, Murray J, McElwee-Whitmer S, Bai C, Kunapuli P, Johnson EN (2009) A combination of ultrahigh throughput PathHunter and cytokine secretion assays to identify glucocorticoid receptor agonists. Anal Biochem 385 (2):286-292. doi:10.1016/j.ab.2008.11.005

190. Corp RB (2017) Reaction Biology Corp Kinase Assay Protocol. http://www.reactionbiology.com/webapps/site/Kinase_Assay_Protocol.aspx. 2017

191. DiscoverX (2017) PathHunter Nuclear Translocation Assays. https://www.discoverx.com/technologies-platforms/enzyme-fragment-complementation-technology/cell-based-efc-assays/protein-translocation/nuclear-translocation-assays. 2017

192. Kanehisa M, Goto S (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic acids research 28 (1):27-30. doi:DOI 10.1093/nar/28.1.27

193. Kline CL, Van den Heuvel AP, Allen JE, Prabhu VV, Dicker DT, El-Deiry WS (2016) ONC201 kills solid tumor cells by triggering an integrated stress response dependent on ATF4 activation by specific eIF2alpha kinases. Sci Signal 9 (415):ra18. doi:10.1126/scisignal.aac4374

194. Aksoy BA, Gao J, Dresdner G, Wang W, Root A, Jing X, Cerami E, Sander C (2013) PiHelper: an open source framework for drug-target and antibody-

target data. Bioinformatics 29 (16):2071-2072. doi:10.1093/bioinformatics/btt345

195. Khurana E, Fu Y, Chen J, Gerstein M (2013) Interpretation of genomic variants using a unified biological network approach. PLoS computational biology 9 (3):e1002886. doi:10.1371/journal.pcbi.1002886

196. Das J, Yu H (2012) HINT: High-quality protein interactomes and their applications in understanding human disease. BMC systems biology 6:92. doi:10.1186/1752-0509-6-92

197. Csardi G, Nepusz T (2006) The igraph software package for complex network research. InterJournal Complex Systems:1695