Loyola University Chicago
Loyola eCommons

Master's Theses | Theses and Dissertations

2021

The Utilization and Optimization of Omics Trait Prediction Models Within and Across Diverse Populations

Ashley Mulford

Follow this and additional works at: https://ecommons.luc.edu/luc_theses

 Part of the Bioinformatics Commons

Recommended Citation

Mulford, Ashley, "The Utilization and Optimization of Omics Trait Prediction Models Within and Across Diverse Populations" (2021). *Master's Theses*. 4362.
https://ecommons.luc.edu/luc_theses/4362

This Thesis is brought to you for free and open access by the Theses and Dissertations at Loyola eCommons. It has been accepted for inclusion in Master's Theses by an authorized administrator of Loyola eCommons. For more information, please contact ecommons@luc.edu.

This work is licensed under a Creative Commons Attribution-Noncommercial-No Derivative Works 3.0 License.
Copyright © 2021 Ashley Mulford

LOYOLA UNIVERSITY CHICAGO

THE UTILIZATION AND OPTIMIZATION OF OMICS TRAIT PREDICTION MODELS WITHIN AND ACROSS DIVERSE POPULATIONS

A THESIS SUBMITTED TO
THE FACULTY OF THE GRADUATE SCHOOL
IN CANDIDACY FOR THE DEGREE OF
MASTER OF SCIENCE

PROGRAM IN BIOINFORMATICS

BY
ASHLEY MULFORD
CHICAGO, IL
MAY 2021

Copyright by Ashley Mulford, 2021
All rights reserved.

ACKNOWLEDGMENTS

I would like to first specially thank Dr. Wheeler for giving me the opportunity to join her computational human genetics lab my first year at Loyola, despite my lack of programming skills at the time. Her support and guidance have deepened my understanding of diverse population genetics, statistical analyses, and pharmacogenomic phenotypes, and her mentorship over the past three years has shaped me into the passionate and dedicated scientist I am today. I would also like to thank all the current and former members of Wheeler lab who have made my thesis possible, with a special thanks to Ryan Schubert for answering all my statistics and coding questions and Elyse Geoffrey for providing comradery and support as we completed the BS/MS program together.

Thank you to all my past professors for providing me the vast foundation of scientific knowledge enabling my success in my research, with a special thanks to Dr. Kanzok and Dr. Banerjee for their participation on my thesis committee and their guidance and expertise in proteomics and statistics, respectively. Additionally, I would like to thank the head of the Bioinformatics Program, Dr. Putonti, for meeting with me many times to discuss and advise on my progress and plans as I completed my coursework. I would also like to thank the Biology Department and Bioinformatics Program for awarding me with the Biology Summer, Mulcahy, and Graduate Research Fellowships throughout my time at Loyola, all of which made my research possible.

Finally, I would like to thank my family and friends, and in particular my parents, who have unconditionally supported me and my goals my entire life. Thank you for always believing that I could do anything I set my mind to, for encouraging me to have big dreams, and for helping me make them happen.

Somewhere, something incredible is waiting to be known.

— Carl Sagan

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AFA	African American individuals from the MESA cohort
ALL	All individuals from 1000 Genomes Project populations YRI, CEU, and ASN
ALL-M	All individuals from MESA populations AFA, CHN, EUR, and HIS
ara-C	Cytarabine arabinoside
ASN	Han Chinese from Beijing, China and Japanese from Tokyo, Japan
AUC	Area under the dose-response curve
CEU	Individuals of European ancestries from Utah, USA
CHN	Chinese American individuals from the MESA cohort
EUR	European American individuals from the MESA cohort
GTEx	Genotype-Tissue Expression
GWAS	Genome-wide association studies
HIS	Hispanic American individuals from the MESA cohort
IC_{50}	Half-maximal inhibitory concentration
LCLs	Lymphoblastoid cell lines
MAF	Minor allele frequency
MESA	Multi-Ethnic Study of Atherosclerosis
PAS	Protein-based Association Studies
PCA	Principal components analysis
PXR	Pregnane X receptor

RN	Rank-normalized
SNP	Single nucleotide polymorphism
TOPMed	Trans-Omics for Precision Medicine
TWAS	Transcriptome-wide association studies
YRI	Yoruba from Ibadan, Nigeria

ABSTRACT

Most cancer chemotherapeutic agents are ineffective in a subset of patients; thus, it is important to consider the role of genetic variation in drug response. Lymphoblastoid cell lines (LCLs) derived from 1000 Genomes Project populations of diverse ancestries are a useful model for determining how genetic factors impact variation in cytotoxicity. In our study, LCLs from three 1000 Genomes Project populations of diverse ancestries were previously treated with increasing concentrations of eight chemotherapeutic drugs and cell growth inhibition was measured at each dose with half-maximal inhibitory concentration (IC_{50}) or area under the dose-response curve (AUC) as our phenotype for each drug. We conducted genome-wide (GWAS), transcriptome-wide (TWAS), protein-based association studies (PAS) within and across ancestral populations. We identified four unique loci with GWAS, three genes with TWAS, and seven proteins with PAS significantly associated with chemotherapy-induced cytotoxicity within and across ancestral populations. For etoposide, increased *STARD5* predicted expression associated with decreased etoposide IC_{50} ($p = 8.5 \times 10^{-8}$). Functional studies in A549, a lung cancer cell line, revealed that knockdown of *STARD5* expression resulted in decreased sensitivity to etoposide following exposure for 72 ($p = 0.033$) and 96 hours ($p = 0.0001$). By identifying loci, genes, and proteins associated with cytotoxicity across ancestral populations, we strive to understand the genetic factors impacting the effectiveness of chemotherapy drugs and to contribute to the development of future cancer treatment.

CHAPTER ONE

INTRODUCTION

Cancer Genomics and Treatments

The Cancer Genome and Common Variants

Cancer is a complex disease with genetic, environmental, and lifestyle-based risk factors and in recent years it has become a leading cause of death globally (Torre et al. 2016). There are more than 100 distinct types of cancer that can occur across tissues, each with unique genetic characteristics (Stratton, Campbell, and Futreal 2009). The most common cancer types worldwide are prostate and lung cancer in men and breast cancer in women (Torre et al. 2016). Cancer arises when a series of somatic mutations occur within a cell, allowing it to proliferate without regulation and, in many cases, metastasize (Stratton, Campbell, and Futreal 2009; Shibata 2012). Currently, more than 350 protein-coding genes in the human genome have been found to be mutated in various cancer types and likely contribute to cancer development (Stratton, Campbell, and Futreal 2009). Of these mutations, around 90% have been found to be dominant in effect, meaning mutation in only one allele will contribute to the cell becoming cancerous (Stratton, Campbell, and Futreal 2009). Additionally, some types of cancer emerge when a cell incorporates viral DNA, such as the development of cervical cancer in individuals that contracted human papilloma virus (Stratton, Campbell, and Futreal 2009).

Of the protein-coding genes that have been implicated in cancer development, some occur more frequently across cancer types while others are unique to specific tumors. Somatic

mutations in *TP53,* a tumor suppressor gene, are found in more than half of all human cancers spanning many tissues including brain, breast, lung, ovarian, and colorectal carcinomas (Olivier, Hollstein, and Hainaut 2010; Leroy, Anderson, and Soussi 2014). The gene *TP53* encodes the protein p53; wildtype p53 functions to suppress tumor development by regulating transcription and inducing apoptosis (Ko et al. 2019). Mutations in *TP53* commonly occur in the DNA-binding domain of p53, resulting in a reduction in the ability to bind DNA and mediate transcription in the mutated protein (Baugh et al. 2018). These mutations occur across approximately 190 codons, most often as missense mutations resulting in single-amino acid changes rather than as frameshift or nonsense mutations, which are more common in other tumor suppressor genes (Olivier, Hollstein, and Hainaut 2010; Baugh et al. 2018). Additionally, a greater number of mutations in *TP53* is correlated with increasingly altered structure of the p53 protein, resulting in functional changes that promote a cancerous phenotype (Baugh et al. 2018).

Other tumor suppressor genes commonly implicated in cancer are *BRCA1* and *BRCA2,* which both regulate transcription and DNA repair in response to damage (Yoshida and Miki 2004). The proteins encoded by *BRCA1* and *BRCA2* have been found in complexes to repair double stranded breaks in DNA in addition to having independent functions in transcription mediation and cell cycle regulation (Yoshida and Miki 2004; Varol et al. 2018). *BRCA1* and *BRCA2* mutations are associated with increased susceptibility to breast, ovarian, and prostate cancers (Yoshida and Miki 2004). As some *BRCA* mutations are germline, increased cancer susceptibility is hereditary; women with inherited *BRCA* mutations therefore have a 45% to 75% chance of developing breast cancer within their lifetime (Baretta et al. 2016). Breast cancers with *BRCA* mutations have also been found to be more aggressive and are correlated with higher mortality rates (Baretta et al. 2016).

Although common mutations in tumor suppressor and other cancer-associated genes have been widely studied, much is still unknown about the mechanisms through which these mutations promote cancer development and progression. By conducting studies on the cancer genome, the functions of common mutants associated with cancer development, such as those arising from *TP53* and *BRCA*, can be better understood. Additionally, genetic studies exploring the effectiveness of cancer treatments allow for the identification of new variants and genes associated with treatment phenotypes.

Chemotherapeutic Drugs and Mechanisms

Chemotherapy-based treatments for cancer emerged in the early 1900s; however, use of chemotherapeutics did not become widespread until the 1960s when studies demonstrated they could be used to cure more advanced cancers that were less responsive to surgery and radiation therapy (DeVita and Chu 2008). The discoveries of various chemotherapeutics allowed for targeted treatments to emerge and adjuvant chemotherapy methods to arise, using multiple methods of treatment in conjunction to produce better patient outcomes (DeVita and Chu 2008). A common example of this is the use of chemotherapeutics to reduce the size of the tumor before surgery, in effort to improve the likelihood of complete extraction and preserve more of the surrounding healthy tissue (DeVita and Chu 2008). Subsequently, the advancements provided by chemotherapy have caused cancer mortality rates to continually decline since 1990 (DeVita and Chu 2008).

Platinum-based drugs are a common class of chemotherapeutics; these include cisplatin, carboplatin, and oxaliplatin, all of which are widely used to treat various cancer types (Hato et al. 2014). The reactive platinum in these drugs is able to covalently bind to DNA to form platinum-

DNA adducts, which disrupt DNA repair mechanisms, causing cancerous cells to induce apoptosis (Dasari and Tchounwou 2014; Hato et al. 2014). Recent studies have found that platinum-based chemotherapeutics may also have anticancer effects as a result of immune system modulation (Hato et al. 2014). Treatments with platinum-based drugs have been found to enhance T-cell activation, strengthening the immune response towards cancerous cells, and to regulate the phosphorylation of STAT signaling proteins that then interact with programmed death receptors to induce cell death (Hato et al. 2014). However, platinum-based chemotherapeutics also come with challenges. For cisplatin in particular, negative side effects can occur, including severe kidney problems, hearing loss, gastrointestinal disorders, and hemorrhage (Dasari and Tchounwou 2014). Additionally, cisplatin-resistance is common; thus, combination therapies with radiation or other chemotherapeutics are used to provide effective treatment of resistant tumors (Dasari and Tchounwou 2014).

One drug often used jointly with cisplatin to treat resistant tumors is paclitaxel. Paclitaxel was found to be an effective anticancer drug in the 1980s when a clinical study found 30% of patients with advanced ovarian cancer responded positively to treatment (Weaver 2014). Currently, paclitaxel is used primarily to treat breast, ovarian, and lung cancers (Weaver 2014; Zhu and Chen 2019). Paclitaxel inhibits microtubule production by reducing the concentration of tubulin subunits in the cell and it also binds to existing microtubules and interferes with their function in cell division, leading to mitotic arrest and, ultimately, cell death (Weaver 2014; Abu Samaan et al. 2019). Paclitaxel also has positive immunological effects, as it promotes the activation and proliferation of T cells and natural killer cells, bolstering the body's own immune response to cancer cells (Zhu and Chen 2019). Resistant ovarian cancers treated with a

combination of cisplatin and paclitaxel had a 73% better response rate than those treated with cisplatin alone (Dasari and Tchounwou 2014).

Another common class of chemotherapeutics are antineoplastic drugs; these inhibit DNA topoisomerases, which are responsible for cutting and pasting both single- and double-stranded DNA (Hande 1998). The antineoplastic drug etoposide inhibits topoisomerase II, disrupting DNA replication, recombination, and transcription in malignant cells, resulting in increased DNA degradation and apoptosis (Hande 1998). Etoposide is used to treat both small and non-small cell lung cancers, gastric and testicular cancers, and lymphoma, with response rates ranging from 10% to 45% (Hande 1998).

Although chemotherapy is a widely effective treatment for various cancer types, limitations exist. Varied patient responses, including the development of drug-resistant tumors that require combination therapies, and the degree of tumor progression both impact the success of chemotherapy treatments (Galmarini, Galmarini, and Galmarini 2012; Stordal et al. 2012; Marin et al. 2009). Moreover, finding effective treatments for metastatic cancer is especially challenging, despite recent developments in targeted therapy and cancer immunology. (Roy and Saikia 2016; Galmarini, Galmarini, and Galmarini 2012). Therefore, personalized approaches to cancer medicine that deepen our understanding of the genetic variants and biological mechanisms impacting a patient's response to treatment are necessary in order to successfully cure advanced cancers (Jackson and Chester 2015).

Lymphoblastoid Cell Lines

One method for identifying factors that impact drug efficacy and patient response is to conduct pharmacogenomic studies of chemotherapeutics, which involve treatment with drug, quantitative measurement of response or cytotoxicity, and statistical analysis of a response or

cytotoxicity phenotype in relation to genomic, transcriptomic, or proteomic variation. Cancer pharmacogenomic studies are often performed using *in vitro* human cell lines models, including lymphoblastoid cell lines (LCLs) and cancer cell lines from various tissues (Niu and Wang 2015). LCLs are derived by infecting blood lymphocytes with the Epstein-Barr virus; this immortalizes the cell population, providing a model that continuously proliferates without becoming tumorigenic (Hussain and Mulherkar 2012). The widespread availability and relative affordability of cell lines makes it easier to conduct initial studies with *in vitro* models rather than clinically in patients (Niu and Wang 2015; Heather E. Wheeler and Dolan 2012).

LCLs from the International HapMap and 1000 Genomes Projects serve as one effective model for determining genetic factors contributing to chemotherapeutic cytotoxicity because they have extensive genetic information and environmental factors can be controlled (Heather E. Wheeler and Dolan 2012). There are also LCLs derived from a multitude of ancestral populations making them particularly useful for studying how cytotoxicity varies across ancestral populations (Heather E. Wheeler and Dolan 2012; International HapMap Consortium 2003; 1000 Genomes Project Consortium et al. 2015). Studies conducted in LCLs do have limitations though, as complex drug effects and interactions that exist in the body cannot be fully determined *in vitro* and treatment with a single drug does not allow for analysis of the factors contributing to the effectiveness of combination therapies, which are commonly used on less-responsive tumors (Heather E. Wheeler and Dolan 2012; Roell et al. 2019). Overall, LCLs provide a promising model for pharmacogenomic studies due to their vast utility, and they have enabled the identification of variants involved in cancer progression and may contribute to the development of more effective and personalized cancer treatments.

Multi-Omics Approaches in Genetic Studies

Genome-Wide Association Studies

Genome-wide association studies (GWAS), which emerged in the early 2000s, are a powerful computational tool used to identify genotypic variants in the form of single nucleotide polymorphisms (SNPs) associated with a given phenotype (Bush and Moore 2012; Ku et al. 2010). The human genome contains millions of SNPs that can have significant phenotypic implications as they can impact RNA transcript stability and cause amino acid changes that could potentially alter protein structure and function (Bush and Moore 2012). The majority of SNPs have two alleles, with the major allele occurring with greater frequency than the minor allele in a given population (Bush and Moore 2012). Commonly occurring alleles generally have lower penetrance, meaning they have smaller genetic effects (Bush and Moore 2012). Consequently, the heritability of complex diseases is determined through the combination of a multitude of alleles, which can be identified with GWAS (Bush and Moore 2012).

Conducting GWAS requires both genotype and phenotype data for a group of individuals; phenotype data must be measured quantitively and can either be continuous or in the form of cases and controls (Bush and Moore 2012). GWAS implement linear modeling to test the null hypothesis that there is no significant difference in phenotype between alleles of a SNP; millions of SNPs are analyzed and those found to be significantly associated with the phenotype can then be further investigated (Bush and Moore 2012). As a result of linkage disequilibrium, which is the non-random correlation of alleles at a given locus, not all SNPs identified through GWAS will be causal; false positives that appear to associate with the phenotype may occur due to linkage to the causal SNP (Bush and Moore 2012). Thus, while GWAS are useful for identifying novel variants associated with complex traits, additional studies are necessary to

better understand and validate findings so that they may one day be applied to improve treatment.

As GWAS have become more established, many software tools have been developed to allow for greater utility and more accurate results. Genome-wide efficient mixed-model association (GEMMA), which uses linear mixed modeling, is one of those tools (Zhou and Stephens 2012). GEMMA rapidly produces results even with large sample sizes (Zhou and Stephens 2012). Additionally, GEMMA can adjust for population-based covariates, including ancestry and relatedness, which allows for admixed populations to be analyzed and related individuals to remain in samples rather than be filtered out as they would skew results if not accounted for (Zhou and Stephens 2012).

Cancer GWAS

The emergence of GWAS provided a novel approach for investigating the role of genetic variants in cancer. As of 2017, more than 700 SNPs associated with increased risk for various malignancies had been identified, providing new insight into the heritability of cancer (Sud, Kinnersley, and Houlston 2017). More than 90% of these variants are located within non-coding regions of the genome, such as intergenic and intronic regions, rather than in protein-coding regions, making them challenging to interpret (Chen et al. 2019). However, when the SNPs are located within protein-coding regions the results can be promising, as further research can then be conducted on the possible role of gene expression levels, protein functions, and chemical pathways on cancer development (Sud, Kinnersley, and Houlston 2017; Liang et al. 2020).

In addition to providing insight into the genetics of and biochemical mechanisms involved in cancer risk, GWAS can also help to contextualize known environmental factors that can lead to cancer development. Several GWAS identified significant SNPs associated with both

nicotine dependence and lung cancer susceptibility within the genes *CHRNA3*, *CHRNA5*, and *CHRNB4*, all of which encode nicotinic acetylcholine receptor subunits (Bossé and Amos 2018). These findings demonstrate the relationship between smoking, a well-known environmental risk factor, and lung cancer development, adding to our understanding of how environmental and genetic components impacting cancer risk are related (Bossé and Amos 2018). While many significant loci associated with cancer risk have been found, these variants generally have low penetrance and only account for a small percentage of heritability (Liang et al. 2020). In order to better understand the genetic factors impacting cancer risk, additional association studies can be performed to directly identify significant gene expression and protein levels that play a role in malignancy.

Transcriptome-Wide Association Studies

Although GWAS identify associations at the SNP level, they do not provide insight into the underlying biochemical mechanisms that regulate traits (Gamazon et al. 2015). Transcriptome-wide association studies (TWAS) are another method for analyzing factors impacting phenotype as they identify genes with significant expression levels that can then be further studied to determine their role in regulating traits (Gamazon et al. 2015; Barbeira et al. 2019; Mogil et al. 2018). One widely used tool for conducting TWAS is PrediXcan, which employs statistical modeling to predict transcript expression levels from genotypes and determine associations between predicted gene expression and phenotype (Gamazon et al. 2015). Through predictive modeling, PrediXcan provides an accessible method to analyze gene expression levels and their impact on phenotype as the user does not need to have transcript data but only genomic data, as they would for GWAS, or GWAS summary statistics; this is notable as it eases the process of studying the transcriptome, which historically has been more challenging

due to the rapid rate of degradation of RNA samples and human tissue accessibility (Gamazon et al. 2015; Barbeira et al. 2018).

The prediction models used in PrediXcan were trained with cross-validated Elastic Net regularization of genotype and transcriptomic data from approximately 20,000 samples from 48 tissue types primarily from the Genotype-Tissue Expression (GTEx) Project (Gamazon et al. 2015). These models can be used to predict tissue-specific gene expression levels from genotypes and identify associations with phenotypes. Additional predictive models also derived with Elastic Net were trained with transcriptomic data from monocytes from diverse populations from the Multi-Ethnic Study of Atherosclerosis (MESA) cohort and tested in independent cohorts (Mogil et al. 2018; Bild et al. 2002). These models differ from the GTEx models as they can be used to predict population-specific gene expression levels. Another tool for conducting TWAS is MulTiXcan, which uses the same GTEx models as PrediXcan but derives results by aggregating expression levels to find associations across tissues rather than to find tissue-specific associations (Barbeira et al. 2019). Most importantly, both PrediXcan and MulTiXcan can aid in contextualizing GWAS results, as they implicate gene regulation in relation to phenotype and provide the direction of effect for each association. Thus, conducting TWAS in addition to GWAS enables researchers to better identify the biochemical mechanisms impacting phenotype, as the combination of associations with SNPs and gene expression levels creates a more cohesive understanding of the factors regulating traits.

Advantages of Studying Proteomic Variants

Both GWAS and TWAS have become prominent computational tools in the field of human genetics, enabling scientists to expand their knowledge of the variants impacting complex traits. Yet, a truly holistic understanding of the biological processes regulating phenotypes

requires a multi-omics approach where genomic, transcriptomic, and proteomic variants are all analyzed (Hasin, Seldin, and Lusis 2017; I. Subramanian et al. 2020). While the transcriptome has been more widely studied due to the larger and more complete nature of transcriptomic data sets, the proteome has become the subject of more recent analyses as high-throughput technologies have amassed large proteomic datasets (Liu 2008; Aslam et al. 2017). Proteomic data is far more dynamic than genomic and even transcriptomic data, as protein expression levels, structure, and function vary depending on cell type, conditions, and conformations, whereas genomic data is consistent across cell type and transcriptomic data accounts for primarily tissue-based expression differences (Manzoni et al. 2018). Moreover, analyzing the proteome is vital in understanding gene function, as many proteins undergo post-translational modifications, resulting in complexities in regulation and protein function that studying the genome and transcriptome alone will not account for (Aslam et al. 2017). Thus, the intricacies of the proteome can provide clarity into the biological mechanisms underlying disease development and progression, while also challenging us to create methods of analysis accounting for greater degrees of complexity.

Computational omics studies all rely on statistical testing to identify significant associations with phenotype; when testing integrates multi-omics data the results can be compared across the genome, transcriptome, and proteome to identify novel regulating pathways and find commonalties that further implicate and contextualize mechanisms (Hasin, Seldin, and Lusis 2017; I. Subramanian et al. 2020). Although progress have been made in the development of software tools designed for proteomic studies, there are still advancements needed to improve performance and expand the degree with which the full proteome can be studied (Aslam et al. 2017). Protein-based association studies (PAS), for example, take statistical analysis a step

beyond TWAS to identify significant proteins associated with a given phenotype; however, the software tools for performing PAS are still being developed and necessary data is still being collected, so they are not truly proteome-wide yet, as only a subset of proteins have been included in predictive modelling or other analysis methods (Okada et al. 2016; Brandes, Linial, and Linial 2020). Nonetheless, proteomic studies have versatile applications, as their results not only provide greater insight into the biochemical factors regulating disease risk, but also enable further analyses into how proteomic variation impacts treatment (Manzoni et al. 2018). Significant protein associations identified through PAS have more therapeutic application than significant SNPs or transcripts from genomic and transcriptomic studies, as the functions and relevant mechanisms of significant proteins can be more directly explored through clinical experimentation (Doll, Gnad, and Mann 2019; Ahmed 2020). Consequently, when specific biochemical pathways are implicated, scientists can begin developing more personalized treatments that effectively target the proteins involved (Ahmed 2020).

One organization seeking to expand access to proteomic data for its utilization in computational analyses of disease traits is the NHLBI Trans Omics for Precision Medicine (TOPMed) Consortium (Raffield et al. 2020). The TOPMed Consortium includes proteomic data from various studies, including the MESA cohort (Bild et al. 2002; Raffield et al. 2020). Proteomic data was collected for approximately 1,300 proteins from blood plasma samples using SOMAscan aptamer-based arrays, which measure protein levels through the binding of the target protein to a specific aptamer (Gold et al. 2010; Raffield et al. 2020). Looking forward, this data can be used in future studies to find associations between protein levels and diseases, providing new insight into how omics traits regulate phenotype and their larger role in human health.

Several studies have investigated the potential applications of proteomic analyses on cancer precision medicine (Tyanova and Cox 2018; Uzozie and Aebersold 2018; Doll, Gnad, and Mann 2019; Giudice and Petsalaki 2019). While cancer has been the focus of many other genetic studies, including GWAS and TWAS that have identified hundreds of significant SNPs and transcript associations, proteomic studies greatly expand on previous findings, as determining the functionalities of implicated proteins is more relevant in understanding the mechanisms regulating complex cancer phenotypes (Doll, Gnad, and Mann 2019). The characterization of proteins associated with cancer risk and prognosis enables the option of preventative measures for high-risk patients and the determination of the best course of treatment for patients with cancer (Tyanova and Cox 2018; Sellami and Bragazzi 2020). Proteomic studies also provide insight into cancer-specific biochemical pathways, which could potentially be useful in the development of targeted therapies (Uzozie and Aebersold 2018). Cancer precision medicine has slowly advanced as computational and clinical pharmacogenomic studies have made beneficial discoveries; the first cancer drug based on genetic factors rather than tumor or tissue type was approved by the FDA in 2017 (Doll, Gnad, and Mann 2019). Overall, the use of computational methods for analyzing the role of proteomic variants in disease risk and treatment is vital, as future clinical studies can further explore relevant proteins to enable the development of more effective and personalized treatments.

Diversity in Genetic Studies

In the past two decades, genetic studies have identified and contextualized a myriad of genomic, transcriptomic, and proteomic variants impacting phenotypes; however, these studies are often lacking the diversity, as the vast majority of participants are of European ancestries.

This discrepancy can be illustrated with GWAS, as 81% of participants across the more than 3,000 studies published as of 2018 were of European ancestries (Hindorff et al. 2018). This is detrimental as alleles and allele frequencies differ across human populations; thus, disproportionately analyzing data from one ancestral population over others results in fewer significant variants being identified and some rare variants found only within certain populations not being included at all (Hindorff et al. 2018). Consequently, this lack of representation hinders our understanding of how genetic differences affect disease and treatment, limiting the clinical application of findings, as the bias from studying predominantly European populations yields incomplete results (Sirugo, Williams, and Tishkoff 2019).

The 1000 Genomes Project (phase 3) aimed to expand diversity in human genetic research by performing whole-genome sequencing on 26 ancestral populations from around the world and creating a publicly available platform where the data collected could be accessed and utilized in genetic studies (1000 Genomes Project Consortium et al. 2015). Through this project, more than 88 million SNPs were genotyped; notably, African ancestral populations had the highest proportions of population- and continent-specific SNPs, as well as the greatest total numbers of SNPs, at about 5 million per genome (1000 Genomes Project Consortium et al. 2015). These populations have since been used in hundreds of studies, which subsequently implicated a plethora of novel variants in phenotypic regulation (S. L. Park, Cheng, and Haiman 2018). These findings demonstrate that the development of precision medical treatments is dependent on greater diversity in genetic studies.

Summary

There have been a number of previous studies demonstrating the impacts of genomic

variation on chemotherapeutic drug response (Niu and Wang 2015; R. S. Huang, Duan, Bleibel, et al. 2007; H. E. Wheeler et al. 2013; R. S. Huang, Duan, Shukla, et al. 2007; Bleibel et al. 2009; R. S. Huang, Duan, Kistner, Hartford, et al. 2008; R. S. Huang, Duan, Kistner, Bleibel, et al. 2008; O'Donnell et al. 2012; Hartford et al. 2009). In this project, we sought to expand on prior findings by conducting GWAS, TWAS and PAS on drug-response phenotypes from eight chemotherapeutics measured in HapMap LCLs derived from three ancestral populations consisting of individuals with African, Asian, and European ancestries. By including individuals of diverse backgrounds in this study, we identified associations both within and across ancestral populations. Previous GWAS were conducted on subsets of these individuals before the 1000 Genomes Project was complete, thus at that time many individuals had been genotyped through the HapMap Project but not sequenced (R. S. Huang, Duan, Bleibel, et al. 2007; Bleibel et al. 2009; Komatsu et al. 2015; Gamazon et al. 2013; H. E. Wheeler et al. 2013; R. S. Huang, Duan, Shukla, et al. 2007; R. S. Huang, Duan, Kistner, Hartford, et al. 2008; O'Donnell et al. 2012; R. S. Huang, Duan, Kistner, Bleibel, et al. 2008; Gamazon et al. 2018; Hartford et al. 2009; 1000 Genomes Project Consortium et al. 2015; International HapMap Consortium 2003). In this study, all individuals were either sequenced or imputed with the 1000 Genomes as reference, allowing more SNPs to be analyzed. We also performed TWAS and PAS on these data for the first time to discover gene- and protein-based associations and gain further insight into the underlying mechanisms involved in regulating drug response. Moreover, for the most significant gene identified, *STARD5*, we validated our results by performing knockdown experiments in a lung cancer cell line treated with the associated chemotherapeutic, etoposide. By conducting GWAS, TWAS, and PAS, confirming our results experimentally, and incorporating diverse ancestral

populations, we aimed to cultivate a deeper understanding of the genomic factors and biochemical mechanisms impacting chemotherapy drug response and contribute to the development of future precision cancer treatment.

CHAPTER TWO

METHODS

Publication disclaimer

This work was previously published in Human Molecular Genetics (2021) doi.org/10.1093/hmg/ddab029 with the following authors:

Ashley J. Mulford[1,2], Claudia Wing[3], M. Eileen Dolan[3], Heather E. Wheeler[1,2]

[1]Department of Biology, Loyola University Chicago, Chicago, IL, USA, [2]Program in Bioinformatics, Loyola University Chicago, Chicago, IL, USA, [3]Section of Hematology/Oncology, Department of Medicine, University of Chicago, Chicago, IL, USA

Data Preparation

We procured cytotoxicity phenotypes measured in HapMap LCLs from previous studies of eight chemotherapy drugs, including ara-C, capecitabine, carboplatin, cisplatin, daunorubicin, etoposide, paclitaxel, and pemetrexed (R. S. Huang, Duan, Bleibel, et al. 2007; Bleibel et al. 2009; Komatsu et al. 2015; Gamazon et al. 2013; H. E. Wheeler et al. 2013; R. S. Huang, Duan, Shukla, et al. 2007; R. S. Huang, Duan, Kistner, Hartford, et al. 2008; O'Donnell et al. 2012; R. S. Huang, Duan, Kistner, Bleibel, et al. 2008; Gamazon et al. 2018; Hartford et al. 2009). These LCLs were derived from 178 individuals from the Yoruba population in Ibadan, Nigeria (YRI), 178 individuals with European ancestries from Utah, United States (CEU), and 90 individuals from a combined population of Han Chinese from Beijing, China and Japanese from Tokyo, Japan (ASN). The YRI population contained 58 parent-child trios and the CEU population contained 52 parent-child trios, which we accounted for when conducting our genetic analyses.

The numbers of LCLs with measured phenotypes varied for each drug (Table 1). Cellular sensitivity to each drug was recorded as the area under the dose-response curve (AUC) for ara-C, capecitabine, paclitaxel, and pemetrexed, and as the half-maximal inhibitory concentration (IC_{50}) for carboplatin, cisplatin, daunorubicin, and etoposide. These concentrations were all measured after 72 hours of exposure to the corresponding chemotherapeutic. We rank-normalized (RN) the AUC or IC_{50} for use in our subsequent genetic analyses. Additionally, once phenotypic data was collected for each ancestral population and drug, genotypic data were imputed using BEAGLE; all genotypes were in Genome Build 37 and only autosomal variants were analyzed (Browning and Browning 2007).

Table 1. Individuals with genotype and phenotype data. Counts given for each ancestral population and drug combination.

		Population			
		CEU	**YRI**	**ASN**	**ALL**
Drug	**Ara-C (RN AUC)**	165	177	90	432
	Capecitabine (RN AUC)	165	175	90	424
	Carboplatin (RN IC_{50})	168	172	84	430
	Cisplatin (RN IC_{50})	166	175	90	431
	Daunorubicin (RN IC_{50})	86	173	0	259
	Etoposide (RN IC_{50})	84	171	0	255
	Paclitaxel (RN AUC)	77	87	0	164
	Pemetrexed (RN AUC)	84	176	0	260

Genome-Wide Association Studies

GWAS with Ancestral Populations

Some individuals with HapMap LCLs used in this study were sequenced in the 1000 Genomes Project and some had genotypes only. Individuals genotyped in HapMap r28, but not sequenced, were previously imputed to 1000 Genomes (Komatsu et al. 2015). Imputation was performed using BEAGLE version 3.3.2, which considers the relatedness of the trios in the imputation (Browning and Browning 2007). We used SNPs with imputation $R^2 > 0.8$, population minor allele frequency (MAF) > 0.05, and in Hardy–Weinberg equilibrium ($P > 1 \times 10^{-6}$) in our studies.

Prior to conducting GWAS, we created a relatedness matrix for each of the ancestral populations, YRI, CEU, and ASN, using GEMMA. For each ancestral population we used the genotype dosages, with a minimum MAF of 0.05, to calculate the centered relatedness matrix. We then used GEMMA version 0.98.1 to conduct GWAS using the linear mixed model Wald test for each ancestral population and corresponding phenotypes (Table 1) (Zhou and Stephens 2012). After conducting GWAS, we created QQ, Manhattan, and LocusZoom plots to aid in visualizing our results. We made the QQ and Manhattan plots in R using the package qqman and created the LocusZoom plots with the single plot service on http://locuszoom.org/ (Turner 2014; Pruim et al. 2010). We made LocusZoom plots for all SNPs with genome-wide significance ($p < 5 \times 10^{-8}$) and we used the corresponding 1000 Genomes Nov. 2014 ancestral population when generating the LocusZoom plots.

GWAS with Combined Population

To organize data for the ALL population, we combined the BIMBAM files for both the genotype and phenotype data from each ancestral population into single files. We then used a

subset of 100,000 SNPs to convert the BIMBAM files into PLINK files, which we needed to conduct principal components analysis (PCA) with KING (Manichaikul et al. 2010; Purcell et al. 2007). We used the covariates calculated by KING to account for population stratification in the ALL population. We also plotted the first three principal components to demonstrate that they accounted for population-based variation (Figure 1). Once these covariates were obtained, we generated a relatedness matrix for ALL and then conducting GWAS using the same methods as described for the ancestral populations, with the only difference being the inclusion of the covariates generated with PCA when conducting GWAS. We generated QQ, Manhattan, and LocusZoom plots as well, using the same methods (Pruim et al. 2010). As the ALL population does not correspond to a single 1000 Genomes Nov. 2014 population, we made multiple LocusZoom plots for each genome-wide significant SNP, each with a different ancestral population included in the ALL.

Transcriptome-Wide Association Studies

We conducted TWAS with PrediXcan on both the ancestral and combined populations for all applicable phenotypes, using the GTEx v7 and MESA prediction models (Gamazon et al. 2015; Mogil et al. 2018; Barbeira et al. 2018). PrediXcan was used to calculate the predicted expression levels for each gene. We then used GEMMA to perform a total of 7,487,956 association tests, as this enabled us to account for relatedness within the populations with the matrices created previously. To use GEMMA for this purpose, we reformatted the predicted expression matrices outputted by PrediXcan into a readable format for GEMMA, so the association tests could be performed. This produced results specific to each prediction model for each population and phenotype combination. Additionally, we conducted TWAS with MulTiXcan for the same populations and phenotypes, using the GTEx v7 prediction models only

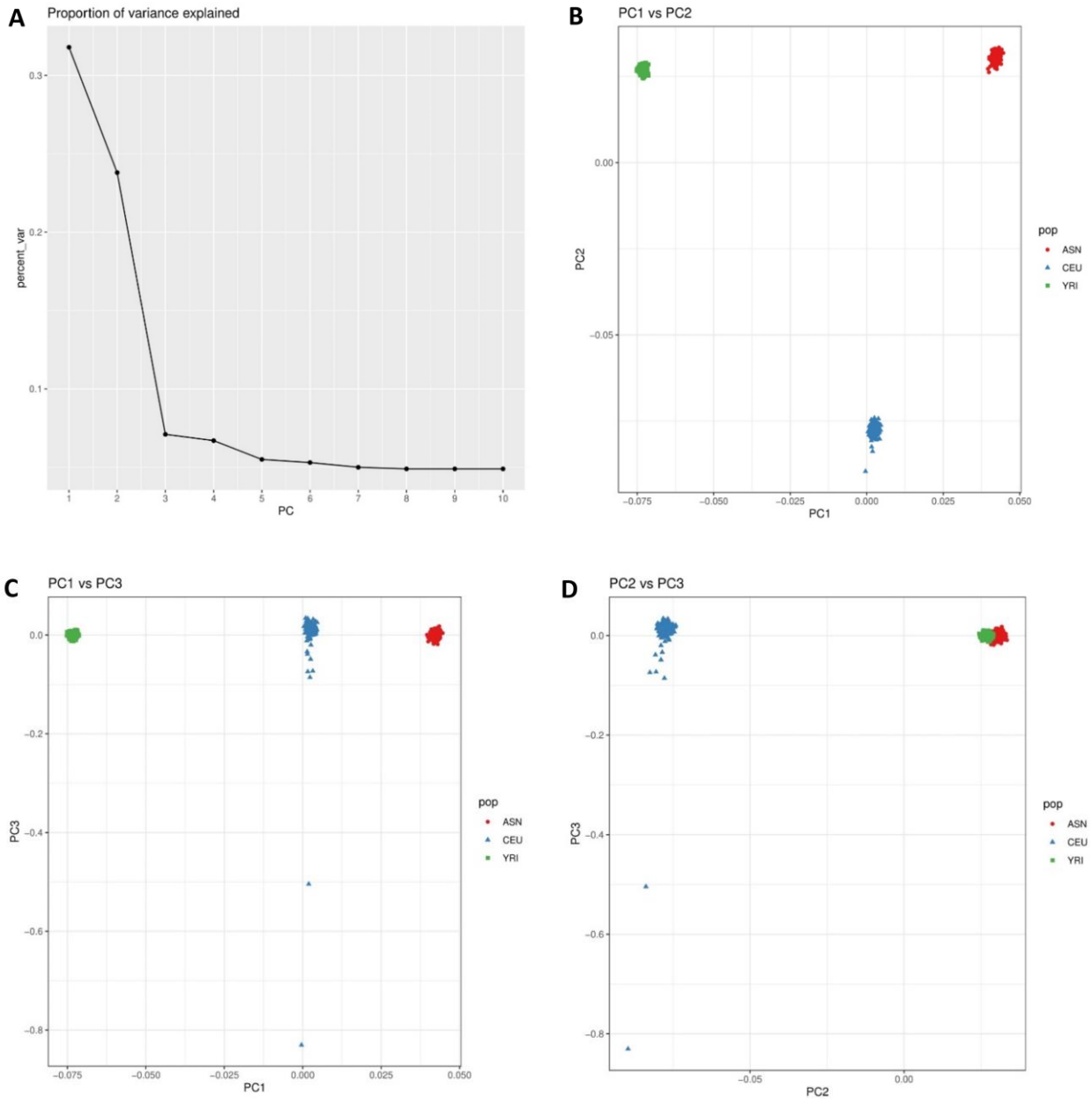

Figure 1. Principal component analysis (PCA) of genotype data. (A) Scree plot showing the percentage of variance accounted for by each of the ten PCs. (B) PC1 plotted against PC2 for each individual, colored by ancestral population: ASN, CEU, or YRI. (C) PC1 plotted against PC3 for each ancestral population. (D) PC2 plotted against PC3 for each ancestral population.

(Barbeira et al. 2019). We did not use GEMMA to conduct these association tests, as MulTiXcan aggregates across prediction models to find overall associations and GEMMA does not conduct

the association tests in this manner. Using MulTiXcan, we performed 727,944 association tests and produced a single set of results for each population and phenotype combination, containing overall rather than model-specific associations. For the ALL population, we included the covariates generated from PCA when performing the association tests with both GEMMA and MulTiXcan to account for population stratification. We then adjusted the p-values derived from both GEMMA and MulTiXcan using Bonferroni correction, to determine which genes had significant predicted expression levels associated with drug cytotoxicity. For each significant gene, we then created predicted expression plots in R using the package ggplot2, which plot the gene's predicted expression level against the chemotherapy phenotype (either IC_{50} or AUC) for each individual (Wickham 2016).

Gene Set Enrichment Analyses

After performing TWAS on each population and cytotoxicity phenotype, we used the FUMA tool GENE2FUNC to perform gene set enrichment analysis of the results from PrediXcan and MulTiXcan (Watanabe et al. 2017). One GENE2FUNC query was made for each ancestral population and phenotype combination. We submitted two lists of genes for each query, one for background genes, which contained all the genes analyzed during TWAS, and one for genes of interest, which contained a significant subset of genes based on either the PrediXcan or MulTiXcan results we generated previously. To achieve a subset of approximately 100 genes in each genes of interest list, we used a significance threshold of unadjusted p-value < 0.0005 for all the PrediXcan results and unadjusted p-value < 0.005 for all the MulTiXcan results. The PrediXcan results, which were derived from multiple prediction models, were combined so that the top genes across all models were selected for each ancestral population and phenotype. For the GENE2FUNC optional parameters, we used all the default options except for gene

expression data sets, for which we selected GTEx v7: 53 tissue types and GTEx v7: 30 general tissue types, as these correspond to the prediction models we used when conducting TWAS. We report significant gene sets that are enriched in each run of PrediXcan or MulTiXcan for each ancestral population and phenotype with adjusted p (Benjamini-Hochberg FDR) < 0.05.

Gene Knockdown Experiments

Cancer Cell Lines

We obtained non-small cell lung cancer line A549 (CCL-185) from ATCC (Manassas, VA). IDEXX BioResearch (Columbia, MO) performed authentication of the cancer cell line, Case # 12135-2020, by using the Promega CELL ID System (Madison, WI) with 8 short tandem repeat markers (CSF1PO, D13S317, D16S539, D5S818, D7S820, TH01, TPOX, vWA) and amelogenin (for sex).

Compound preparations

We dissolved etoposide (Sigma-Aldrich, St. Louis, MO) in DMSO to obtain a stock solution of 10 mM and filtered using a 0.22 µm solvent resistant filter (EMD Millipore, Billerica, MA, USA) for sterility. We serially diluted the stock in media for final concentrations of 5 to 100 µM for treatment of the A549 cancer cell line. Vehicle control was 0.1% DMSO in media.

Cellular Assay with *STARD5* knockdown

We maintained A549 cells in F-12K media (Life Technologies; Carlsbad, CA) supplemented with 10% FBS (Hyclone, Fisher Scientific; Hanover Park, IL) and 1% Penicillin-Streptomycin (Life Technologies). We incubated cultures in a humidified incubator at 37°C with 5% CO_2. We performed knockdown of *STARD5* using a modified reverse transfection method (Thermo Fisher "Literature Code: 00189-08-C-01-U"). We mixed ON-TARGETplus SMARTpool siSTARD5 or ON-TARGETplus non-targeting pool (siSCR) purchased from

Dharmacon Inc. (Lafayette, CO) with DharmaFECT1 (Dharmacon Inc.) as per manufacturer's recommendations to create the transfection mix. We added complete media siSTARD5 or siSCR complex to produce 25nM final concentrations of each, then added the mixture to a cell pellet such that the final concentration of cells was 6000 cells/100 µL volume and plated into 96-well flat bottom tissue culture plates (Cell Star; Quality Biologicals Inc., Gaithersburg, MD). As a quality control check of the effect of siRNA on cell growth rates, we assayed cell viability using CellTiter-Glo 2.0 (Promega; Madison, WI), which measures cellular ATP from 0 to 96 hours in control wells. At 24 hours, we replaced transfection media with media containing increasing concentrations of etoposide (5 to 100 µM). To determine cellular sensitivity to etoposide in presence of siSTARD5 or siSCR, we incubated cells with drug for 72 and 96 hours followed by cell viability assays using CellTiter-Glo 2.0.

Quantitative reverse transcription PCR analysis

At 0, 72, and 96 hours post-drug treatment, we added trypsin to wells of A549 cells (6,000 cells/well) containing siSTARD5 or siSCR and combined, pelleted, and stored the cells at -80^{o}C. We extracted RNA using RNeasy Plus (Qiagen; Valencia, CA) and prepared cDNA from 500 ng RNA/sample with the High Capacity cDNA kit (Life Technologies). To determine *STARD5* knockdown in A549 cells, we performed quantitative reverse transcription PCR (qRT-PCR) for *STARD5*, Hs01075234_m1 and a control gene *B2M*, 4326319E (Life Technologies) using TaqMan Fast Gene Expression mix (Applied Biosystems; Foster City, CA). We ran each qRT-PCR in triplicate and determined gene expression levels using the relative standard curve method on the Viia7 (Life Technologies). We calculated percent knockdown by dividing the relative *STARD5* expression levels in the siSTARD5 sample by the *STARD5* expression in the non-targeting control (siSCR).

Derivation of Protein-based Prediction Models

We derived new prediction models using protein level data from the MESA cohort obtained from the TOPMed Consortium. We trained population-based prediction models using genotype and plasma protein data from a SOMAscan aptamer-based assay of 1335 proteins from individuals of African (AFA, n = 183), European (EUR, n = 416), Chinese (CHN, n = 71), and Hispanic/Latino (HIS, n = 301) ancestries in the TOPMed MESA multi-omics pilot study (Bild et al. 2002; Raffield et al. 2020). A total of five model groups were created from this data, corresponding to each separate population and one combined population (ALL-M). We used cross-validated elastic net regularization (alpha mixing parameter=0.5) using the R package glmnet with genetic variants within 1Mb of the gene encoding each protein as predictors for protein levels (Friedman, Hastie, and Tibshirani 2010). The models we derived were then tested in a separate population comprised of individuals of predominately European ancestries. We created database files, one for each population group, containing all protein models with Spearman correlation > 0.1 between predicted and observed levels, which were used as the models in the PAS we conducted. These models are referred to as the TOPMed prediction models in subsequent sections.

Protein-based Association Studies

We conducted protein-based association studies (PAS) with PrediXcan on both the ancestral and combined populations for all applicable phenotypes, using the TOPMed prediction models. As with TWAS, we used PrediXcan to calculate the predicted levels for each protein. We then reformatted the prediction matrices derived with PrediXcan for GEMMA, which we used to perform a total of 10,931 association tests, while accounting for relatedness in each ancestral population. This produced results specific to each prediction model for each population

and phenotype combination. We used Bonferroni correction to adjust the p-values in each set of results for multiple testing across models, to identify proteins with predicted levels significantly associated with cytotoxicity. For each significant protein we created plots in R using the package ggplot2, displaying the predicted protein levels versus the cytotoxicity phenotype (either IC_{50} or AUC) for each individual (Wickham 2016).

CHAPTER THRE

RESULTS

Publication disclaimer

This work was previously published in Human Molecular Genetics (2021)

doi.org/10.1093/hmg/ddab029 with the following authors:

Ashley J. Mulford[1,2], Claudia Wing[3], M. Eileen Dolan[3], Heather E. Wheeler[1,2]

[1]Department of Biology, Loyola University Chicago, Chicago, IL, USA, [2]Program in Bioinformatics, Loyola University Chicago, Chicago, IL, USA, [3]Section of Hematology/Oncology, Department of Medicine, University of Chicago, Chicago, IL, USA

Overview of Analyses

In order to investigate genetic and transcriptomic effects on chemotherapeutic toxicity, we gathered and analyzed previously published dose-response data from LCLs of three diverse ancestral populations (Komatsu et al. 2015; R. S. Huang, Duan, Bleibel, et al. 2007; Bleibel et al. 2009; Hartford et al. 2009; R. S. Huang, Duan, Shukla, et al. 2007; R. S. Huang, Duan, Kistner, Hartford, et al. 2008; O'Donnell et al. 2012; Gamazon et al. 2018; H. E. Wheeler et al. 2013; R. S. Huang, Duan, Kistner, Bleibel, et al. 2008; Gamazon et al. 2013). These LCLs were derived from 178 individuals from the Yoruba population in Ibadan, Nigeria (YRI), 178 individuals with European ancestries from Utah, United States (CEU), and 90 individuals from a combined population of Han Chinese from Beijing, China and Japanese from Tokyo, Japan (ASN). Both the YRI and CEU populations included parent-child trios. We used phenotypes from eight

chemotherapy drugs in our study. Depending on the drug, the cytotoxicity phenotype from each individual's LCL was calculated either with the half-maximal inhibitory concentration (IC_{50}) or the area under the dose-response curve (AUC). We rank-normalized (RN) these measurements for use in our genetic analyses. The total counts for individuals with both genotype and phenotype data varied for each drug and ancestral population. We then performed GWAS, TWAS, PAS, and gene set enrichment analyses to identify multi-omic traits significantly associated with chemotherapy-induced cytotoxicity (see overview in Figure 2).

GWAS reveal four loci associated with chemotherapy-induced cytotoxicity

We conducted GWAS using 1000 Genomes Project sequenced and imputed genotypes to identify genome-wide significant associations between SNPs and the cytotoxicity of each drug for each ancestral population (YRI, CEU, and ASN) and in all three ancestral populations combined (ALL) (1000 Genomes Project Consortium et al. 2015). We used GEMMA to perform univariate linear mixed model GWAS while accounting for relatedness in each ancestral population and population stratification in the ALL population using covariates generated with PCA (Zhou and Stephens 2012).We used a threshold p-value = 5×10^{-8} to determine genome-wide significance. We found twelve unique SNPs at four independent loci to be significantly associated with cytotoxicity of four distinct chemotherapeutics, all of which were not previously implicated in any other GWAS as they do not appear in the GWAS catalog (Table 2) (MacArthur et al. 2017).

We found two SNPs located in a noncoding region of chromosome four, rs61079639 ($p = 2.3 \times 10^{-9}$) and rs60507300 ($p = 2.3 \times 10^{-9}$), to be associated with daunorubicin cytotoxicity in the YRI population (Figure 3). We found three SNPs on chromosome nine, rs2100011 ($p = 4.7 \times 10^{-9}$), rs2254812 ($p = 4.7 \times 10^{-9}$), and rs2254813 ($p = 4.7 \times 10^{-9}$), to be associated with carboplatin

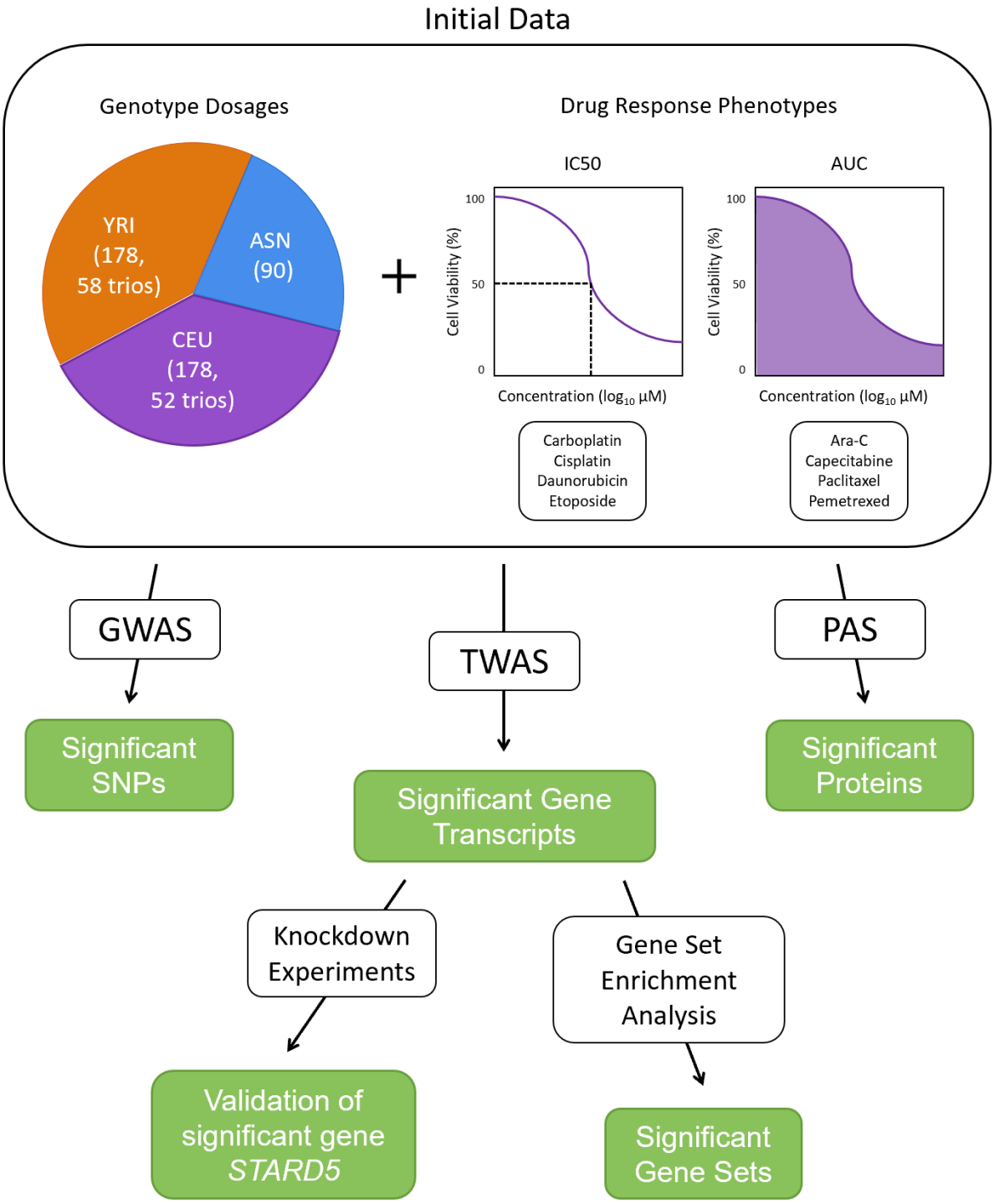

Figure 2. Overview of Analyses.

cytotoxicity in the ASN population (Figure 4). These SNPs are located in the gene *PPP1R26*; rs2100011 is an intron variant and rs2254812 and rs2254813 are 5' untranslated region variants. Additionally, we found six SNPs located in a noncoding region of chromosome twelve, led by rs7971310 ($p = 1.1 \times 10^{-8}$), to be associated with etoposide cytotoxicity in the YRI population (Table 2). Two of these SNPs, rs2711729 ($p = 4.9 \times 10^{-8}$), rs2711728 ($p = 4.9 \times 10^{-8}$), were also found to be associated with etoposide cytotoxicity in the ALL population (Figure 5). We found

Table 2. Genome-wide significant SNPs (Genome Build 37) from all GWAS performed.

Pop.	Drug	SNP	Chr.	Position	A1	A2	P-value	Beta
YRI	Daunorubicin	rs61079639	4	96611494	T	A	2.3×10^{-9}	0.79
YRI	Daunorubicin	rs60507300	4	96611493	T	G	2.3×10^{-9}	0.79
ASN	Carboplatin	rs2100011	9	138376145	A	G	4.7×10^{-9}	0.77
ASN	Carboplatin	rs2254812	9	138375872	C	G	4.7×10^{-9}	0.77
ASN	Carboplatin	rs2254813	9	138375861	G	A	4.7×10^{-9}	0.77
YRI	Etoposide	rs7971310	12	47428174	G	A	1.1×10^{-8}	-0.85
YRI	Etoposide	rs7960974	12	47424034	A	G	1.1×10^{-8}	-0.85
YRI	Etoposide	rs7979399	12	47424033	G	T	1.3×10^{-8}	-0.85
YRI	Etoposide	rs2711729	12	47409824	A	G	1.5×10^{-8}	0.88
YRI	Etoposide	rs2711728	12	47411926	C	A	1.5×10^{-8}	0.88
YRI	Etoposide	rs11183699	12	47426533	A	G	2.6×10^{-8}	-0.79
YRI	Cisplatin	rs10510241	3	2907097	A	G	4.7×10^{-8}	0.65
ALL	Etoposide	rs2711729	12	47409824	A	G	4.9×10^{-8}	0.80
ALL	Etoposide	rs2711728	12	47411926	C	A	4.9×10^{-8}	0.80

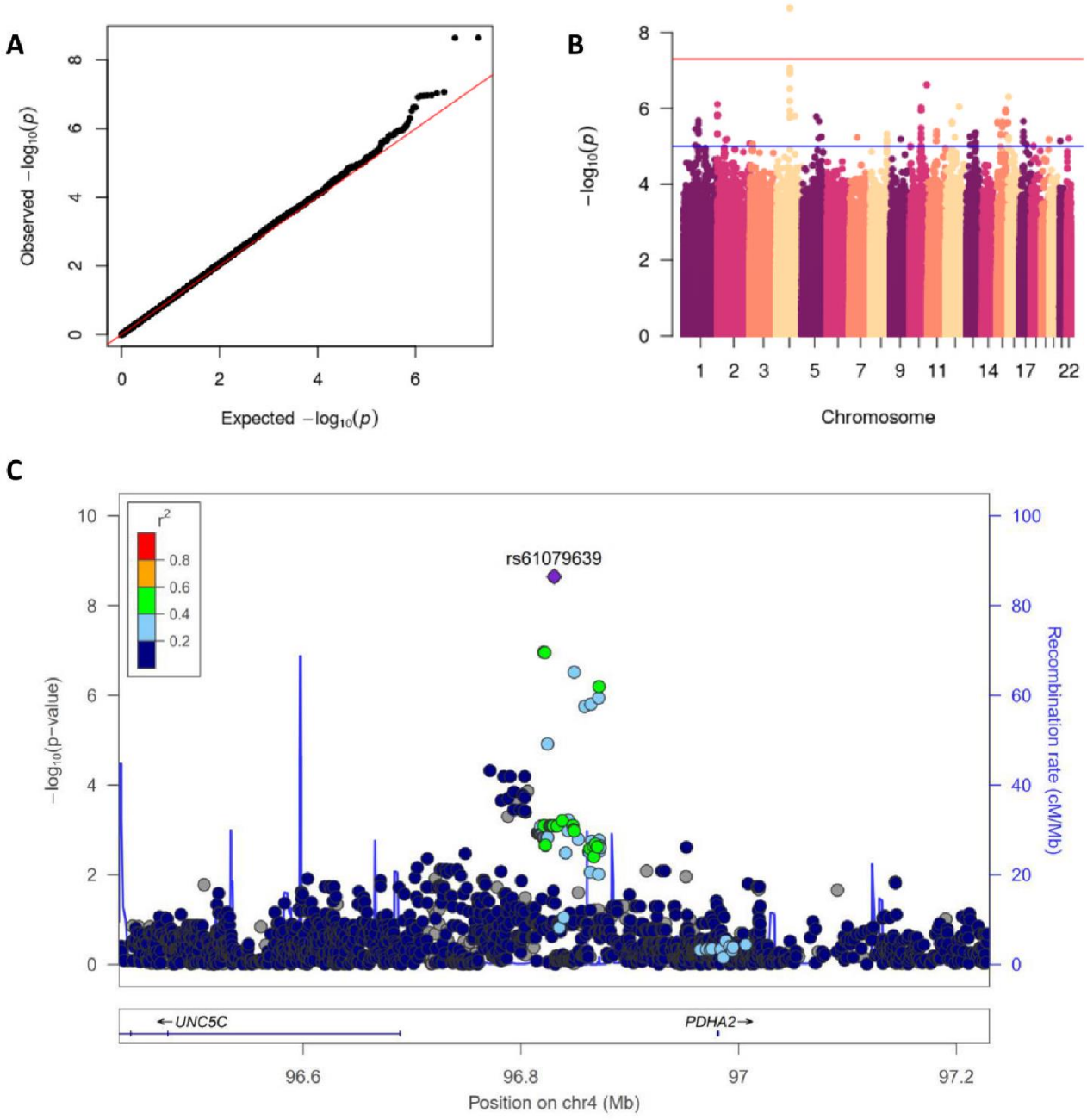

Figure 3. GWAS results for YRI and Daunorubicin cytotoxicity phenotype. (A) QQ plot of GWAS results showing expected vs observed p-values, red line at x=y. (B) Manhattan plot of GWAS results, red line at genome-wide significance threshold. (C) LocusZoom plot of rs61079639 ($p = 2.3 \times 10^{-9}$), the blue line measures the recombination rate at a certain position and each point is colored to indicate linkage disequilibrium (r^2) with rs61079639 in the 1000 Genomes Nov. 2014 AFR population.

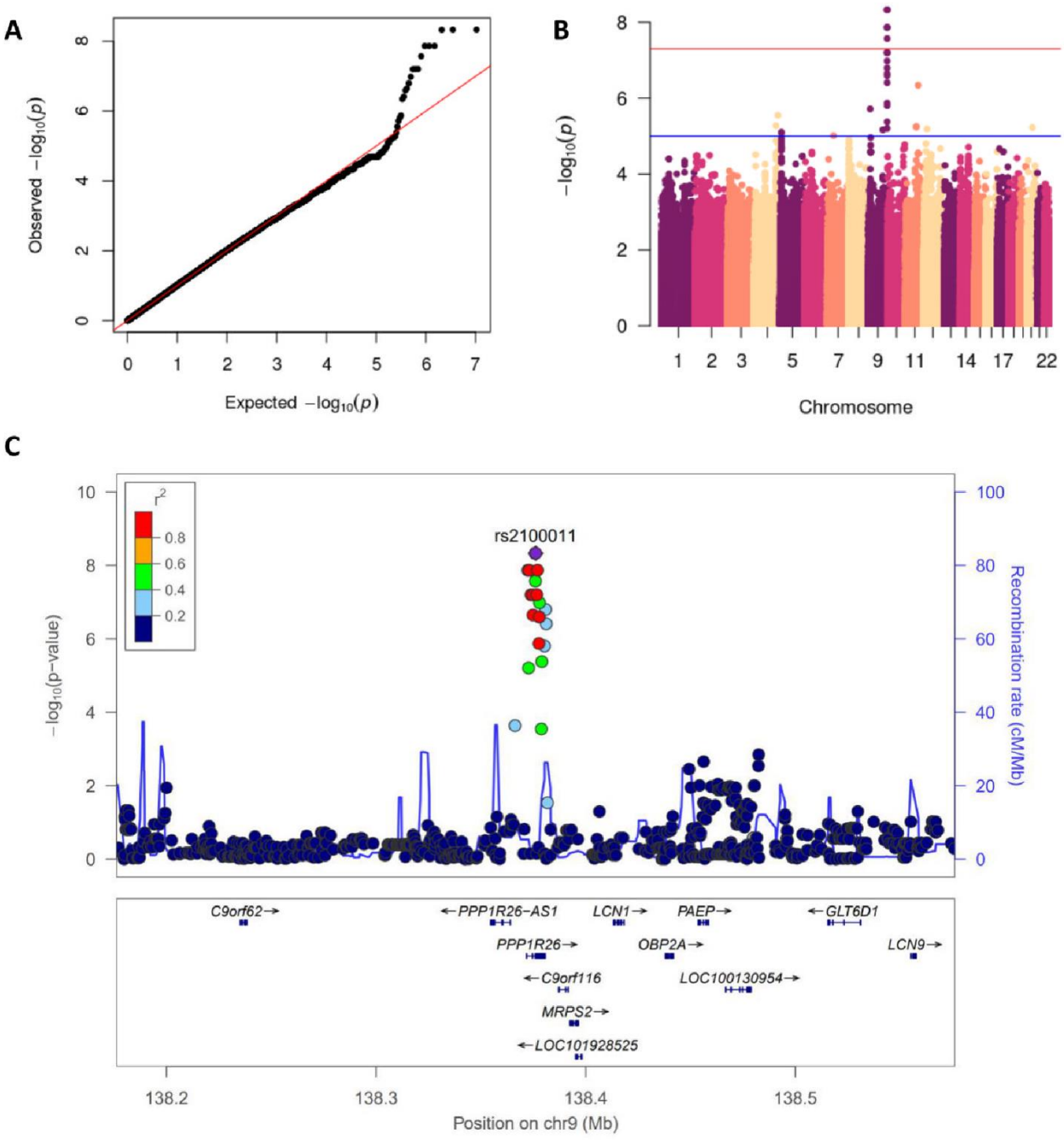

Figure 4. GWAS results for ASN and Carboplatin cytotoxicity phenotype. (A) QQ plot of GWAS results showing expected vs observed p-values, red line at x=y. (B) Manhattan plot of GWAS results, red line at genome-wide significance threshold. (C) LocusZoom plot of rs2100011 ($p = 4.7 \times 10^{-9}$), the blue line measures the recombination rate at a certain position and each point is colored to indicate linkage disequilibrium (r^2) with rs2100011 in the 1000 Genomes Nov. 2014 ASN population.

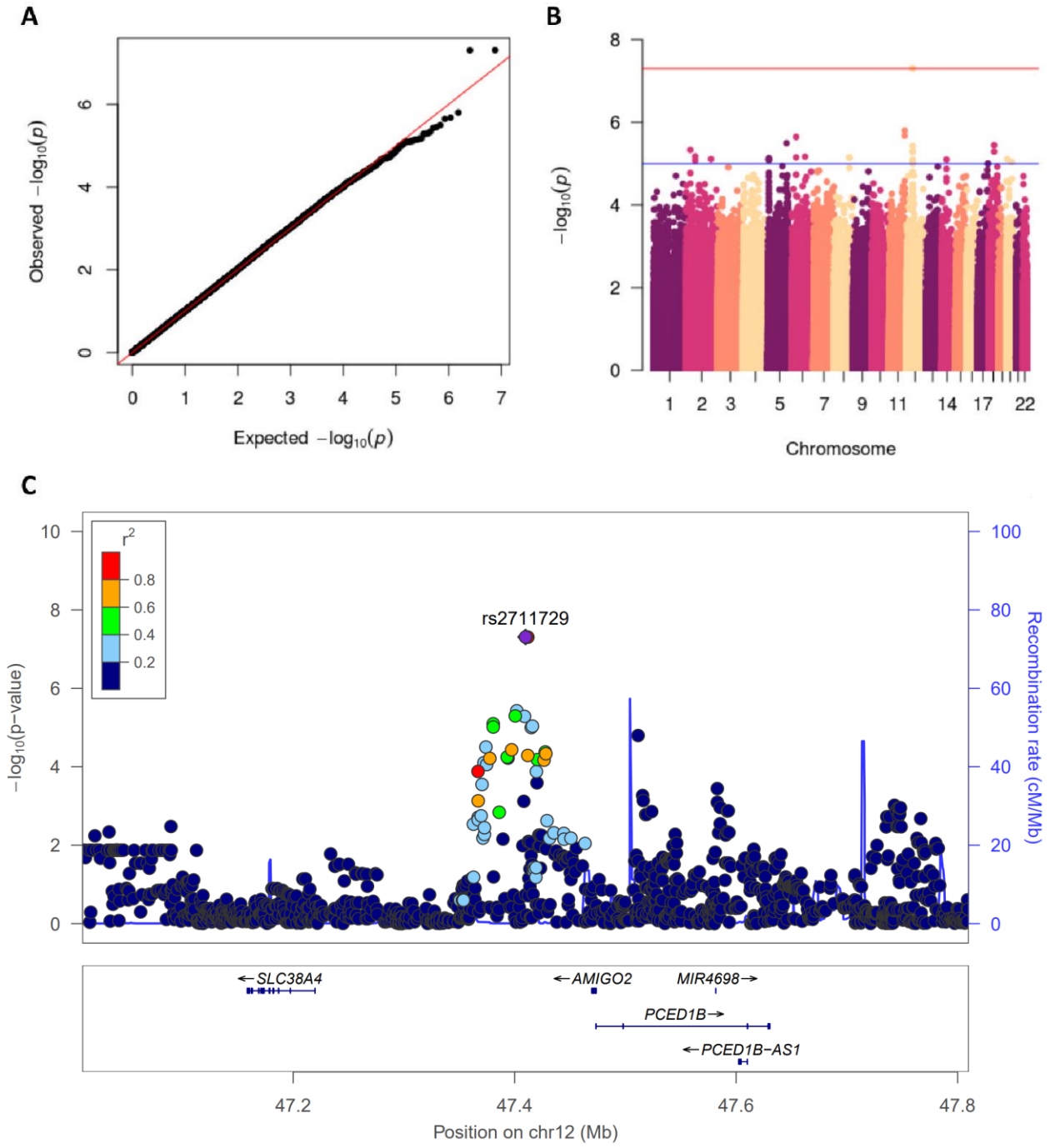

Figure 5. GWAS results for ALL and Etoposide cytotoxicity phenotype. (A) QQ plot of GWAS results showing expected vs observed p-values, red line at x=y. (B) Manhattan plot of GWAS results, red line at genome-wide significance threshold. (C) LocusZoom plot of rs2711729 ($p = 4.9 \times 10^{-8}$), the blue line measures the recombination rate at a certain position and each point is colored to indicate linkage disequilibrium (r^2) with rs2711729 in the 1000 Genomes Nov. 2014 AFR population.

one SNP located on chromosome three, rs10510241 ($p = 4.7 \times 10^{-8}$), to be associated with cisplatin cytotoxicity in the YRI population (Figure 6). This SNP is an intron variant in the gene *CNTN4*. No genome-wide significant associations were found for CEU. Through conditional analysis we found that the SNPs in each chromosomal region were not independent, thus each set of SNPs represents one association between the corresponding cytotoxicity phenotype and locus. None of the significant SNPs identified in one ancestral population replicated in another ancestral population (Table 3).

Table 3. Genome-wide significant SNP results (Genome Build 37) across populations from all GWAS performed. See Table 2 for chromosome, position, and alleles.

SNP	Drug	YRI P-value	YRI Beta	ASN P-value	ASN Beta	CEU P-value	CEU Beta	ALL P-value	ALL Beta
rs61079639	Daunorubicin	2.3×10^{-9}	0.79	N/A	N/A	0.59	0.98	3.6×10^{-6}	0.84
rs60507300	Daunorubicin	2.3×10^{-9}	0.79	N/A	N/A	0.59	0.98	3.6×10^{-6}	0.84
rs2100011	Carboplatin	0.29	0.66	4.7×10^{-9}	0.77	0.35	0.82	0.0096	0.66
rs2254812	Carboplatin	0.24	0.66	4.7×10^{-9}	0.77	0.35	0.82	0.012	0.66
rs2254813	Carboplatin	0.24	0.66	4.7×10^{-9}	0.77	0.35	0.82	0.012	0.66
rs7971310	Etoposide	1.1×10^{-8}	-0.85	N/A	N/A	0.62	0.95	4.6×10^{-5}	0.78
rs7960974	Etoposide	1.1×10^{-8}	-0.85	N/A	N/A	0.61	0.60	N/A	N/A
rs7979399	Etoposide	1.3×10^{-8}	-0.85	N/A	N/A	0.60	0.60	N/A	N/A
rs2711729	Etoposide	1.5×10^{-8}	0.88	N/A	N/A	0.17	0.07	4.9×10^{-8}	0.80
rs2711728	Etoposide	1.5×10^{-8}	0.88	N/A	N/A	0.17	0.07	4.9×10^{-8}	0.80
rs11183699	Etoposide	2.6×10^{-8}	-0.79	N/A	N/A	0.64	0.60	6.8×10^{-5}	0.77
rs10510241	Cisplatin	4.7×10^{-8}	0.65	0.94	0.22	0.94	0.70	7.5×10^{-4}	0.61

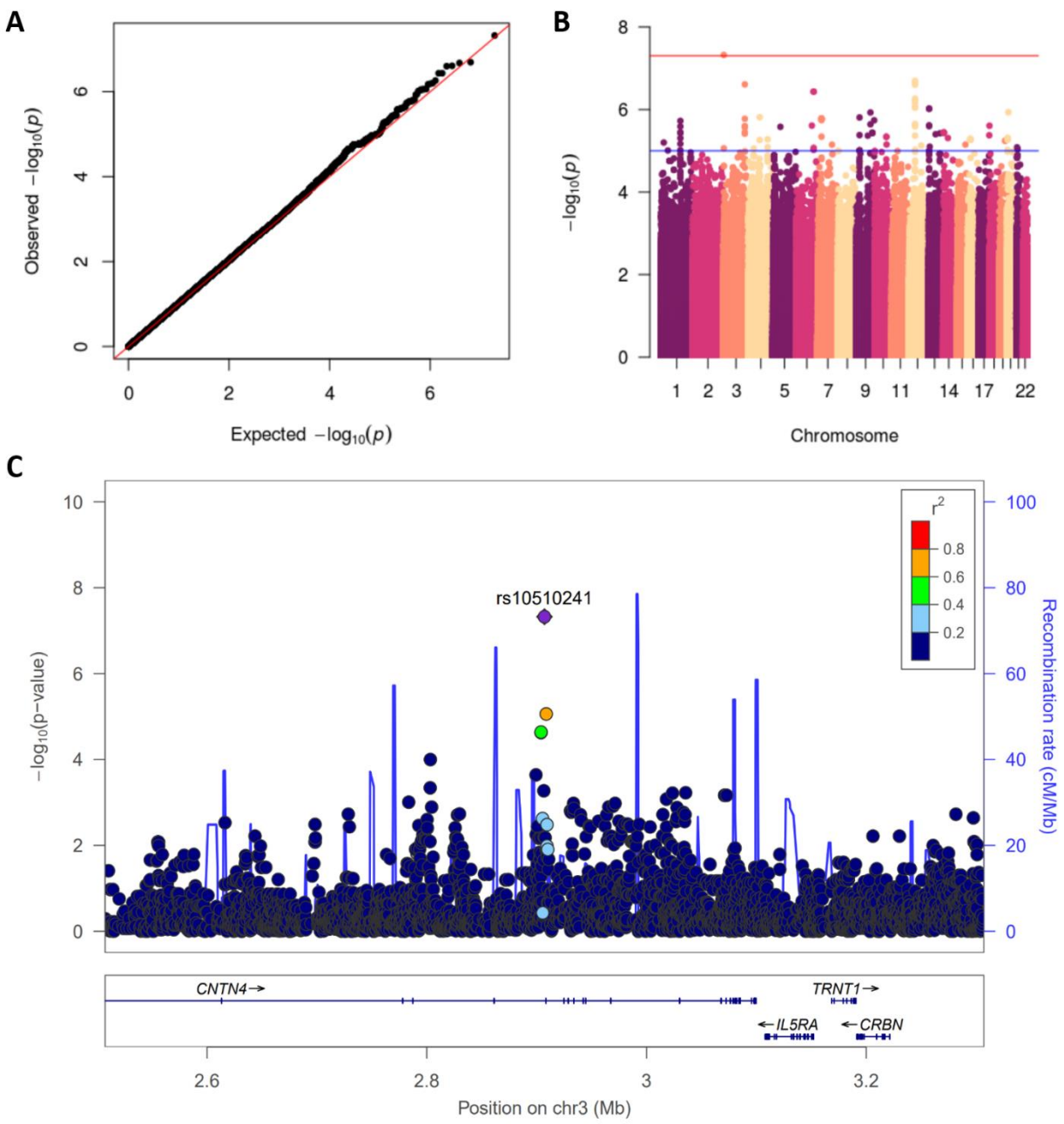

Figure 6. GWAS results for YRI and Cisplatin cytotoxicity phenotype. (A) QQ plot of GWAS results showing expected vs observed p-values, red line at x=y. (B) Manhattan plot of GWAS results, red line at genome-wide significance threshold. (C) LocusZoom plot of rs10510241 ($p = 4.7 \times 10^{-8}$), the blue line measures the recombination rate at a certain position and each point is colored to indicate linkage disequilibrium (r^2) with rs10510241 in the 1000 Genomes Nov. 2014 AFR population.

TWAS predict expression of three genes are associated with chemotherapy-induced cytotoxicity

Following GWAS, we conducted TWAS using both PrediXcan and MulTiXcan to identify significant associations between predicted gene expression levels and the cytotoxicity of each drug for each ancestral population (Gamazon et al. 2015; Barbeira et al. 2019). PrediXcan and MulTiXcan utilize prediction models to calculate predicted expression levels for various genes and identify associations between predicted gene expression levels and phenotype (Gamazon et al. 2015; Barbeira et al. 2019). Both PrediXcan and MulTiXcan calculate predicted gene expression levels for each gene using each model individually, but while PrediXcan then finds model-specific associations between predicted gene expression and phenotype, MulTiXcan aggregates expression to find overall associations and identifies models with the best and worst performance (Gamazon et al. 2015; Barbeira et al. 2019). We used the 48 GTEx version 7 tissue-based prediction models, which each contain approximately 10,000 genes, to run PrediXcan and MulTiXcan (Gamazon et al. 2015; Barbeira et al. 2019). Additionally, for PrediXcan only, we used the 5 MESA population-based prediction models, which each contain approximately 8,000 genes (Mogil et al. 2018). To obtain the PrediXcan results, we used PrediXcan to calculate the predicted gene expression levels and GEMMA to conduct the association tests, as this accounted for relatedness within each ancestral population (Gamazon et al. 2015; Zhou and Stephens 2012). To obtain the MulTiXcan results, we used the same predicted gene expression levels and conducted the association tests with MulTiXcan, as this produced aggregate associations (Barbeira et al. 2019). For the ALL population, we accounted for population stratification with the same covariates as in GWAS.

We found three significant associations (Bonferroni adjusted p-value < 0.05) between gene expression and cytotoxicity, two from PrediXcan and one from MulTiXcan. Using PrediXcan, we determined increased predicted expression of *STARD5* in the brain cortex tissue to be associated with a decrease in the concentration of etoposide required for cytotoxicity (IC_{50}) in the ALL population ($p = 8.5 \times 10^{-8}$) (Figure 7A). Additional results for the YRI population, etoposide phenotype, and *STARD5* derived from other GTEx version 7 and MESA models can be seen in Table 4. We also found increased predicted expression of *USF1* in the liver tissue to

Table 4. *STARD5* results for the ALL population and Etoposide cytotoxicity phenotype derived from GTEx version 7 and MESA models.

Model	P-value	Adj. P	Beta
Brain Cortex	8.5×10^{-8}	0.023	-1.1
MESA AFHI	9.1×10^{-5}	1.00	-2.1
MESA HIS	3.4×10^{-4}	1.00	-1.0
Esophagus Mucosa	3.6×10^{-3}	1.00	-0.66
MESA ALL	4.9×10^{-3}	1.00	-0.63
Stomach	4.9×10^{-3}	1.00	-2.8
Esophagus Muscularis	0.018	1.00	-0.92
Skin Sun Exposed Lower leg	0.073	1.00	-0.51
MESA CAU	0.098	1.00	-0.65
Testis	0.11	1.00	0.70
Artery Tibial	0.14	1.00	1.6
Brain Hippocampus	0.18	1.00	-0.24
Esophagus Gastroesophageal Junction	0.19	1.00	-0.20
Lung	0.23	1.00	-1.6
Muscle Skeletal	0.33	1.00	2.4
Nerve Tibial	0.33	1.00	-0.69
Brain Frontal Cortex	0.71	1.00	-4.4
Colon Sigmoid	0.73	1.00	0.15
Skin Not Sun Exposed Suprapubic	0.91	1.00	0.034
Cells Transformed fibroblasts	1.00	1.00	0.010

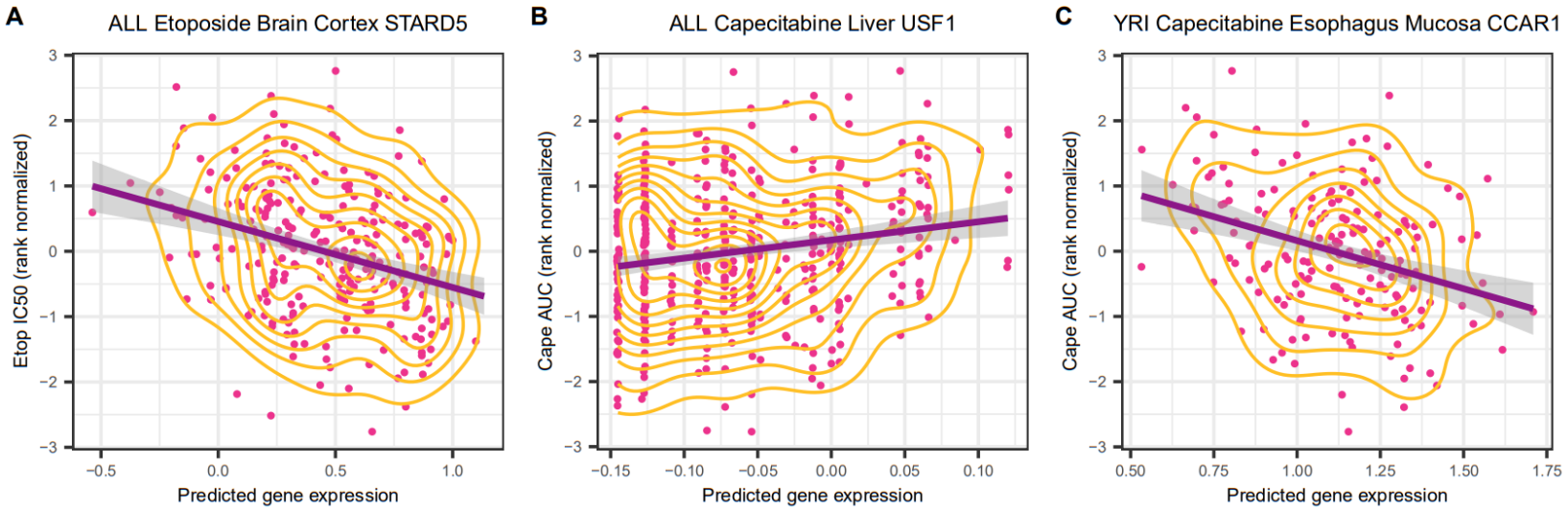

Figure 7. Predicted Expression of significant TWAS gene hits versus measured drug cytotoxicity levels. (A) Predicted expression of *STARD5* in the ALL population as determined by PrediXcan using the GTEx v7 Brain Cortex prediction model plotted against rank-normalized Etoposide IC_{50} levels as measured in LCLs from the ALL population. (B) Predicted expression of *USF1* in the ALL population as determined by PrediXcan using the GTEx v7 Liver prediction model plotted against rank-normalized Capecitabine AUC levels as measured in LCLs from the ALL population. (C) Predicted expression of *CCAR1* in the YRI population as determined by MulTiXcan plotted against rank-normalized Capecitabine AUC levels as measured in LCLs from the YRI population. *CCAR1* expression was best predicted by the GTEx v7 Esophagus Mucosa prediction model. Each point represents an individual, the curved yellow lines convey density in regard to the distribution of the points, and the purple line is the best fit determined by linear regression, which shows the direction of effect.

be associated with an increase in the concentration of capecitabine required for cytotoxicity (AUC) in the ALL population ($p = 8.7 \times 10^{-8}$) (Figure 7B). Using MulTiXcan, we found increased predicted expression of *CCAR1* to be associated with a decrease in the concentration of capecitabine required for cytotoxicity (AUC) in the YRI population ($p = 4.2 \times 10^{-6}$) (Figure 7C).

FUMA identifies enrichment in oncogenic signatures

We performed FUMA gene set enrichment analysis on top PrediXcan results for each ancestral population and drug and found twelve significant gene sets (Table 5) (Watanabe et al. 2017). For the CEU population and cisplatin, we identified one significant gene set WNT_UP.V1_ UP ($p = 1.2 \times 10^{-5}$). This gene set is an oncogenic signature, denoting up-regulation of the listed genes as a result of the over-expression of *WNT1* in mammary epithelial cells (Ziegler et al. 2005). The genes making up this set were all found to have predicted expression levels associated with cisplatin IC_{50}. Cisplatin is often used to treat a variety of cancers, including lung, colon, testicular, and ovarian cancers (Trendowski, El-Charif, et al. 2019; Trendowski, El Charif, et al. 2019). Additionally, for the CEU population and cytarabine arabinoside (ara-C), we identified the gene set P53_DN.V1_DN to be significant ($p = 1.1 \times 10^{-4}$). This is another oncogenic signature, characterized by down-regulation of the genes listed in cancer cell lines with mutated *TP53* from the NCI-60 collection (A. Subramanian et al. 2005). The genes in the set are impacted by mutations in *TP53*, a known tumor suppressor gene that, when mutated, can lead to malignancy (A. Subramanian et al. 2005). The predicted expression levels of these genes are associated with ara-C AUC.

We also performed FUMA gene set enrichment analysis on top MulTiXcan results for each ancestral population and drug, which identified fifteen significant gene sets (Table 6). For the YRI cohort and Daunorubicin, four gene sets, classified as cancer gene neighborhoods, were

Table 5. Significant Gene Sets from FUMA tool GENE2FUNC generated using top genes from PrediXcan results.

Pop.	Drug	Category	Gene Set	N	n	P-value	Adj. P	Genes
CEU	Cisplatin	Oncogenic Signatures	WNT_UP.V1_UP	170	7	1.2×10^{-5}	0.0023	*VAMP1, RPAP3, LTB4R, SERPINF1, AP2S1, POMC, HS3ST1*
ASN	Capecitabine	microRNA Targets (MsigDB c3)	CCCACAT_MIR2993P	48	4	1.7×10^{-5}	0.0038	*RAB6A, ITGAV, ABCE1, TRPM3*
CEU	Capecitabine	Hallmark Gene Sets (MsigDB h)	HALLMARK_PEROXISOME	100	5	1.1×10^{-4}	0.0053	*PRDX5, RETSAT, ABCC5, SEMA3C, GSTK1*
CEU	Paclitaxel	GWAS Catalog Reported Genes	Liver enzyme levels (gamma-glutamyl transferase)	42	4	8.9×10^{-6}	0.015	*GSTT2B, DDTL, KB-226F1.2, DDT, GGT1*
ALL	Carboplatin	Chemical and Genetic Pertubation Gene Sets	NIKOLSKY_BREAST_CANCER_17Q21_Q25_AMPLICON	318	9	4.6×10^{-6}	0.016	*PDK2, CACNA1G, SCPEP1, COG1, FAM104A, C17orf80, BTBD17, GPRC5C, SLC16A3*
CEU	Paclitaxel	Hallmark Gene Sets (MsigDB h)	HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION	190	5	3.2×10^{-4}	0.016	*VCAM1, COL1A1, MATN3, CXCL1, ECM2*
CEU	Ara-C	Oncogenic Signatures	P53_DN.V1_DN	179	6	1.1×10^{-4}	0.020	*AJAP1, KCNAB2, GPRC5B, HOXB2, CBX4, DFNA5*
ASN	Ara-C	Chemical and Genetic Pertubation Gene Sets	SOTIRIOU_BREAST_CANCER_GRADE_1_VS_3_DN	51	5	6.5×10^{-6}	0.022	*PIGV, BBS1, TUBGCP4, SNX1, CRTC3*
ASN	Ara-C	Chemical and Genetic Pertubation Gene Sets	NIKOLSKY_BREAST_CANCER_11Q12_Q14_AMPLICON	153	7	1.5×10^{-5}	0.025	*BBS1, ZDHHC24, CCS, LRFN4, RAD9A, NDUFV1, MTL5*
YRI	Ara-C	Immunological Signatures (MsigDB c7)	GSE39110_DAY3_VS_DAY6_POST_IMMUNIZATION_CD8_TCELL_UP	190	8	5.2×10^{-6}	0.025	*RRP12, TRMT112, ACAT1, BTG1, EVL, MPPE1, FAM161A, MAPK11*
ALL	Daunorubicin	Chemical and Genetic Pertubation Gene Sets	RICKMAN_TUMOR_DIFFERENTIATED_WELL_VS_POORLY_UP	219	8	8.8×10^{-6}	0.030	*LMO4, TRAF3IP3, BCL2L11, ABHD12, IFT122, MSL2, VARS2, CASC7, AGO2*
ASN	Cisplatin	Cancer Gene Modules (MsigDB c4)	MODULE_372	21	3	7.8×10^{-5}	0.034	*ABCC4, TWSG1, CCNE2*

Table 6. Significant Gene Sets from FUMA tool GENE2FUNC generated using top genes from MulTiXcan results.

Pop.	Drug	Category	Gene Set	N	n	P-value	Adj. P	Genes
YRI	Carboplatin	GO Cellular Components (MsigDB c5)	GO_COMPACT_MYELIN	15	4	7.4×10^{-6}	0.0043	*NCMAP, CD59, MPP5, PLLP*
CEU	Daunorubicin	GO Cellular Components (MsigDB c5)	GO_CYTOPLASMIC_DYNEIN_COMPLEX	14	5	9.3×10^{-6}	0.0054	*TPR, DYNLL1, BCL2L11, DYNC1LI1, DCTN4*
CEU	Ara-C	KEGG (MsigDB c2)	KEGG_PENTOSE_PHOSPHATE_PATHWAY	22	4	3.7×10^{-5}	0.0068	*H6PD, PFKM, TKT, TKTL2*
CEU	Carboplatin	Immunological Signatures (MsigDB c7)	GSE4142_GC_BCELL_VS_MEMORY_BCELL_DN	189	9	5.8×10^{-6}	0.017	*STX6, AMPD3, ALOX15B, PIGL, ASAP2, HACL1, ZNF827, UNC5CL, C9orf64*
CEU	Carboplatin	Immunological Signatures (MsigDB c7)	GSE17721_CPG_VS_GARDIQUIMOD_8H_BMDC_DN	193	9	6.9×10^{-6}	0.017	*LRP8, SLK, AMPD3, EEF1G, EMC7, NDRG4, CTDNEP1, LRRC16A, QKI*
YRI	Daunorubicin	Cancer Gene Neighborhoods (MsigDB c4)	GCM_TPT1	66	6	5.3×10^{-5}	0.023	*RPL27A, RPS3, NDUFA12, NPM1, RPS18, RPS10*
ALL	Cisplatin	BioCarta (MsigDB c2)	BIOCARTA_MCM_PATHWAY	18	3	1.2×10^{-4}	0.026	*ORC1, CDC6, MCM6*
YRI	Daunorubicin	Cancer Gene Neighborhoods (MsigDB c4)	GNF2_EIF3S6	113	7	1.5×10^{-4}	0.032	*PNRC2, RPL27A, RPS3, EIF3D, NPM1, RPS18, RPS10*
CEU	Ara-C	WikiPathways	Pathways in clear cell renal cell carcinoma%WikiPathways_20190110%WP4018%Homo sapiens	79	6	6.8×10^{-5}	0.033	*ARNT, TGFB2, TPI1, PFKM, MDH1, TSC1*
YRI	Daunorubicin	Cancer Gene Neighborhoods (MsigDB c4)	MORF_ACTG1	126	7	2.9×10^{-4}	0.034	*TAGLN2, RPL27A, ZFPL1, RPS3, NPM1, RPS18, RPS10*
YRI	Daunorubicin	Cancer Gene Neighborhoods (MsigDB c4)	MORF_TPT1	91	6	3.2×10^{-4}	0.034	*RPL27A, ZFPL1, RPS3, NPM1, RPS18, RPS10*
YRI	Cisplatin	microRNA Targets (MsigDB c3)	ACCAATC_MIR509	43	5	2.6×10^{-4}	0.035	*PCDHA2, PCDHA3, PCDHA4, PCDHA5, ZFAND3*
YRI	Cisplatin	microRNA Targets (MsigDB c3)	GTAGGCA_MIR189	25	4	3.1×10^{-4}	0.035	*CAPRIN1, MBLAC2, SRPK2, MTSS1*
ALL	Capecitabine	GO Cellular Components (MsigDB c5)	GO_BASAL_PLASMA_MEMBRANE	32	4	6.4×10^{-5}	0.037	*SLC27A5, PKD2, ERBB2IP, CAV1*
ALL	Cisplatin	Reactome (MsigDB c2)	REACTOME_G2_M_CHECKPOINTS	41	4	7.1×10^{-5}	0.048	*ORC1, ATM, CDC6, MCM6*

identified: GCM_TPT1 (p = 5.33e-05), GNF2_EIF3S6 (p = 1.49e-04), MORF_ACTG1 (p = 2.92e-04), and MORF_TPT1 (p = 3.18e-04). Cancer gene neighborhoods develop as a result of mutations in multiple genes in an area of the genome and are common to some cancer types, including leukemia. One gene in all four of these sets, RPS18, has been found to be highly expressed acute lymphoblastic leukemia. Another gene, NPM1, which is also included in each of these sets, has been found to be upregulated in both acute myeloid and lymphoblastic leukemia. Daunorubicin is used to treat various subtypes of leukemia, including acute myeloid and lymphoblastic leukemia, thus it is interesting that the predicted expression levels for the genes making up these neighborhoods were identified by MulTiXcan to be associated with Daunorubicin IC_{50}.

Knockdown experiments validate reduced *STARD5* expression is associated with reduced etoposide-induced cytotoxicity

After conducting GWAS and TWAS, we followed up on our results by performing functional experiments for *STARD5*, as this gene had the most significant predicted expression levels from the TWAS results. The predicted expression plot for *STARD5* showed a negative correlation between *STARD5* predicted expression and etoposide IC_{50}. Therefore, for our functional experiments, we hypothesized that the knockdown of *STARD5* expression levels would result in a higher etoposide IC_{50}, which corresponds to lower cellular sensitivity to etoposide. We selected the lung cancer cell line A549 for the knockdown experiments, as etoposide is often used to treat lung cancer (Qiu et al. 2019).

After knocking down *STARD5* with siRNA, we treated A549 cells with increasing concentrations of etoposide and then measured relative viability at 72 and 96 hours after treatment (Figure 8A). siRNA reduced *STARD5* expression to less than 25% of control at 0, 72,

and 96 hours (Figure 8B). At both 72 and 96 hours, reduced *STARD5* expression significantly increased cell viability (Figure 8C-D, p = 0.034 for 72 hours, p = 0.0001 for 96 hours), validating our TWAS results that higher expression of *STARD5* is correlated with greater sensitivity to etoposide.

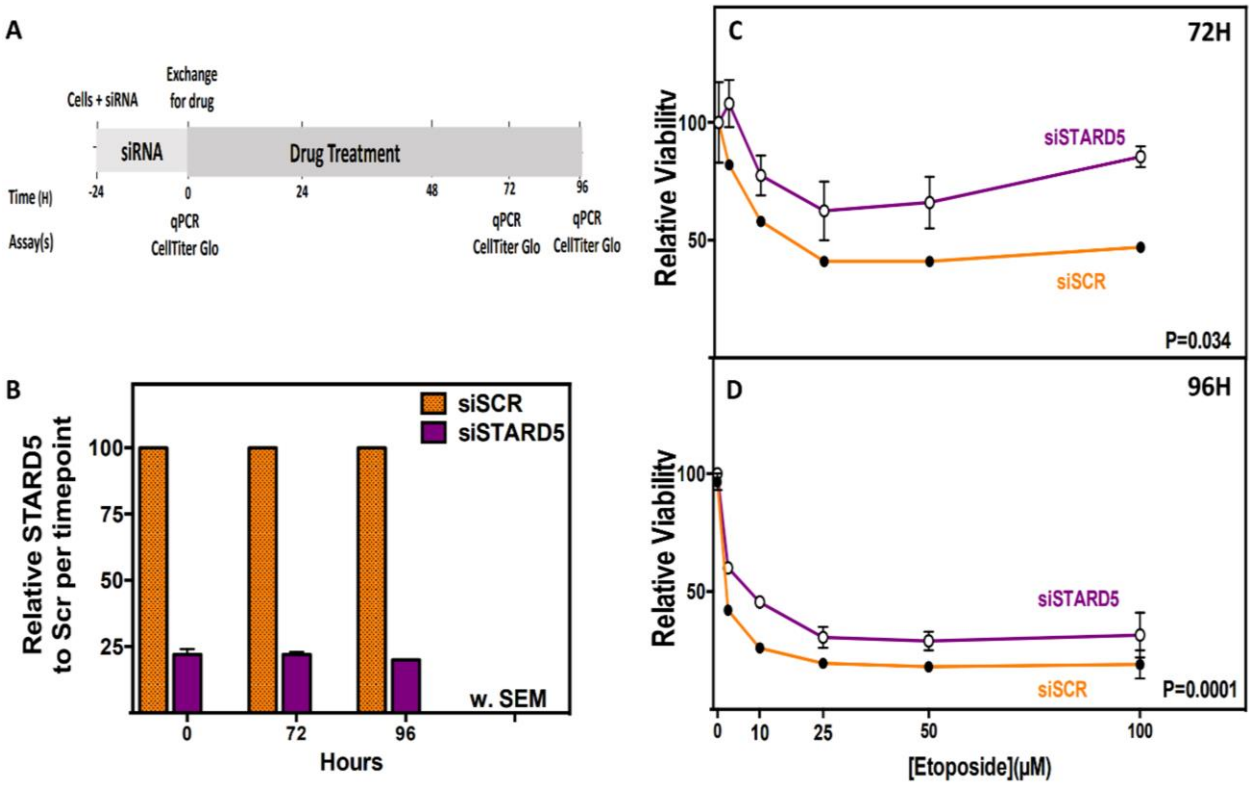

Figure 8. Evaluation of the effect of *STARD5* knockdown on sensitivity of A549 lung cancer cells to etoposide. (A) Experimental scheme for knockdown of *STARD5* in A549 and treatment with etoposide. (B) *STARD5* expression was reduced < 25% for cells treated with siSTARD5 (gray bars) compared to expression in siSCR (black bars) at time of drug treatment (0H) and at 72 and 96 hours as determined by quantitative reverse transcription PCR (qRT-PCR). Relative viability, determined by CellTiter-Glo 2.0 assay, for A549 cells treated with increasing concentrations of etoposide at (C) 72 hours and (D) 96 hours after treatment with siSTARD5 (open circle) or siSCR control (closed circle). Data represents two independent experiments including at least three replicates analyzed by two-way ANOVA showing the SEM.

PAS predict seven unique proteins to be significantly associated with chemotherapy-induced cytotoxicity

In addition to GWAS and TWAS, we conducted PAS to identify significant associations between predicted protein levels and the cytotoxicity of each drug for each ancestral population. We first predicted protein levels with PrediXcan using the TOPMed prediction models and we then used GEMMA to perform the association tests, in order to account for relatedness within each population. We found seven unique proteins with predicted levels significantly associated with chemotherapy-induced cytotoxicity (Bonferroni adjusted p-value < 0.05) in three of the four populations (Table 6). In the ASN population, the most significant association identified was found with the TOPMed EUR model between increased predicted levels of the protein encoded by *NAGK* and increased cisplatin concentration required for cytotoxicity (Figure 9A). In the ALL population, the most significant association identified was found with the TOPMed ALL-M model between increased predicted levels of the protein encoded by *HK2* and decreased daunorubicin concentration required for cytotoxicity (Figure 9B). In the YRI population, the

Table 7. Significant predicted protein levels from all PAS performed.

Pop.	Drug	Model	Protein-coding Gene	Chr.	P-value	Adj. P	Beta
ASN	Cisplatin	TOPMed EUR	*NAGK*	2	1.2×10^{-4}	0.0065	2.4
ALL	Daunorubicin	TOPMed ALL-M	*HK2*	2	1.0×10^{-4}	0.015	-3.3
ALL	Pemetrexed	TOPMed CHN	*IL17RD*	3	5.9×10^{-4}	0.015	-4.5
ALL	Ara-C	TOPMed EUR	*DPT*	1	1.8×10^{-4}	0.016	1.6
YRI	Pemetrexed	TOPMed CHN	*IL17RD*	3	1.3×10^{-3}	0.036	-5.2
ALL	Daunorubicin	TOPMed ALL-M	*EGF*	4	2.6×10^{-4}	0.038	-1.3
ALL	Ara-C	TOPMed AFA	*IL5RA*	3	7.2×10^{-4}	0.039	3.9
YRI	Pemetrexed	TOPMed HIS	*PDE5A*	4	4.3×10^{-4}	0.042	1.8

most significant association identified was found with the TOPMed CHN model between increased predicted levels of the protein encoded by *IL17RD* and decreased pemetrexed concentration required for cytotoxicity.

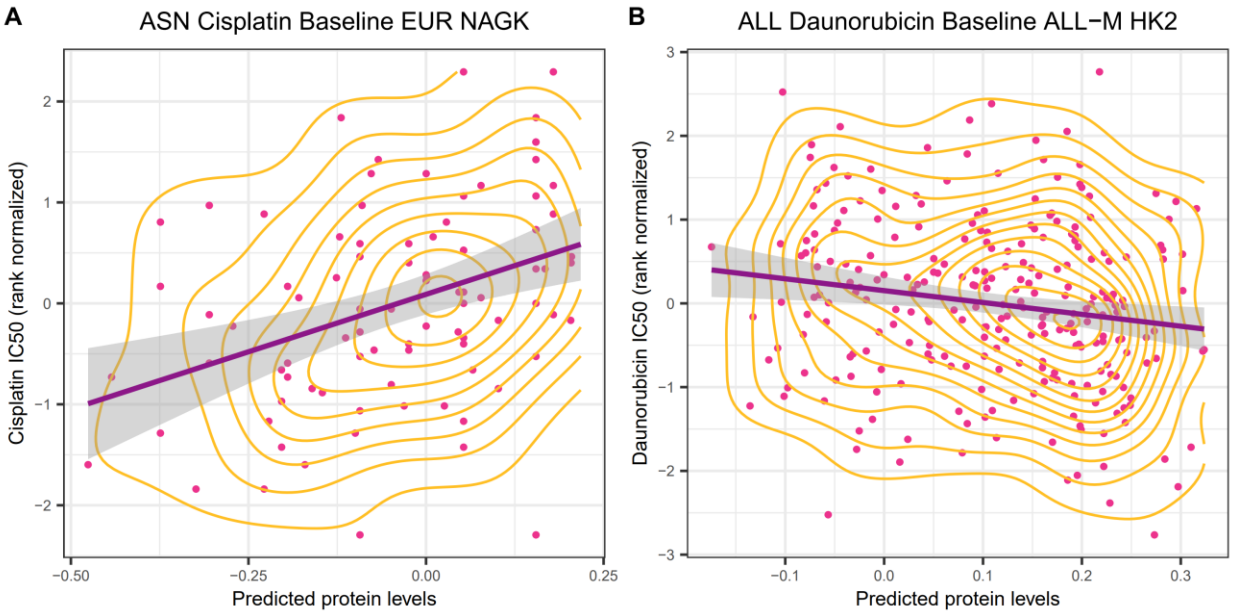

Figure 9. Predicted protein levels of significant PAS hits versus measured drug cytotoxicity levels. (A) Predicted levels of the protein encoded by *NAGK* in the ASN population as determined by PrediXcan using the TOPMed EUR prediction model plotted against rank-normalized Cisplatin IC_{50} levels as measured in LCLs from the ASN population. (B) Predicted levels of the protein encoded by *HK2* in the ALL population as determined by PrediXcan using the TOPMed ALL-M prediction model plotted against rank-normalized Daunorubicin IC_{50} levels as measured in LCLs from the ALL population. Each point represents an individual, the curved yellow lines convey density in regard to the distribution of the points, and the purple line is the best fit determined by linear regression, which shows the direction of effect.

CHAPTER FOUR

DISCUSSION AND CONCLUSION

Publication disclaimer

This work was previously published in Human Molecular Genetics (2021) doi.org/10.1093/hmg/ddab029 with the following authors:

Ashley J. Mulford[1,2], Claudia Wing[3], M. Eileen Dolan[3], Heather E. Wheeler[1,2]

[1]Department of Biology, Loyola University Chicago, Chicago, IL, USA, [2]Program in Bioinformatics, Loyola University Chicago, Chicago, IL, USA, [3]Section of Hematology/Oncology, Department of Medicine, University of Chicago, Chicago, IL, USA

We conducted GWAS, TWAS and PAS for eight chemotherapeutic cytotoxicity phenotypes measured in LCLs from individuals in three ancestral populations (YRI, CEU, and ASN) and one combined population (ALL). We identified twelve SNPs at four unique loci, three genes, and seven proteins significantly associated with chemotherapy-induced cytotoxicity. For the most significant gene, *STARD5*, we performed knockdown experiments to follow up on our finding that increased *STARD5* expression associates with decreased etoposide IC_{50}. These functional experiments validated this result, as knockdown of *STARD5* increased viability of A549 lung cancer cell lines treated with etoposide, demonstrating the positive correlation between *STARD5* expression and cellular sensitivity to etoposide.

The TWAS we conducted identified an association between increased predicted expression of *STARD5* and decreased etoposide IC_{50}, implying a greater cellular sensitivity to etoposide. This finding was then validated through the knockdown experiments we performed,

which demonstrated that a reduction of *STARD5* expression to twenty-five percent that of unaltered expression results in increased viability in A549 lung cancer cell lines treated with etoposide. Etoposide is a chemotherapeutic and antineoplastic drug that targets topoisomerase II, an enzyme that plays an essential role in DNA replication, recombination, and transcription, by cutting and pasting double-stranded DNA (Hande 1998). By interfering in topoisomerase II function in malignant cells, etoposide disrupts necessary biological processes, leading to an increase in DNA breakage that ultimately induces apoptosis (Hande 1998). Etoposide is commonly used to treat lung cancer; this informed our selection of the A549 lung cancer cell line for use in the knockdown experiments to test how etoposide IC_{50} would be impacted by a reduction in *STARD5* expression (Zucchetti et al. 1995). Additionally, previous projects have used A549 cell lines to study factors contributing to etoposide-induced cell death (Litwiniec et al. 2013; Y. Huang et al. 1997).

STARD5 encodes a steroidogenic acute regulatory related lipid transfer domain protein (Rodriguez-Agudo et al. 2005). Studies have found *STARD5* to become more highly expressed as a response to endoplasmic reticulum (ER) stress, which leads to the relocation of the protein encoded by *STARD5* from the nucleus to the cytosol and cell membrane (Rodriguez-Agudo et al. 2012). Etoposide, while disrupting normal topoisomerase II function, often induces ER stress in the process (C. Wang et al. 2016). This could contribute to increased *STARD5* expression in cancer cells. Additionally, increased *STARD5* expression in hepatocytes has been linked to increased cholesterol levels (Rodriguez-Agudo et al. 2005). *STARD5* protein binds and transports cholesterol and other sterol-derived molecules in the liver and thus helps regulate lipid homeostasis and metabolism (Rodriguez-Agudo et al. 2005). The mechanisms for cholesterol

homeostasis and drug metabolism have been found to rely on the same cellular receptors, including pregnane X receptor (PXR) (Rezen et al. 2011). PXR binds etoposide as well as other chemotherapeutics to activate CYP3A4, a key enzyme involved in drug metabolism (Schuetz et al. 2002). The role of *STARD5* in regulating metabolism and other liver functions could be one explanation for the association between etoposide-induced cytotoxicity and increased *STARD5* expression. Etoposide metabolism occurs primarily in the liver, where *STARD5* is highly expressed (Kawashiro et al. 1998; Rodriguez-Agudo et al. 2005). Overall, increased expression of *STARD5*, whether preexisting or prompted by ER stress, may facilitate etoposide metabolism in the liver, in turn promoting etoposide-induced cytotoxicity.

The GWAS we conducted revealed four unique loci associated with cellular sensitivity to either carboplatin, cisplatin, daunorubicin, or etoposide. In the ASN population, we found three SNPs on chromosome 9 located within *PPP1R26* to be associated with carboplatin-induced toxicity. *PPP1R26* has been associated with tumor formation and is upregulated in breast carcinomas, promoting metastasis through the degradation of retinoblastoma protein, a tumor suppressor protein (Zheng et al. 2018; Yang et al. 2005). In the YRI population, we found one SNP on chromosome 3 located within *CNTN4* to be associated with cisplatin-induced toxicity. *CNTN4* encodes a contactin 4, an immunoglobulin that regulates cellular interactions and axonal growth in the nervous system (Garcia et al. 2020; Evenepoel et al. 2018). Overexpression of *CNTN4* has been found to be associated with malignancy in nerve tissue and with cisplatin-induced nephrotoxicity (Garcia et al. 2020; Evenepoel et al. 2018). In the ALL population, we found two SNPs on chromosome 12 in proximity to *AMIGO2* to be associated with etoposide-induced toxicity. *AMIGO2* is a scaffold protein that binds to *PDK1* to regulate the phosphoinositide 3-kinase–Akt signaling pathway, which plays a role in many biological

mechanisms, including cell proliferation and metabolism (H. Park et al. 2015). Overexpression of *AMIGO2* has been found to induce abnormal Akt signaling, which contributes to the onset and progression of various cancers (H. Park et al. 2015). Additionally, *AMIGO2* overexpression is a common characteristic of metastatic tissue, particularly when metastasis occurs in the liver, as *AMIGO2* regulates cell adhesion in liver cells (Kanda et al. 2017).

The PAS we conducted identified seven unique proteins associated with cellular sensitivity to either ara-C, cisplatin, daunorubicin, or pemetrexed. In the ASN population we found N-Acetylglucosamine kinase, encoded by *NAGK*, to be significantly associated with cisplatin cytotoxicity. N-Acetylglucosamine kinase is known to regulate the Wnt signaling pathway, which is involved in metabolism and cell growth and proliferation (Neitzel et al. 2019). In the ALL population we found Hexokinase II, encoded by *HK2*, to be significantly associated with daunorubicin cytotoxicity. Hexokinase II catalyzes the first step in glycolysis and the upregulation of *HK2* in cancer cells has been found to increase the rate of glucose metabolism, aiding in cell growth and inhibiting apoptosis (Rai et al. 2019). Hexokinase II has been implicated in several previous cancer studies and has also been used as a target for some recently developed anticancer therapeutics (Nakajima et al. 2019; S.-J. Wang et al. 2021). Additionally, the inhibition of Hexokinase II has been found to increase cellular sensitivity to daunorubicin in myeloid leukemia cells, as this diminishes the protective effects of Hexokinase II against apoptosis, increasing the likelihood of drug-induced cytotoxicity (Rai et al. 2019). In the ALL population we also identified interleukin-17 receptor D, encoded by *IL17RD*, to be associated with cellular sensitivity to pemetrexed. A previous study found that the downregulation of *IL17RD* is common in certain cancer types, such as colon cancers, and can also promote tumor development (Girondel et al. 2021). We found that the lower predicted levels of interleukin-17

receptor D associate with a higher concentration of pemetrexed need for cytotoxicity; this is consistent with these prior findings, as *IL17RD* functions as a tumor suppressor, thus its inhibition may result in tumors that are more challenging to treat and require higher dosages of chemotherapeutics (Girondel et al. 2021).

Additionally, we performed FUMA gene set enrichment analysis on the top genes identified with TWAS (Watanabe et al. 2017). For CEU and ara-C, we identified enrichment in the oncogenic signature gene set P53_DN.V1_DN, which consists of genes that are down-regulated in cell lines with mutated *TP53* (A. Subramanian et al. 2005). Mutations in *TP53*, which encodes a tumor suppressor protein, are linked to various cancer types, and the genes in this set are often down-regulated in cancers where *TP53* is also mutated (A. Subramanian et al. 2005). *TP53* mutations are known to confer resistance to ara-C (Goldberg et al. 2018; Ko et al. 2019). We also found enrichment in the oncogenic signature WNT_UP.V1_UP for CEU and cisplatin. This gene set consists of upregulated genes in the Wnt signaling pathway, which is involved in cell proliferation (Ziegler et al. 2005). Abnormal activation of this pathway can result in tumor formation and progression (Giles, van Es, and Clevers 2003). For CEU and paclitaxel, enrichment was found in a GWAS Catalog Reported gene set, containing genes associated with liver enzyme levels. *GGT1* encodes gamma-glutamyl transferase, the main enzyme featured in this set, which cleaves extracellular glutathione and transfers its components—glutamic acid, cysteine, and glycine—for intracellular use (Bansal et al. 2019). Upregulation of *GGT1* is a feature of a variety of cancer types, including kidney and ovarian carcinomas (Bansal et al. 2019; Stordal et al. 2012). Ovarian carcinomas often are treated with combination chemotherapy using cisplatin and paclitaxel, as these drugs use different mechanisms to induce cell death; however, a subset of patients develop resistance to one or both of these drugs (Stordal et al. 2012).

Upregulation of *GGT1* was found to be associated with paclitaxel resistance in ovarian cancer cell lines already resistant to cisplatin (Stordal et al. 2012). Thus, the enrichment of genes in this set, which are associated with paclitaxel, and the association with *GGT1* in particular, may be understood in the context of this prior finding.

This study has limitations; only the *STARD5* TWAS association was functionally validated, functional studies of the other discovered GWAS, TWAS, and PAS associations have not yet been attempted. In addition, the functional follow up to the TWAS we conducted utilized the lung cancer cell line A549 rather than patients with lung cancer or another replication population. However, the A549 siRNA experiments we performed validated the association between increased *STARD5* expression and increased etoposide-induced cytotoxicity that we ascertained through TWAS. To fully understand how *STARD5* expression impacts the mechanisms through which etoposide induces cell death, further mechanistic studies are required. Association studies conducted with proteomic data could enhance these findings further, as well as additional functional studies that explore links between *STARD5* and drug metabolism. Moreover, if strides towards precision medicine are to continue, studies must promote greater diversity within participating populations, as currently the majority of human genome-wide studies are conducted on individuals of European ancestries (Hindorff et al. 2018; Landry et al. 2018). By studying diseases and drug response in populations with diverse ancestries data will become more representative of the global population and knowledge of genetic variants and their role in disease and drug response will be expanded (Landry et al. 2018). In summary, this project successfully identified novel genetic variants involved in chemotherapy-induced cytotoxicity in diverse ancestral populations through GWAS, TWAS, PAS, gene set enrichment analysis, and functional gene knockdown experiments.

REFERENCE LIST

1000 Genomes Project Consortium, Adam Auton, Lisa D. Brooks, Richard M. Durbin, Erik P. Garrison, Hyun Min Kang, Jan O. Korbel, et al. 2015. "A Global Reference for Human Genetic Variation." *Nature* 526 (7571): 68–74.

Abu Samaan, Tala M., Marek Samec, Alena Liskova, Peter Kubatka, and Dietrich Büsselberg. 2019. "Paclitaxel's Mechanistic and Clinical Effects on Breast Cancer." *Biomolecules* 9 (12). https://doi.org/10.3390/biom9120789.

Ahmed, Zeeshan. 2020. "Practicing Precision Medicine with Intelligently Integrative Clinical and Multi-Omics Data Analysis." *Human Genomics* 14 (1): 35.

Aslam, Bilal, Madiha Basit, Muhammad Atif Nisar, Mohsin Khurshid, and Muhammad Hidayat Rasool. 2017. "Proteomics: Technologies and Their Applications." *Journal of Chromatographic Science* 55 (2): 182–96.

Bansal, Ankita, Danielle J. Sanchez, Vivek Nimgaonkar, David Sanchez, Romain Riscal, Nicolas Skuli, and M. Celeste Simon. 2019. "Gamma-Glutamyltransferase 1 Promotes Clear Cell Renal Cell Carcinoma Initiation and Progression." *Molecular Cancer Research: MCR* 17 (9): 1881–92.

Barbeira, Alvaro N., GTEx Consortium, Scott P. Dickinson, Rodrigo Bonazzola, Jiamao Zheng, Heather E. Wheeler, Jason M. Torres, et al. 2018. "Exploring the Phenotypic Consequences of Tissue Specific Gene Expression Variation Inferred from GWAS Summary Statistics." *Nature Communications*. https://doi.org/10.1038/s41467-018-03621-1.

Barbeira, Alvaro N., Milton Pividori, Jiamao Zheng, Heather E. Wheeler, Dan L. Nicolae, and Hae Kyung Im. 2019. "Integrating Predicted Transcriptome from Multiple Tissues Improves Association Detection." *PLoS Genetics* 15 (1): e1007889.

Baretta, Zora, Simone Mocellin, Elena Goldin, Olufunmilayo I. Olopade, and Dezheng Huo. 2016. "Effect of BRCA Germline Mutations on Breast Cancer Prognosis: A Systematic Review and Meta-Analysis." *Medicine* 95 (40): e4975.

Baugh, Evan H., Hua Ke, Arnold J. Levine, Richard A. Bonneau, and Chang S. Chan. 2018. "Why Are There Hotspot Mutations in the TP53 Gene in Human Cancers?" *Cell Death and Differentiation* 25 (1): 154–60.

Bild, Diane E., David A. Bluemke, Gregory L. Burke, Robert Detrano, Ana V. Diez Roux, Aaron R. Folsom, Philip Greenland, et al. 2002. “Multi-Ethnic Study of Atherosclerosis: Objectives and Design.” *American Journal of Epidemiology* 156 (9): 871–81.

Bleibel, Wasim K., Shiwei Duan, R. Stephanie Huang, Emily O. Kistner, Sunita J. Shukla, Xiaolin Wu, Judith A. Badner, and M. Eileen Dolan. 2009. “Identification of Genomic Regions Contributing to Etoposide-Induced Cytotoxicity.” *Human Genetics* 125 (2): 173–80.

Bossé, Yohan, and Christopher I. Amos. 2018. “A Decade of GWAS Results in Lung Cancer.” *Cancer Epidemiology, Biomarkers & Prevention: A Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology* 27 (4): 363–79.

Brandes, Nadav, Nathan Linial, and Michal Linial. 2020. “PWAS: Proteome-Wide Association Study-Linking Genes and Phenotypes by Functional Variation in Proteins.” *Genome Biology* 21 (1): 173.

Browning, Sharon R., and Brian L. Browning. 2007. “Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies by Use of Localized Haplotype Clustering.” *American Journal of Human Genetics* 81 (5): 1084–97.

Bush, William S., and Jason H. Moore. 2012. “Chapter 11: Genome-Wide Association Studies.” *PLoS Computational Biology* 8 (12): e1002822.

Chen, Zhishan, Wanqing Wen, Alicia Beeghly-Fadiel, Xiao-Ou Shu, Virginia Díez-Obrero, Jirong Long, Jiandong Bao, et al. 2019. “Identifying Putative Susceptibility Genes and Evaluating Their Associations with Somatic Mutations in Human Cancers.” *American Journal of Human Genetics* 105 (3): 477–92.

Dasari, Shaloam, and Paul Bernard Tchounwou. 2014. “Cisplatin in Cancer Therapy: Molecular Mechanisms of Action.” *European Journal of Pharmacology* 740 (October): 364–78.

DeVita, Vincent T., Jr, and Edward Chu. 2008. “A History of Cancer Chemotherapy.” *Cancer Research* 68 (21): 8643–53.

Doll, Sophia, Florian Gnad, and Matthias Mann. 2019. “The Case for Proteomics and Phospho-Proteomics in Personalized Cancer Medicine.” *Proteomics. Clinical Applications* 13 (2): e1800113.

Evenepoel, Lucie, Francien H. van Nederveen, Lindsey Oudijk, Thomas G. Papathomas, David F. Restuccia, Eric J. T. Belt, Wouter W. de Herder, et al. 2018. “Expression of Contactin 4 Is Associated With Malignant Behavior in Pheochromocytomas and Paragangliomas.” *The Journal of Clinical Endocrinology and Metabolism* 103 (1): 46–55.

Friedman, Jerome, Trevor Hastie, and Rob Tibshirani. 2010. “Regularization Paths for Generalized Linear Models via Coordinate Descent.” *Journal of Statistical Software* 33 (1): 1–22.

Galmarini, Darío, Carlos M. Galmarini, and Felipe C. Galmarini. 2012. “Cancer Chemotherapy: A Critical Analysis of Its 60 Years of History.” *Critical Reviews in Oncology/Hematology* 84 (2): 181–99.

Gamazon, Eric R., GTEx Consortium, Heather E. Wheeler, Kaanan P. Shah, Sahar V. Mozaffari, Keston Aquino-Michaels, Robert J. Carroll, et al. 2015. “A Gene-Based Association Method for Mapping Traits Using Reference Transcriptome Data.” *Nature Genetics*. https://doi.org/10.1038/ng.3367.

Gamazon, Eric R., Jatinder K. Lamba, Stanley Pounds, Amy L. Stark, Heather E. Wheeler, Xueyuan Cao, Hae K. Im, et al. 2013. “Comprehensive Genetic Analysis of Cytarabine Sensitivity in a Cell-Based Model Identifies Polymorphisms Associated with Outcome in AML Patients.” *Blood* 121 (21): 4366–76.

Gamazon, Eric R., Matthew R. Trendowski, Yujia Wen, Claudia Wing, Shannon M. Delaney, Won Huh, Shan Wong, Nancy J. Cox, and M. Eileen Dolan. 2018. “Gene and MicroRNA Perturbations of Cellular Response to Pemetrexed Implicate Biological Networks and Enable Imputation of Response in Lung Adenocarcinoma.” *Scientific Reports* 8 (1): 733.

Garcia, Sara L., Jakob Lauritsen, Zeyu Zhang, Mikkel Bandak, Marlene D. Dalgaard, Rikke L. Nielsen, Gedske Daugaard, and Ramneek Gupta. 2020. “Prediction of Nephrotoxicity Associated With Cisplatin-Based Chemotherapy in Testicular Cancer Patients.” *JNCI Cancer Spectrum* 4 (3): kaa032.

Giles, Rachel H., Johan H. van Es, and Hans Clevers. 2003. “Caught up in a Wnt Storm: Wnt Signaling in Cancer.” *Biochimica et Biophysica Acta* 1653 (1): 1–24.

Girondel, Charlotte, Kim Lévesque, Marie-Josée Langlois, Sarah Pasquin, Marc K. Saba-El-Leil, Nathalie Rivard, Robert Friesel, et al. 2021. “Loss of Interleukin-17 Receptor D Promotes Chronic Inflammation-Associated Tumorigenesis.” *Oncogene* 40 (2): 452–64.

Giudice, Girolamo, and Evangelia Petsalaki. 2019. “Proteomics and Phosphoproteomics in Precision Medicine: Applications and Challenges.” *Briefings in Bioinformatics* 20 (3): 767–77.

Gold, Larry, Deborah Ayers, Jennifer Bertino, Christopher Bock, Ashley Bock, Edward N. Brody, Jeff Carter, et al. 2010. “Aptamer-Based Multiplexed Proteomic Technology for Biomarker Discovery.” *PloS One* 5 (12): e15004.

Goldberg, Aaron D., Chetasi Talati, Pinkal Desai, Christopher Famulare, Sean M. Devlin, Noushin Farnoud, David A. Sallman, et al. 2018. “TP53 Mutations Predict Poorer

Responses to CPX-351 in Acute Myeloid Leukemia." *Blood* 132 (Supplement 1): 1433–1433.

Hande, K. R. 1998. "Etoposide: Four Decades of Development of a Topoisomerase II Inhibitor." *European Journal of Cancer* 34 (10): 1514–21.

Hartford, Christine M., Shiwei Duan, Shannon M. Delaney, Shuangli Mi, Emily O. Kistner, Jatinder K. Lamba, R. Stephanie Huang, and M. Eileen Dolan. 2009. "Population-Specific Genetic Variants Important in Susceptibility to Cytarabine Arabinoside Cytotoxicity." *Blood* 113 (10): 2145–53.

Hasin, Yehudit, Marcus Seldin, and Aldons Lusis. 2017. "Multi-Omics Approaches to Disease." *Genome Biology* 18 (1): 83.

Hato, Stanleyson V., Andrea Khong, I. Jolanda M. de Vries, and W. Joost Lesterhuis. 2014. "Molecular Pathways: The Immunogenic Effects of Platinum-Based Chemotherapeutics." *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research* 20 (11): 2831–37.

Hindorff, Lucia A., Vence L. Bonham, Lawrence C. Brody, Margaret E. C. Ginoza, Carolyn M. Hutter, Teri A. Manolio, and Eric D. Green. 2018. "Prioritizing Diversity in Human Genomics Research." *Nature Reviews. Genetics* 19 (3): 175.

Huang, R. Stephanie, Shiwei Duan, Wasim K. Bleibel, Emily O. Kistner, Wei Zhang, Tyson A. Clark, Tina X. Chen, et al. 2007. "A Genome-Wide Approach to Identify Genetic Variants That Contribute to Etoposide-Induced Cytotoxicity." *Proceedings of the National Academy of Sciences of the United States of America* 104 (23): 9758–63.

Huang, R. Stephanie, Shiwei Duan, Emily O. Kistner, Wasim K. Bleibel, Shannon M. Delaney, Donna L. Fackenthal, Soma Das, and M. Eileen Dolan. 2008. "Genetic Variants Contributing to Daunorubicin-Induced Cytotoxicity." *Cancer Research* 68 (9): 3161–68.

Huang, R. Stephanie, Shiwei Duan, Emily O. Kistner, Christine M. Hartford, and M. Eileen Dolan. 2008. "Genetic Variants Associated with Carboplatin-Induced Cytotoxicity in Cell Lines Derived from Africans." *Molecular Cancer Therapeutics* 7 (9): 3038–46.

Huang, R. Stephanie, Shiwei Duan, Sunita J. Shukla, Emily O. Kistner, Tyson A. Clark, Tina X. Chen, Anthony C. Schweitzer, John E. Blume, and M. Eileen Dolan. 2007. "Identification of Genetic Variants Contributing to Cisplatin-Induced Cytotoxicity by Use of a Genomewide Approach." *American Journal of Human Genetics* 81 (3): 427–37.

Huang, Y., A. M. Chan, Y. Liu, X. Wang, and N. J. Holbrook. 1997. "Serum Withdrawal and Etoposide Induce Apoptosis in Human Lung Carcinoma Cell Line A549 via Distinct Pathways." *Apoptosis: An International Journal on Programmed Cell Death* 2 (2): 199–206.

Hussain, Tabish, and Rita Mulherkar. 2012. “Lymphoblastoid Cell Lines: A Continuous in Vitro Source of Cells to Study Carcinogen Sensitivity and DNA Repair.” *International Journal of Molecular and Cellular Medicine* 1 (2): 75–87.

International HapMap Consortium. 2003. “The International HapMap Project.” *Nature* 426 (6968): 789–96.

Jackson, Sarah E., and John D. Chester. 2015. “Personalised Cancer Medicine.” *International Journal of Cancer. Journal International Du Cancer* 137 (2): 262–66.

Kanda, Yusuke, Mitsuhiko Osaki, Kunishige Onuma, Ayana Sonoda, Masanobu Kobayashi, Junichi Hamada, Garth L. Nicolson, Takahiro Ochiya, and Futoshi Okada. 2017. “Amigo2-Upregulation in Tumour Cells Facilitates Their Attachment to Liver Endothelial Cells Resulting in Liver Metastases.” *Scientific Reports* 7 (March): 43567.

Kawashiro, T., K. Yamashita, X. J. Zhao, E. Koyama, M. Tani, K. Chiba, and T. Ishizaki. 1998. “A Study on the Metabolism of Etoposide and Possible Interactions with Antitumor or Supporting Agents by Human Liver Microsomes.” *The Journal of Pharmacology and Experimental Therapeutics* 286 (3): 1294–1300.

Ko, Ya-Chen, Chung-Yi Hu, Zheng-Hau Liu, Hwei-Fang Tien, Da-Liang Ou, Hsiung-Fei Chien, and Liang-In Lin. 2019. “Cytarabine-Resistant FLT3-ITD Leukemia Cells Are Associated with TP53 Mutation and Multiple Pathway Alterations—Possible Therapeutic Efficacy of Cabozantinib.” *International Journal of Molecular Sciences* 20 (5): 1230.

Komatsu, Masaaki, Heather E. Wheeler, Suyoun Chung, Siew-Kee Low, Claudia Wing, Shannon M. Delaney, Lidija K. Gorsic, et al. 2015. “Pharmacoethnicity in Paclitaxel-Induced Sensory Peripheral Neuropathy.” *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research* 21 (19): 4337–46.

Ku, Chee Seng, En Yun Loy, Yudi Pawitan, and Kee Seng Chia. 2010. “The Pursuit of Genome-Wide Association Studies: Where Are We Now?” *Journal of Human Genetics* 55 (4): 195–206.

Landry, Latrice G., Nadya Ali, David R. Williams, Heidi L. Rehm, and Vence L. Bonham. 2018. “Lack Of Diversity In Genomic Databases Is A Barrier To Translating Precision Medicine Research Into Practice.” *Health Affairs* 37 (5): 780–85.

Leroy, Bernard, Martha Anderson, and Thierry Soussi. 2014. “TP53 Mutations in Human Cancer: Database Reassessment and Prospects for the next Decade.” *Human Mutation* 35 (6): 672–88.

Liang, Baiqiang, Hongrong Ding, Lianfang Huang, Haiqing Luo, and Xiao Zhu. 2020. “GWAS in Cancer: Progress and Challenges.” *Molecular Genetics and Genomics: MGG* 295 (3): 537–61.

Litwiniec, Anna, Lidia Gackowska, Anna Helmin-Basa, Agnieszka Żuryń, and Alina Grzanka. 2013. “Low-Dose Etoposide-Treatment Induces Endoreplication and Cell Death Accompanied by Cytoskeletal Alterations in A549 Cells: Does the Response Involve Senescence? The Possible Role of Vimentin.” *Cancer Cell International*. https://doi.org/10.1186/1475-2867-13-9.

Liu, Edison T. 2008. “Functional Genomics of Cancer.” *Current Opinion in Genetics & Development* 18 (3): 251–56.

MacArthur, Jacqueline, Emily Bowler, Maria Cerezo, Laurent Gil, Peggy Hall, Emma Hastings, Heather Junkins, et al. 2017. “The New NHGRI-EBI Catalog of Published Genome-Wide Association Studies (GWAS Catalog).” *Nucleic Acids Research* 45 (D1): D896–901.

Manichaikul, Ani, Josyf C. Mychaleckyj, Stephen S. Rich, Kathy Daly, Michèle Sale, and Wei-Min Chen. 2010. “Robust Relationship Inference in Genome-Wide Association Studies.” *Bioinformatics* 26 (22): 2867–73.

Manzoni, Claudia, Demis A. Kia, Jana Vandrovcova, John Hardy, Nicholas W. Wood, Patrick A. Lewis, and Raffaele Ferrari. 2018. “Genome, Transcriptome and Proteome: The Rise of Omics Data and Their Integration in Biomedical Sciences.” *Briefings in Bioinformatics* 19 (2): 286–302.

Marin, Jose J. G., Marta R. Romero, Alba G. Blazquez, Elisa Herraez, Emma Keck, and Oscar Briz. 2009. “Importance and Limitations of Chemotherapy among the Available Treatments for Gastrointestinal Tumours.” *Anti-Cancer Agents in Medicinal Chemistry* 9 (2): 162–84.

Mogil, Lauren S., Angela Andaleon, Alexa Badalamenti, Scott P. Dickinson, Xiuqing Guo, Jerome I. Rotter, W. Craig Johnson, Hae Kyung Im, Yongmei Liu, and Heather E. Wheeler. 2018. “Genetic Architecture of Gene Expression Traits across Diverse Populations.” *PLoS Genetics* 14 (8): e1007586.

Nakajima, Kei, Ichiro Kawashima, Megumi Koshiisi, Takuma Kumagai, Megumi Suzuki, Jun Suzuki, Toru Mitsumori, and Keita Kirito. 2019. “Glycolytic Enzyme Hexokinase II Is a Putative Therapeutic Target in B-Cell Malignant Lymphoma.” *Experimental Hematology* 78 (October): 46-55.e3.

Neitzel, Leif R., Zachary T. Spencer, Anmada Nayak, Christopher S. Cselenyi, Hassina Benchabane, Cheyanne Q. Youngblood, Alya Zouaoui, et al. 2019. “Developmental Regulation of Wnt Signaling by Nagk and the UDP-GlcNAc Salvage Pathway.” *Mechanisms of Development* 156 (April): 20–31.

Niu, Nifang, and Liewei Wang. 2015. “In Vitro Human Cell Line Models to Predict Clinical Response to Anticancer Drugs.” *Pharmacogenomics* 16 (3): 273–85.

O'Donnell, Peter H., Amy L. Stark, Eric R. Gamazon, Heather E. Wheeler, Bridget E. McIlwee, Lidija Gorsic, Hae Kyung Im, R. Stephanie Huang, Nancy J. Cox, and M. Eileen Dolan. 2012. "Identification of Novel Germline Polymorphisms Governing Capecitabine Sensitivity." *Cancer* 118 (16): 4063–73.

Okada, Hirokazu, H. Alexander Ebhardt, Sibylle Chantal Vonesch, Ruedi Aebersold, and Ernst Hafen. 2016. "Proteome-Wide Association Studies Identify Biochemical Modules Associated with a Wing-Size Phenotype in Drosophila Melanogaster." *Nature Communications* 7 (September): 12649.

Olivier, Magali, Monica Hollstein, and Pierre Hainaut. 2010. "TP53 Mutations in Human Cancers: Origins, Consequences, and Clinical Use." *Cold Spring Harbor Perspectives in Biology* 2 (1): a001008.

Park, Hyojin, Sungwoon Lee, Pravesh Shrestha, Jihye Kim, Jeong Ae Park, Yeongrim Ko, Young Ho Ban, et al. 2015. "AMIGO2, a Novel Membrane Anchor of PDK1, Controls Cell Survival and Angiogenesis via Akt Activation." *The Journal of Cell Biology* 211 (3): 619–37.

Park, Sungshim L., Iona Cheng, and Christopher A. Haiman. 2018. "Genome-Wide Association Studies of Cancer in Diverse Populations." *Cancer Epidemiology, Biomarkers & Prevention: A Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology* 27 (4): 405–17.

Pruim, Randall J., Ryan P. Welch, Serena Sanna, Tanya M. Teslovich, Peter S. Chines, Terry P. Gliedt, Michael Boehnke, Gonçalo R. Abecasis, and Cristen J. Willer. 2010. "LocusZoom: Regional Visualization of Genome-Wide Association Scan Results." *Bioinformatics* 26 (18): 2336–37.

Purcell, Shaun, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A. R. Ferreira, David Bender, Julian Maller, et al. 2007. "PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses." *American Journal of Human Genetics* 81 (3): 559–75.

Qiu, Zhengang, Anqi Lin, Kun Li, Weiyin Lin, Qiongyao Wang, Ting Wei, Weiliang Zhu, Peng Luo, and Jian Zhang. 2019. "A Novel Mutation Panel for Predicting Etoposide Resistance in Small-Cell Lung Cancer." *Drug Design, Development and Therapy*. https://doi.org/10.2147/dddt.s205633.

Raffield, Laura M., Hong Dang, Katherine A. Pratte, Sean Jacobson, Lucas A. Gillenwater, Elizabeth Ampleford, Igor Barjaktarevic, et al. 2020. "Comparison of Proteomic Assessment Methods in Multiple Cohort Studies." *Proteomics* 20 (12): e1900278.

Rai, Yogesh, Priyanshu Yadav, Neeraj Kumari, Namita Kalra, and Anant Narayan Bhatt. 2019. "Hexokinase II Inhibition by 3-Bromopyruvate Sensitizes Myeloid Leukemic Cells K-

562 to Anti-Leukemic Drug, Daunorubicin." *Bioscience Reports* 39 (9). https://doi.org/10.1042/BSR20190880.

Rezen, Tadeja, Damjana Rozman, Jean-Marc Pascussi, and Katalin Monostory. 2011. "Interplay between Cholesterol and Drug Metabolism." *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*. https://doi.org/10.1016/j.bbapap.2010.05.014.

Rodriguez-Agudo, Daniel, Maria Calderon-Dominguez, Miguel Angel Medina, Shunlin Ren, Gregorio Gil, and William M. Pandak. 2012. "ER Stress Increases StarD5 Expression by Stabilizing Its MRNA and Leads to Relocalization of Its Protein from the Nucleus to the Membranes." *Journal of Lipid Research* 53 (12): 2708–15.

Rodriguez-Agudo, Daniel, Shunlin Ren, Phillip B. Hylemon, Kaye Redford, Ramesh Natarajan, Antonio Del Castillo, Gregorio Gil, and William M. Pandak. 2005. "Human StarD5, a Cytosolic StAR-Related Lipid Binding Protein." *Journal of Lipid Research* 46 (8): 1615–23.

Roell, Kyle R., Tammy M. Havener, David M. Reif, John Jack, Howard L. McLeod, Tim Wiltshire, and Alison A. Motsinger-Reif. 2019. "Synergistic Chemotherapy Drug Response Is a Genetic Trait in Lymphoblastoid Cell Lines." *Frontiers in Genetics* 10 (October): 829.

Roy, P. S., and B. J. Saikia. 2016. "Cancer and Cure: A Critical Analysis." *Indian Journal of Cancer* 53 (3): 441–42.

Schuetz, Erin, Lubin Lan, Kazuto Yasuda, Richard Kim, Thomas A. Kocarek, John Schuetz, and Stephen Strom. 2002. "Development of a Real-Time in Vivo Transcription Assay: Application Reveals Pregnane X Receptor-Mediated Induction of CYP3A4 by Cancer Chemotherapeutic Agents." *Molecular Pharmacology* 62 (3): 439–45.

Sellami, Maha, and Nicola Luigi Bragazzi. 2020. "Nutrigenomics and Breast Cancer: State-of-Art, Future Perspectives and Insights for Prevention." *Nutrients* 12 (2). https://doi.org/10.3390/nu12020512.

Shibata, Tatsuhiro. 2012. "Cancer Genomics and Pathology: All Together Now." *Pathology International* 62 (10): 647–59.

Sirugo, Giorgio, Scott M. Williams, and Sarah A. Tishkoff. 2019. "The Missing Diversity in Human Genetic Studies." *Cell* 177 (1): 26–31.

Stordal, Britta, Marion Hamon, Victoria McEneaney, Sandra Roche, Jean-Pierre Gillet, John J. O'Leary, Michael Gottesman, and Martin Clynes. 2012. "Resistance to Paclitaxel in a Cisplatin-Resistant Ovarian Cancer Cell Line Is Mediated by P-Glycoprotein." *PloS One* 7 (7): e40717.

Stratton, Michael R., Peter J. Campbell, and P. Andrew Futreal. 2009. “The Cancer Genome.” *Nature* 458 (7239): 719–24.

Subramanian, Aravind, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, et al. 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): 15545–50.

Subramanian, Indhupriya, Srikant Verma, Shiva Kumar, Abhay Jere, and Krishanpal Anamika. 2020. “Multi-Omics Data Integration, Interpretation, and Its Application.” *Bioinformatics and Biology Insights* 14 (January): 1177932219899051.

Sud, Amit, Ben Kinnersley, and Richard S. Houlston. 2017. “Genome-Wide Association Studies of Cancer: Current Insights and Future Perspectives.” *Nature Reviews. Cancer* 17 (11): 692–704.

Torre, L. A., R. L. Siegel, E. M. Ward, and A. Jemal. 2016. “Global Cancer Incidence and Mortality Rates and Trends--An Update.” *Cancer Epidemiology Biomarkers & Prevention*. https://doi.org/10.1158/1055-9965.epi-15-0578.

Trendowski, Matthew R., Omar El Charif, Paul C. Dinh Jr, Lois B. Travis, and M. Eileen Dolan. 2019. “Genetic and Modifiable Risk Factors Contributing to Cisplatin-Induced Toxicities.” *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research* 25 (4): 1147–55.

Trendowski, Matthew R., Omar El-Charif, Mark J. Ratain, Patrick Monahan, Zepeng Mu, Heather E. Wheeler, Paul C. Dinh Jr, et al. 2019. “Clinical and Genome-Wide Analysis of Serum Platinum Levels after Cisplatin-Based Chemotherapy.” *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research* 25 (19): 5913–24.

Turner, S. D. 2014. “Qqman: An R Package for Visualizing GWAS Results Using QQ and Manhattan Plots.” *Biorxiv*. https://www.biorxiv.org/content/10.1101/005165v1.full-text.

Tyanova, Stefka, and Juergen Cox. 2018. “Perseus: A Bioinformatics Platform for Integrative Analysis of Proteomics Data in Cancer Research.” *Methods in Molecular Biology* 1711: 133–48.

Uzozie, Anuli Christiana, and Ruedi Aebersold. 2018. “Advancing Translational Research and Precision Medicine with Targeted Proteomics.” *Journal of Proteomics* 189 (October): 1–10.

Varol, Umut, Yuksel Kucukzeybek, Ahmet Alacacioglu, Isil Somali, Zekiye Altun, Safiye Aktas, and Mustafa Oktay Tarhan. 2018. “BRCA Genes: BRCA 1 and BRCA 2.” *Journal of B.U.ON.: Official Journal of the Balkan Union of Oncology* 23 (4): 862–66.

Wang, Chen, Feng Zhang, Yu Cao, Mingming Zhang, Aixiu Wang, Mingcui Xu, Min Su, Ming Zhang, and Yuzheng Zhuge. 2016. “Etoposide Induces Apoptosis in Activated Human Hepatic Stellate Cells via ER Stress.” *Scientific Reports* 6 (September): 34330.

Wang, Shuang-Jia, Xiu-Dong Li, Lu-Peng Wu, Ping Guo, Liu-Xing Feng, and Bin Li. 2021. “MicroRNA-202 Suppresses Glycolysis of Pancreatic Cancer by Targeting Hexokinase 2.” *Journal of Cancer* 12 (4): 1144–53.

Watanabe, Kyoko, Erdogan Taskesen, Arjen van Bochoven, and Danielle Posthuma. 2017. “Functional Mapping and Annotation of Genetic Associations with FUMA.” *Nature Communications* 8 (1): 1826.

Weaver, Beth A. 2014. “How Taxol/Paclitaxel Kills Cancer Cells.” *Molecular Biology of the Cell* 25 (18): 2677–81.

Wheeler, H. E., E. R. Gamazon, A. L. Stark, P. H. O’Donnell, L. K. Gorsic, R. S. Huang, N. J. Cox, and M. E. Dolan. 2013. “Genome-Wide Meta-Analysis Identifies Variants Associated with Platinating Agent Susceptibility across Populations.” *The Pharmacogenomics Journal* 13 (1): 35–43.

Wheeler, Heather E., and M. Eileen Dolan. 2012. “Lymphoblastoid Cell Lines in Pharmacogenomic Discovery and Clinical Translation.” *Pharmacogenomics* 13 (1): 55–70.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer.

Yang, Lin, Jun Zhao, Wenqing Lü, Yong Li, Xiaojuan Du, Tao Ning, Guirong Lu, and Yang Ke. 2005. “KIAA0649, a 1A6/DRIM-Interacting Protein with the Oncogenic Potential.” *Biochemical and Biophysical Research Communications* 334 (3): 884–90.

Yoshida, Kiyotsugu, and Yoshio Miki. 2004. “Role of BRCA1 and BRCA2 as Regulators of DNA Repair, Transcription, and Cell Cycle in Response to DNA Damage.” *Cancer Science* 95 (11): 866–71.

Zheng, Tong, Min Lu, Ting Wang, Chunfeng Zhang, and Xiaojuan Du. 2018. “NRBE3 Promotes Metastasis of Breast Cancer by Down-Regulating E-Cadherin Expression.” *Biochimica et Biophysica Acta, Molecular Cell Research* 1865 (12): 1869–77.

Zhou, Xiang, and Matthew Stephens. 2012. “Genome-Wide Efficient Mixed-Model Analysis for Association Studies.” *Nature Genetics* 44 (7): 821–24.

Zhu, Linyan, and Liqun Chen. 2019. “Progress in Research on Paclitaxel and Tumor Immunotherapy.” *Cellular & Molecular Biology Letters* 24 (June): 40.

Ziegler, Slava, Sonja Röhrs, Lara Tickenbrock, Tarik Möröy, Ludger Klein-Hitpass, Ingrid R. Vetter, and Oliver Müller. 2005. “Novel Target Genes of the Wnt Pathway and Statistical Insights into Wnt Target Promoter Regulation.” *The FEBS Journal* 272 (7): 1600–1615.

Zucchetti, M., O. Pagani, V. Torri, C. Sessa, M. D’Incalci, M. De Fusco, J. de Jong, D. Gentili, G. Martinelli, and A. Tinazzi. 1995. “Clinical Pharmacology of Chronic Oral Etoposide in Patients with Small Cell and Non-Small Cell Lung Cancer.” *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research* 1 (12): 1517–24.

VITA

Ashley Mulford was born and raised in Tampa, Florida. She began attending Loyola University Chicago in August 2017 and began working as an undergraduate research assistant in the Wheeler lab in January, 2018. She was awarded the Biology Summer Fellowship and the Mulcahy Fellowship for her undergraduate research project in 2019. After earned her Bachelor of Science in Bioinformatics, *summa cum laude* from Loyola University Chicago in May 2020 Mulford continued her education, pursuing a Master of Science in Bioinformatics through the accelerated BS/MS Bioinformatics Program. She was awarded a Graduate Research Assistantship and Fellowship in 2020 and completed her Master of Science in Bioinformatics in May 2021. Moving forward, Mulford is excited to begin her professional career as a scientist in the fields of precision medicine and computational biology.