



# Integration of genetic and functional genomics data to uncover chemotherapeutic induced cytotoxicity

Ruowang Li <sup>1,2</sup> · Dokyoon Kim<sup>3</sup> · Heather E. Wheeler <sup>4</sup> · Scott M. Dudek<sup>2,5</sup> · M. Eileen Dolan <sup>6</sup> · Marylyn D. Ritchie <sup>1,2,5</sup>

Received: 18 July 2016 / Revised: 1 November 2017 / Accepted: 12 February 2018 / Published online: 25 May 2018  
© Macmillan Publishers Limited, part of Springer Nature 2018

## Abstract

Identifying genetic variants associated with chemotherapeutic induced toxicity is an important step towards personalized treatment of cancer patients. However, annotating and interpreting the associated genetic variants remains challenging because each associated variant is a surrogate for many other variants in the same region. The issue is further complicated when investigating patterns of associated variants with multiple drugs. In this study, we used biological knowledge to annotate and compare genetic variants associated with cellular sensitivity to mechanistically distinct chemotherapeutic drugs, including platinating agents (cisplatin, carboplatin), capecitabine, cytarabine, and paclitaxel. The most significantly associated SNPs from genome wide association studies of cellular sensitivity to each drug in lymphoblastoid cell lines derived from populations of European (CEU) and African (YRI) descent were analyzed for their enrichment in biological pathways and processes. We annotated genetic variants using higher-level biological annotations in efforts to group variants into more interpretable biological modules. Using the higher-level annotations, we observed distinct biological modules associated with cell line populations as well as classes of chemotherapeutic drugs. We also integrated genetic variants and gene expression variables to build predictive models for chemotherapeutic drug cytotoxicity and prioritized the network models based on the enrichment of DNA regulatory data. Several biological annotations, often encompassing different SNPs, were replicated in independent datasets. By using biological knowledge and DNA regulatory information, we propose a novel approach for jointly analyzing genetic variants associated with multiple chemotherapeutic drugs.

## Introduction

A better understanding of genetic variation contributing to cellular sensitivity to chemotherapeutic drugs can lead to more precise and personalized treatment of cancer patients [1]. Lymphoblastoid cell lines (LCLs) have been established as a model system to study the genetic components of drug-induced cytotoxicity by measuring cell growth inhibition following drug exposure [2]. Previous genome-wide association studies (GWAS) have identified numerous genetic variants and gene expression variables associated with drug cytotoxicity [3–6]. However, a comprehensive study of multiple drugs in different populations can reveal new insights into the genetic susceptibility of cytotoxicity.

We studied genetic factors associated with cytotoxicity of five mechanistically distinct chemotherapeutic drugs: cisplatin, carboplatin, capecitabine, cytarabine, and paclitaxel (Fig. 1). Cytotoxicities were measured for all drugs in two HapMap populations: Utah Residents with European ancestry (CEU) and African individuals from Yoruba in

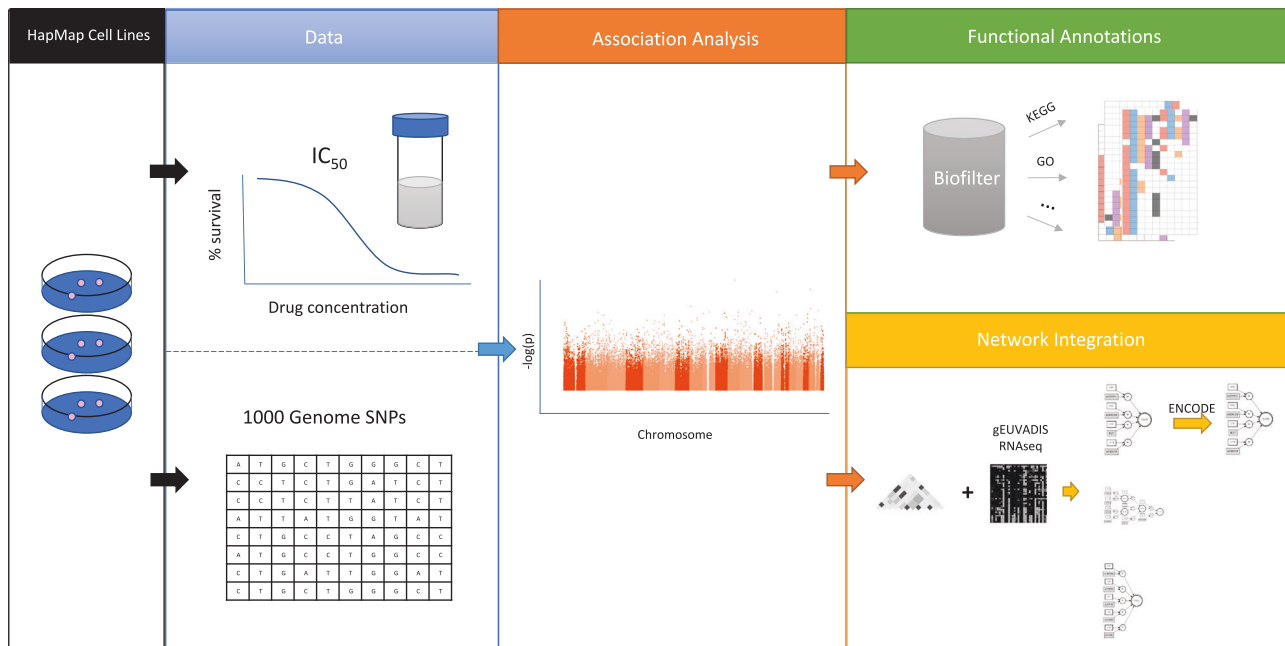
---

**Electronic supplementary material** The online version of this article (<https://doi.org/10.1038/s41397-018-0024-6>) contains supplementary material, which is available to authorized users.

---

✉ Marylyn D. Ritchie  
marylyn@pennmedicine.upenn.edu

- <sup>1</sup> Bioinformatics and Genomics program, Pennsylvania State University, University Park, Pennsylvania, USA
- <sup>2</sup> Institute for Biomedical Informatics, University of Pennsylvania, Perelman School of Medicine, Philadelphia, PA, USA
- <sup>3</sup> Biomedical and Translational Informatics, Geisinger, Danville, Pennsylvania, USA
- <sup>4</sup> Departments of Biology and Computer Science, Loyola University Chicago, Chicago, Illinois, USA
- <sup>5</sup> Department of Genetics, University of Pennsylvania, Perelman School of Medicine, Philadelphia, PA, USA
- <sup>6</sup> Department of Medicine, University of Chicago, Chicago, Illinois, USA



**Fig. 1** Using functional annotations and network analysis to model chemotherapeutic induced cytotoxicity, chemotherapeutic drugs'  $IC_{50}$  and SNPs' genotype were measured on the HapMap CEU and YRI cell lines. GWAS were used to identify the most significantly associated SNPs for each drug and population; the resulting SNPs were

annotated using Biofilter. To integrate SNP and gene expression, the associated SNPs and gene expression variables were trained using the grammatical evolution neural network. The resulting network models were prioritized using the ENCODE functional data.

Ibadan, Nigeria (YRI). Platinum-based compounds, including cisplatin and carboplatin are the most widely applied group of cytotoxic drugs worldwide, are used to treat head and neck, testicular, lung, endometrial and ovarian cancers [7–9]. Capecitabine is mainly used to treat colorectal and breast cancers [10]. Patients with acute myeloid leukemia have long been treated with cytarabine [11]. Paclitaxel is commonly used for the treatment of lung, breast, and ovarian cancers [12]. Previous studies have shown that drugs in the same class have common genetic loci associated with drug induced cytotoxicity, for example, cisplatin and carboplatin [4]. An individual's ancestral background has also been linked to differential risks for cytotoxicity [13]. Thus, a more comprehensive understanding of the distinct and shared genetic components associated with cytotoxicity between drugs and populations would be valuable to identify new treatment options.

However, a molecular understanding of individual genetic variations is challenging because there are a large number of genetic variations that can be associated with drug cytotoxicity and each variant is a surrogate for many other variants in the same region. To address these issues, we evaluated genetic variants using higher-level biological annotations in efforts to group variants into more interpretable biological modules. Comparing CEU to YRI, we found population specific annotations for each drug. Within individual populations, we observed drugs that treat similar

types of cancers are enriched for the same biological annotations. In some cases, we identified similar biological annotations across CEU and YRI, as well as across multiple drugs.

Previous studies relied on GWAS to identify genetic variants that have the strongest independent genetic effects on drug-induced cytotoxicity and incorporated gene expression levels through studies of expression quantitative trait loci (eQTL) analysis [14]. This work led to the important observation that pharmacological GWAS SNPs are enriched in eQTLs for many cytotoxic drugs [15]. While the eQTL method can capture a linear relationship between SNPs and gene expression, it overlooks the possibility that interactions among SNPs or gene expression could also play a crucial role in drug cytotoxicity. To identify these non-linear interactions, we applied the grammatical evolution neural network (GENN) algorithm to build interaction networks consisting of SNPs and gene expression variables. Although the identification of associated SNPs and gene expression variables is an important first step in understanding drug cytotoxicity, a challenge remains on how to interpret the functional relevance of the interaction models. It has been shown that many regulatory elements can aid in identifying important functional SNPs [16, 17]. To this end, we used DNaseI and genome segmentation data published by the ENCODE consortium to prioritize the network models.

In this study, our goal is to better understand chemotherapeutic drug responses from three new avenues: First, while most of the existing studies have analyzed one drug at a time [5, 9, 18], our study investigated multiple drugs in two different populations. Second, to overcome the bias of selecting individual causal variants, we grouped the associated SNPs to higher biological modules including genes, protein families, and other biological annotations. While individual variant's association is affected by the linkage disequilibrium patterns, it should have minimal impact on the higher-level annotations. Third, we integrated SNPs and gene expression variables using a network approach and quantified the importance of the additional data source in terms of prediction accuracy. We believe that a comparative study of multiple drugs and integration of different data sources could be a useful approach for interpreting genetic factors contributing to chemotherapeutic drug responses.

## Methods

### Genetic variants and gene expression data

Genetic variants data for Utah residents with Northern and Western European ancestry (CEU) and African individuals from the Yoruba in Ibadan, Nigeria (YRI) were downloaded from the 1000 Genome project (phase1\_release\_v3.20101123) [19]. RNAseq gene expressions on the same individuals were downloaded from the gEUVADIS project [20]. The gene expression data was normalized by library depth and transcripts length (RPKM). Gene expressions with 0 counts in more than half the samples were removed and technical variations were adjusted by PEER normalization. The detailed normalization process was described in [20].

### Cytotoxicity data

Lymphoblastoid cell lines from HapMap phase 1 CEU and YRI populations were treated with increasing concentrations of capecitabine [18], carboplatin [3], cisplatin [6], cytarabine [21], and paclitaxel [22] as previously reported. For carboplatin and cisplatin, their  $IC_{50}$ , concentration required to inhibit 50% of the cell growth, were calculated and log2 transformed to normality. The areas under the survival curve (AUC) were calculated for capecitabine, cytarabine, and paclitaxel. To satisfy the regression assumptions, all AUC values were also log2 transformed to allow for normal distribution. For replication studies, HapMap phase 3 YRI and CEU cell lines were treated with four of the drugs under the same concentrations: capecitabine, carboplatin, cisplatin, and cytarabine.

**Table 1** SNPs and gene expression variables count after quality control

Drugs	Population	Sample size	SNP (million)	Expression
Cisplatin	CEU	72	3.87	19,919
	YRI	77	4.69	20,380
Carboplatin	CEU	72	3.87	19,923
	YRI	75	4.64	20,427
Cytarabine	CEU	72	3.87	19,911
	YRI	77	4.68	20,380
Capecitabine	CEU	73	3.88	19,859
	YRI	76	4.66	20,421
Paclitaxel	CEU	29	2.71	19,683
	YRI	29	2.99	20,045

Autosome SNPs that have 100% call rate and gene expression variables that have 90% non-zero values were retained. To increase statistical power, SNPs were further filtered so that all three genotypes are present and each genotype has a minimum of two samples.

### Quality control for genetic variants and gene expression data

The SNP data was first transformed into the variant call format. Only SNP data from the autosomes were used for the GWAS analyses. To minimize error accompanied with the sequencing technology, only SNPs with 100% call rate were retained using GATK [23]. To remove extreme outliers and increase statistical power, we limited our analysis to SNPs that have all three possible genotypes and each genotype has at least two representing samples [5]. This effectively removed all of the rare variants, which do not have the power to be detected given the small sample size. Between 2.7 and 4.7 million SNPs have passed the quality control. Gene expressions were filtered so that 90% samples have non-zero expression values. This resulted in around 20,000 gene expression probes being retained (Table 1).

### GWAS analyses of drug susceptibility

In order to perform subsequent integration analyses using genetic variants and gene expression data, only samples that are common between cytotoxicity data, 1000 Genome genetic variant data, and gEUVADIS gene expression data were used for GWAS analyses. As a result, the number of samples is different for each drug (Table 1) and all of the study samples are unrelated. To control for potential confounding effects due to population structure, SNPs that passed quality control criteria were first LD-pruned ( $-r^2 < 0.5$ ) using PLINK software [24]. The principal components of the pruned SNP data were estimated using Eigenstrat [25]. Along with individual's sex, significant principal

components (2 or 3) were adjusted in the association analysis for each SNP. For gene expression data, individual's sex was adjusted for each expression probe.

### Functional meta-analysis of associated SNPs

To determine the biological annotations that are associated across populations and drugs, Biofilter (v2.2) [26] was used to separately map the associated SNPs of each cytotoxicity phenotype to functional groups including genes regions [27], protein family (Pfam) [28], gene ontology (GO) term [29], KEGG pathway [30], and Reactome [31]. Then, for each of the functional groups, we investigated whether any of its functional terms were shared in multiple populations and drugs. To evaluate the significance of the sharing, we carried out 1000 permutation tests, where we permuted each drug's cytotoxicity and performed GWAS on the permuted outcome. If less than 5 out of 1000 permutations resulted in equal or larger number of sharing for a function term, the term was deemed significant ( $p < 0.005$ ). After permutation, 63 genes, 35 GO terms, 2 KEGG pathways, 12 Pfam, and 39 Reactome were determined to be significant.

### Integration analysis using ATHENA

The Analysis Tool for Heritable and Environmental Network Associations (ATHENA) is a multifunctional software package that provides machine learning tools to analyze genomics data. The software has been extensively tested and applied in simulation data and real world data with great success [32, 33]. The software and its modeling processes have been described previously [34]. Briefly, we used an evolutionary algorithm, grammatical evolution neural network (GENN), to optimize artificial neural networks (ANNs), which are used to integrate genetic variants and gene expression data. The evolution process initiates a set of random models and these random models compete with each other through generations. The "fittest" models, or the models that maximize desired target function, can exchange components of themselves. Through transferring of the components, some models may acquire beneficial components and eventually take over the population pool. This evolution process mimics natural selection where the "fittest model" will survive at the end of evolution. The algorithm is described below.

Step 1: The data is divided into five parts for five cross validations with 4/5 for training and 1/5 for testing.

Step 2: Under population size constraint, a random population of models (ANNs) is generated.

Step 3: All models are evaluated with training data. The models with highest fitness are selected for crossover, mutation, reproduction and migration.

Step 4: Step 3 is repeated for a set number of generations.

Step 5: The best solution at the final generation is tested on the testing data and saved.

Step 6: Steps 2–5 are repeated for each cross validation.

The fitness of the model aims to measure how well the variables can explain the cytotoxicity, a continuous value. We used  $R$ -squared as our fitness metric to represent the percentage of cytotoxicity variation explained by SNPs and gene expression. We scaled the cytotoxicity to be between 0 and 1 using min-max scaling so that it matches the output of neural networks, where

$$\text{Normalized } D_i = \frac{D_i - \min(D)}{\max(D) - \min(D)} \quad (1)$$

\* $D_i$  is the  $IC_{50}$  value for the  $i_{th}$  sample,  $D$  is the vector of  $IC_{50}$  values.

And the  $R^2$  is calculated as:

$$R^2 = 1 - \frac{\sum_i^n (D_{\text{predict } i} - D_i)^2}{\sum_i^n (D_i - \bar{D})^2} \quad (2)$$

$D_{\text{predict } i}$  is the predicted  $IC_{50}$  for the  $i_{th}$  sample,  $D_i$  is the  $IC_{50}$  value for the  $i_{th}$  sample,  $\bar{D}$  is the average value of  $IC_{50}$ .

Linkage disequilibrium patterns exist in the associated SNPs because many are proximately located. Even though the SNPs may have distinct biological functions, they are indistinguishable in regards to their association with cytotoxicity because they are highly correlated. To reduce the correlated signals resulting from LD, for each cytotoxicity phenotype, pairwise LD among all associated SNPs were estimated.  $r^2 > 0.7$  was used as a threshold to form LD clusters among the associated SNPs and if a cluster has more than one SNP, the SNP that is the most significantly associated with cytotoxicity was selected as the tag SNP for the cluster. To reduce multi-collinearity in the gene expression data, Pearson correlation was calculated for all possible gene pairs. Genes that have correlation coefficient  $r > 0.8$  were grouped into a cluster. The most significantly associated gene from each cluster was selected as the tag gene for the cluster.

We first used ATHENA to perform variable selection on tagging SNPs and gene expressions. SNPs and gene expressions were integrated together to build neural networks that model the data. We selected SNPs and gene expressions that were consistent between cross validations [33]. The variable selection step used training  $R$  squared to select candidate SNP and gene expression variables; however, the final models are selected based on testing  $R$  squared. Using the selected SNPs and gene expressions, we used ATHENA to build five models, one for each cross validation, and for each cytotoxicity phenotype.

**Table 2** Genotype and gene expression associations with chemotherapeutic drugs

Drugs	Population	Discovery associated SNPs	Discovery associated expressions	Hapmap 3 replication LCLs	Replicated SNPs
Cisplatin	CEU	1945	121	40	292
	YRI	2157	76 <sup>a</sup>	46	266
Carboplatin	CEU	2530	169 <sup>a</sup>	40	302
	YRI	2364	194	44	247
Cytarabine	CEU	2156	126	40	275
	YRI	2749	106 <sup>a</sup>	46	722
Capecitabine	CEU	2014	65 <sup>a</sup>	40	122
	YRI	2485	295	46	297
Paclitaxel	CEU	1230	94	NA	NA
	YRI	1466	80	NA	NA

Associated SNPs and gene expressions were identified in the discovery HapMap phase 1 samples. Replication of SNPs were performed in the HapMap phase 3 LCLs. SNPs association threshold was at  $p < 5 \times 10^{-4}$ . Gene expressions association threshold were at  $p < 5 \times 10^{-4}$  or  $p < 5 \times 10^{-3}$ .

## Using functional data to prioritize neural network models

In order to distinguish Neural Network models that have similar predictable power of cytotoxicity, we utilized functional data produced by the ENCODE project [35] to quantify the functional relevance of each model. We downloaded 128 DNase-I hypersensitivity samples from the ENCODE project ([http://ftp.ebi.ac.uk/pub/databases/ensembl/encode/integration\\_data\\_jan2011/byDataType/openchrom/jan2011/combined\\_peaks/](http://ftp.ebi.ac.uk/pub/databases/ensembl/encode/integration_data_jan2011/byDataType/openchrom/jan2011/combined_peaks/)). The data contains merged DNase-I peaks from UW and Duke that passed FDR 1% cutoff. Genome segmentations of six ENCODE cell lines (GM12878, K562, H1-hESC, HeLa-S3, HepG2, and HUVEC) were obtained from ([http://ftp.ebi.ac.uk/pub/databases/ensembl/encode/integration\\_data\\_jan2011/byDataType/segmentations/jan2011/hub/](http://ftp.ebi.ac.uk/pub/databases/ensembl/encode/integration_data_jan2011/byDataType/segmentations/jan2011/hub/)). We used the combined segmentations calls based on the consensus calling of ChromHMM and Segway algorithms. The combined segmentations split the genome into non-overlapping regions of CTCF enriched element, enhancer, weak enhancer, promoter flanking region, promoter region including TSS, transcribed region, and repressed region. For every SNP in the neural network model, we determined whether it is located in DNase-I hypersensitive regions or genome segmentation regions across all cell types. Because the network models only include the tagging SNPs, we also determined the functional region overlaps for SNPs that are in LD with the tagging SNP. The functional score for each model is calculated as the sum of overlap for each individual SNP, normalized by the model size. In the case where SNPs in LD with the tagging SNP has a higher number of overlaps, the tagging SNP was replaced with the LD SNP. In order to select the final model, we first selected 3 models that have the best testing accuracy ( $R^2$ ). Of those, we selected the model with highest functional score as the final model.

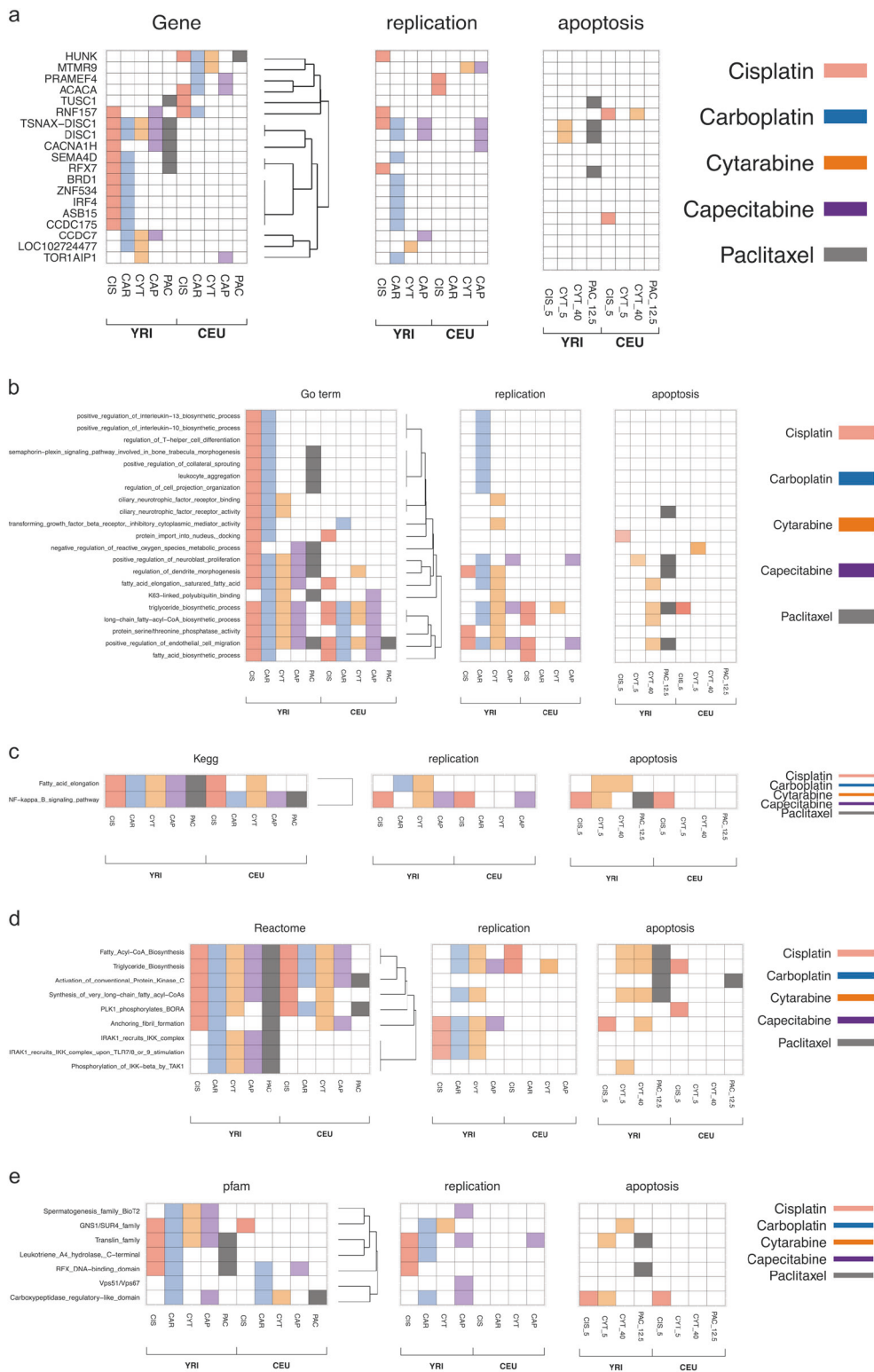
Once we had the final model, we used SNPs and gene expressions to separately build SNP and gene expression only models. In the case where the models have worse prediction than predicting using the mean  $IC_{50}$ , the  $R^2$  value was replaced with 0, thus effectively removed.

## Results

### Chemotherapeutic drug genetic associations

Cell growth inhibition was measured previously on unrelated CEU and YRI LCLs following treatment with increasing concentrations of cisplatin [6], carboplatin [3], cytarabine [21], capecitabine [18] or paclitaxel [36]. Their dose-dependent inhibition was calculated as  $IC_{50}$ , concentration required to inhibit 50% of cell growth, or AUC, area under the survival-drug concentration curve for up to 77 LCLs (Table 1).

Genome-wide SNP data for the LCLs were obtained from the 1000 Genomes Project (<http://www.1000genomes.org/>) and were evaluated for their association with each drug's cytotoxicity. We adjusted for sex and significant principal components of ancestry (2 or 3) in the linear regression model. The quantile-quantile plots and the genomic inflation factors ( $\lambda$ ) showed that there was minimal inflation (S1 Fig). We identified between 1230 and 2749 SNPs significantly associated with each drug response at  $p < 0.0005$ , respectively (Table 2). Gene expression levels for the LCLs, measured by RNA-Seq, were downloaded from the gEUVADIS consortium (<http://www.geuvadis.org/>). Normalized RPKM (reads per kilobase per million) values for ~20,000 genes were tested for association with each drug's  $IC_{50}$  or AUC. To keep the number of associated genes similar across drugs, we used  $p < 0.005$  or  $p < 0.0005$



**Fig. 2** Pan-drug analysis of functional annotations. For each drug in CEU and YRI, associated SNPs were mapped to various functional annotations. A colored square indicates SNP(s) were mapped to that functional term (Cisplatin: Red, Carboplatin: Blue, Cytarabine: Orange, Capecitabine: Purple, Paclitaxel: Black). Only functional

terms that have significant enrichment across drugs and populations (permutation analysis  $p < 0.005$ ) were shown. Functional terms were grouped using hierarchical clustering according to its enrichment. **a** Gene; **b** GO term; **c** KEGG pathway; **d** REACTOME; **e** Pfam

to select candidate genes. We identified between 65 and 295 genes whose expression levels were associated with drug outcome (Table 2). A list of all associated SNPs and gene expression levels can be found in the supplemental materials (S1 Table).

To replicate the SNP associations, we exposed an independent set of HapMap phase 3 LCLs to four of the five chemotherapeutic drugs: cisplatin, carboplatin, cytarabine, and capecitabine. We performed an association analysis on the independent LCLs and using the same  $p$ -value threshold ( $p < 0.0005$ ) and direction of effect, we replicated between 122 and 722 SNPs that were associated in the original samples (Table 2).

### Pan-drug analysis of associated SNPs reveals distinct patterns of functional enrichment

To get a better understanding of the biological processes involved in the differential cytotoxicity, we annotated all the SNPs that are associated with each drug response using gene regions, KEGG pathways, GO terms, REACTOME, and Pfam using Biofilter [26]. We observed that many biological annotations were shared across different drugs and/or populations. To remove annotations that were shared due to random chance, we performed a permutation test (1000 $\times$ ) for each drug's IC<sub>50</sub> or AUC. Using the permuted IC<sub>50</sub> or AUC, we identified associated SNPs using the same criteria as our original analysis. For each permutation, we calculated how many times an annotation is shared across the drug and population. We then removed any annotations that are over-represented in the permutations ( $p < 0.005$ ).

Cellular sensitivity to drugs is a broad phenotype that includes cell cycle arrest, cell damage, and cell death through apoptotic and non-apoptotic mechanisms [37, 38]. Cytarabine (5 and 40  $\mu$ M), cisplatin (5  $\mu$ M) and paclitaxel (12.5 nM) were evaluated for chemotherapeutic-induced apoptosis because they cause a significant increase in cellular caspase-3/7 activation, a measure of apoptosis [22]. For each drug, apoptosis was measured in 30 CEU LCLs and 35 YRI LCLs. We identified SNPs that are associated with drug-induced caspase 3/7 activation (S2 Table) and mapped them using biological annotations. To obtain the most stringent list of biological annotations that are shared between different drugs and populations, we kept only the annotations that passed the permutation test and were also identified in the replication or apoptosis dataset (Fig. 2).

When we compared the associated functional annotations across CEU and YRI LCLs, we observed that some annotations are population specific. For gene annotations, a group of genes including *HUNK*, *MTMR9*, *PRAMEF4*, and *ACACA* were only associated in the CEU population for at least two chemotherapeutics (Fig. 2a). Meanwhile, Spermatogenesis family BioT2, GNS1/SUR4 family, Translin

family, and Leukotriene A4 hydrolase C-terminal in pfam (Fig. 2e), *IKK* related terms in REACTOME (Fig. 2d), and several neuronal development and leukocytes GO terms (Fig. 2b) were only identified in the YRI population. On the other hand, there is a common group of functional terms associated in both CEU and YRI populations. This group consists of mostly fatty acid related functional terms clustered together in GO term, REACTOME, and KEGG pathway. One notable example is the NF-kappa B signaling pathway in the KEGG pathway. This pathway was associated with all of the drugs in both populations (Fig. 2c).

Within each population, we observed that drugs within the same class have similar associated annotation patterns. In particular, cisplatin and carboplatin, both platinating agents have many functional annotations in common. Cytarabine and capecitabine, both antimetabolites, have a number of overlapping annotations (Fig. 2).

We also observed overlapping annotations between drug cytotoxicity and apoptosis. *TSNAX-DISCI* and *DISCI* gene was associated with cytarabine and paclitaxel for both cell cytotoxicity and apoptosis. A number of triglyceride and fatty acid GO terms and REACTOME pathways were shared for cytarabine, paclitaxel and cisplatin. Both *Fatty acid elongation* and NF-kappa B signaling pathway in KEGG are enriched for both processes. In Pfam, GNS1/SUR4 family, Translin family, and RFX DNA binding domain were enriched for cytotoxicity and apoptosis.

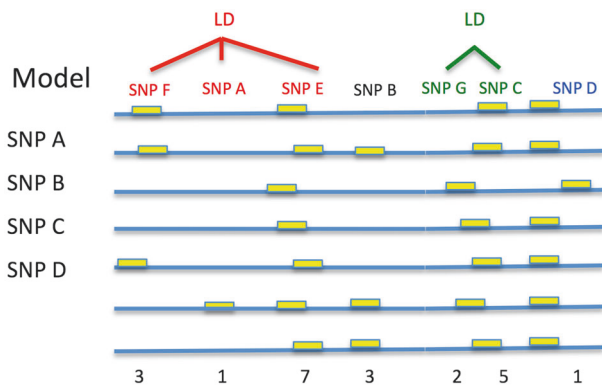
### Network modeling identified interactions between SNPs and gene expression variables important in cytotoxicity

Starting with the SNPs and gene expression variables that were associated with each drug's cytotoxicity, we calculated pairwise correlations among SNPs or gene expression. Using cutoffs of  $r^2 > 0.7$  for SNPs and Pearson's  $r > 0.8$  for gene expression, we grouped SNPs and gene expression variables that are highly correlated to the same clusters. To reduce multi-collinearity for the network analysis, we selected one tag SNP or tag expression that had the highest association with cytotoxicity to represent each cluster.

We integrated the tag SNPs and gene expressions using GENN and built interaction network models for each drug and population combinations.

### Using ENCODE data to prioritize network models

It is possible that a number of network models can be similarly predictive for each drug's cytotoxicity. To prioritize these models, we selected the model that contains variables with evidence of functional relevance from the ENCODE. Previous studies suggested that SNPs that lie in the open chromatin and regulatory regions are more



**Fig. 3** Schematic for functional score calculation. Functional score of a model is calculated as the sum of scores of individual SNP or SNPs in LD normalized by the model size. Individual score was determined by its positional overlap with functional regions. In this example, yellow squares represent DNaseI or genome segmentation regions. The score for a network model of SNP A, B, C, D is  $(7 + 3 + 5 + 1)/4 = 4$

likely to be functional [35]. Thus, we used DNaseI hypersensitivity sites from 124 cell lines and genome segmentation data from 6 cell lines produced by the ENCODE project to give functional relevance for each model. The DNaseI data marks genomic regions that are not occupied by heterochromatin and the genome segmentation data divides the genome into enhancer, transcription start sites, promoter-flanking regions, CTCF binding sites, and repressed regions. For each network model, we first identify the full set of features by including SNPs that are in the same clusters as the tag SNPs in the model. We then calculated a functional score for each feature that is proportion to the number of functional elements it overlaps with in all of the cell lines. The final score for a network model is the summation of the individual score for each feature normalized by network size (Fig. 3). Using the functional score, we were able to prioritize models that have similar predictive power in terms of  $R^2$  (amount of variability explained by the model) and identified one final model for each drug and population (Table 3 & S2-10 Fig). As an example, the multi-layer neural network for capecitabine in CEU population uses 6 SNPs to model  $IC_{50}$  response (Fig. 4).

## Discussion

Understanding a patient's genetic susceptibility to chemotherapeutic drugs will provide important information for precision medicine. Previous studies have evaluated genotype associations to an individual chemotherapeutic drug; however, a comparative study of multiple drugs in different populations could reveal common or unique mechanisms that can be exploited in terms of therapy. Here, we present the first study to analyze the genetic associations of

cytotoxicity induced by five chemotherapeutic drugs (cisplatin, carboplatin, capecitabine, cytarabine, and paclitaxel) in LCLs derived from two populations (CEU and YRI). To comparatively analyze the associated genetic variants across multiple drugs in two populations, higher-level biological knowledge was used to group variants into functional modules. We discovered that mechanistically distinct drugs are enriched in the same functional modules such as NF- $\kappa$ B pathway. We also set to identify biomarkers that are predictive of the drug cytotoxicity. To this end, we found that integrated networks of SNP and gene expression performed better than either data type alone. Finally, we used DNA regulatory information to select network models that are both predictive and functionally important.

We performed genome-wide SNP association analysis for each of the five drugs in both populations to identify significant genetic associations with drug-induced cytotoxicity. A major challenge to interpreting significant SNP associations across different drugs and populations is that comparing individual SNPs alone can be misleading. A slight change in allele frequency could result in any of the SNPs in linkage disequilibrium to be identified, however, SNPs in LD are likely located in the same genes or regions. We, therefore, annotated the associated SNPs to higher-level biological processes using gene regions, GO term, KEGG pathway, REACTOME pathway, and Pfam. We found that biological annotations are considerably different between LCLs derived from individuals of European and African ancestry. Interestingly, ancestry has also been reported to affect gene expression [14], modified cytosines [39] and sensitivity to chemotherapy [13]. The disparities might lie in the differences in population susceptibility to cancer, which could also affect cytotoxicity-induced response. *HUNK* and *ACACA* genes were associated only in the CEU population and are both related to breast cancer [40, 41] (Fig. 2a). A previous report has shown that differences exist between African American and European American women in the nature of breast cancer [42]. *SEMA4D* and *CCDC7* genes were associated in the YRI population (Fig. 2a). Expressions of the genes have been reported to correlate with poor outcome in cervical cancer [43, 44]. In addition, a recent survey has found that Black American females are more likely to develop cervical cancer than White American females [45] and Black American males have higher lung cancer incidence compared to White American males [46]. These candidate genes could be further validated in their respective population. Several *IKK* related REACTOME pathways were associated with the YRI population (Fig. 2d). *IKK* is a central regulator of NF- $\kappa$ B pathway [47] and activation of NF- $\kappa$ B pathway has been observed in many solid tumors [48]. Interestingly, NF- $\kappa$ B pathway is associated in both CEU and YRI populations (Fig. 2c), but *IKK* is only associated with the YRI



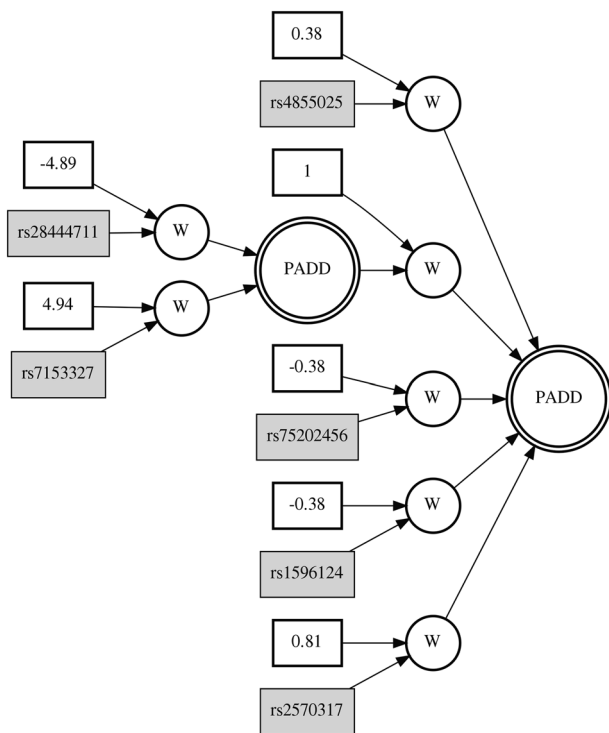
**Table 3** Network model identified by GENN

Drugs	Population	$R^2$			SNPs (LD)	DNaseI	Genome segmentation	Gene
		Integration	SNP	Expression				
Capecitabine	CEU	67.9	67.9	NA	rs4855025	NA	R, R, R, R, R, R	NA
					rs28444711	NA	R, R, R, R, R, R	
					rs7153327	11	R, R, R, R, R	
					rs75202456	NA	R, R, R, R, R	
					rs1596124	NA	R, R, R, R, R, R	
	YRI	64.3	64.3	NA	rs2570317	NA	R, R, R, R, R, R	
					rs11204113	NA	R, R, R, R, R, R	
					rs10760086	NA	R, R, R, R, R, R	
					rs9303059	NA	R, R, R, R, R, R	
					rs9661131	NA	R, R, R, R, R, R	
Carboplatin	CEU	60.4	62.1	23.1	rs6671214	NA	CTCF, CTCF, CTCF, CTCF, CTCF, R	TMEM14E
					rs11233413	9	T, E	
					rs12816395	NA	T, T, R, R, R, R	
					rs79062064	NA	T, T, R, R, R, R	
					rs16823342	NA	R, R, R, R, R, R	
	YRI	66.2	66.2	NA	rs2553650	5	WE, R, R, R	
					rs2079192	3	T, T, T, T, WE	
					rs7325063	NA	T, R, R, R, R, R	
					rs916396	NA	T, T, R, R, R, R	
					rs11715866	NA	T, R, R, R, R, R	
Cisplatin	CEU	66.6	46.3	14.0	rs344946	NA	R, R, R, R, R, R	FABP6 HCFC1 TAS2R30 ZNF192P1
					rs11628331	NA	R, R, R, R	
					rs77859257	NA	R, R, R, R, R, R	
					rs557453	NA	T, T, R, R, R, R	
					rs9422887	9	CTCF, CTCF, CTCF, CTCF, CTCF, CTCF	
	YRI	52.4	36.8	12.7	rs8074638	5	R, R, R, R, R, R	
					rs557453	NA	T, T, R, R, R, R	
					rs812652	NA	R, R, R, R, R, R	
					rs4750139	5	TSS, TSS, R	
					rs7257166	2	WE, T, R, R, R, R	
Cytarabine	CEU	47.7	42.9	0	rs12255911	NA	T, T, T, T, R, R	RP11-463J10.3 IL11RA
					rs6814234	9	WE, T, T, R, R	
					rs10426529	NA	E, R, R, R, R, R	
					rs1281461	NA	R, R, R, R, R, R	
					rs2780788	NA	T, R, R, R, R, R	
YRI	72.2	28.2	45.4	rs593525	11	T, T, T		
				rs4910512	2	T, R, R, R		
				rs7962806	NA	R, R, R, R, R, R		
				rs7666224	NA	R, R, R, R, R, R		
				rs9564627	NA	R, R, R, R, R, R		
Paclitaxel	CEU	67.1	67.1	NA	rs2216926	NA	R, R, R, R, R, R	MAB21L3 RP11-134G8.8
					rs10913404	NA	R, R, R, R, R, R	
					rs2116796	NA	R, R, R, R, R, R	
					rs28634858	2	WE, R, R, R, R, R	
					rs10773683	3	R, R, R, R, R, R	
YRI	87.8	57.0	19.0	rs10478863	NA	R, R, R, R, R, R		

**Table 3** (continued)

Drugs	Population	$R^2$			SNPs (LD)	DNaseI	Genome segmentation	Gene
		Integration	SNP	Expression				
					rs10094960	NA	R, R, R, R, R, R	MAPKBP1
					rs446139	NA	R, R, R, R, R, R	LPP
					rs9905351	8	T, T, T, R, R, R	
					rs28570663	NA	R, R, R, R, R, R	
					rs10948390	NA	T, T, R, R, R, R	

The testing  $R^2$  for integration models, SNP only models, and gene expression only models were obtained through cross validations. NA for gene expression only models were due to no expression variables being identified in the integration model. SNPs identified by the network as well as SNPs that are in LD with the network SNPs ( $r^2 > 0.7$ ) are shown. Genome segmentations abbreviations are: Enhancer (E), weak enhancer (WE), CTCF binding (CTCF), transcribed region (T), repressed region (R), transcription start site (TSS)



**Fig. 4** Neural Network model for capecitabine’s chemotherapeutic response in CEU. W is a weight node, PADD is an addition activation node. In this network, the SNP genotype values are multiplied with their associated weights and then added together in the PADD node

population. This suggests a possible alternate regulator of NF-kB pathway for cytotoxic response.

Many annotation terms were also associated in both populations. Fatty acid and triglyceride related functional terms were identified in GO term, KEGG pathway, and REACTOME (Fig. 2b, c, d). In Pfam, GNS1/SUR4 family is also involved in fatty acid elongation systems [49]. Fatty acid synthase is an important process for cancer cells to expand and proliferate and has been found to be associated with cytotoxicity in cells [50, 51]. High expression of fatty acid synthase was also observed in colon, prostate, ovarian,

breast and endometrium cancers [52, 53]. Recent studies have shown that fatty acid synthase inhibitors are potent therapeutic agents for cancer [54, 55]. Altered growth is one of the direct results of cytotoxic response, so it is likely that fatty acid synthase is also involved in the observed differential drug responses. Positive regulation of endothelial cell migration was associated with all five drugs. In addition, it was reported that during metastasis, cancer cells extravasate metastasis sites by attaching to endothelial cells [56]. We also observed drugs that were known to treat similar cancers have high overlap of biological annotations. In particular, cisplatin and carboplatin are both platinum compounds that treat lung, head and neck, testicular, and ovarian cancer [4, 9]. It can be seen that cisplatin and carboplatin have high overlap in all annotations, especially in the YRI population (Fig. 2).

LCLs’ cellular sensitivity to drugs is a broad phenotype that encompasses many sub-phenotypes including drug-induced apoptosis. Cell apoptosis, as measured by caspase activity, was shown to be weakly correlated with cytotoxicity [22]. Despite the weak correlation at the phenotypic level, we found that many functional terms enriched for cell cytotoxicity are also associated with cell apoptosis (Fig. 2), indicating shared biological mechanism for the two responses. As an example, SNPs in RFX2 gene were identified in a clinical trial evaluating paclitaxel-induced neuropathy of breast cancer patients and shown to be functionally important in paclitaxel-induced cytotoxicity using siRNA [36]. In our analysis, RFX DNA binding domain was associated with both paclitaxel-induced cytotoxicity and apoptosis (Fig. 2e).

The integration of SNP and gene expression data yielded higher predictive  $R^2$  than SNP or gene expression data alone (Table 3), which supports the potential value for combining multiple types of genomics data [57–59]. Because we prioritized our model based on overlaps with DNA regulatory regions, many of our models contain SNPs that are located in the DNaseI region and functional genome

segmentation regions. This information can provide additional interpretability to our models compared with using  $R^2$  alone.

Due to the small sample size of LCLs in the original analysis, we sought for replication in independent HapMap3 LCLs to confirm our result. Of note, we found a large number of biological annotations were replicated in the independent datasets. Of annotations/drug pairs identified in the discovery analysis, between 15–100% were also significant in the respective HapMap3 replication population. This confirms that the associated SNPs might not be identical between discovery and replication studies, but the underlying biological mechanisms are the same. Our results show that many genetic variants and genes are involved in chemotherapeutic drugs cytotoxicity. By mapping genetic variants to higher-level biological processes, we were able to encapsulate variants in the same genomic region into more informative units. Comparing biological processes groups showed population specific patterns between CEU and YRI. However, as CEU LCLs were derived from an earlier time point [60], further studies are needed to verify whether some of the observed differential patterns might be due to time in culture. Nonetheless, a previous study showed that the cellular proliferation rate was not significantly different between CEU and YRI and no widespread genetic differences on common SNPs were observed between phase 2 and phase 3 YRI LCLs [60]. Also, there are common processes across all drugs as well as between drugs that belong to the same class. These results could identify new drug repositioning candidates based on sharing of biological processes. Using the ENCODE data, we built predictive network models for drug cytotoxicity that are also functionally relevant. Future work could explore the use of additional types of functional data to better reflect the functional relevance of the models.

**Acknowledgements** This work is supported, in part, by NSF graduate fellowship DGE1255832 (RL), NIH HG006389 (MDR), NIH P50GM115318 (MDR), and NIH/NIGMS Pharmacogenomics of Anticancer Agents Research Grant U01 GM61393 (MED).

## Compliance with ethical standards

**Conflict of interest** The authors declare that they have no conflict of interest.

## References

1. FACT SHEET: investing in the National Cancer Moonshot | whitehouse.gov. 2016. <https://www.whitehouse.gov/the-press-office/2016/02/01/fact-sheet-investing-national-cancer-moonshot>.
2. Wheeler HE, Dolan ME. Lymphoblastoid cell lines in pharmacogenomic discovery and clinical translation. *Pharmacogenomics*. 2012;13:55–70.
3. Huang RS, Duan S, Kistner EO, Hartford CM, Dolan ME. Genetic variants associated with carboplatin-induced cytotoxicity in cell lines derived from Africans. *Mol Cancer Ther*. 2008;7:3038–46. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2743011&tool=pmcentrez&rendertype=abstract>.
4. Wheeler HE, Gamazon ER, Stark aL, O'Donnell PH, Gorsic LK, Huang RS, et al. Genome-wide meta-analysis identifies variants associated with platinating agent susceptibility across populations. *Pharm J*. 2011;13:35–43. <http://www.nature.com/doi/10.1038/tj.2011.38>.
5. Huang RS, Duan S, Bleibel WK, Kistner EO, Zhang W, Clark Ta, et al. A genome-wide approach to identify genetic variants that contribute to etoposide-induced cytotoxicity. *Proc Natl Acad Sci USA*. 2007;104:9758–63. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1887589&tool=pmcentrez&rendertype=abstract>.
6. Huang RS, Duan S, Shukla SJ, Kistner EO, Clark Ta, Chen TX, et al. Identification of genetic variants contributing to cisplatin-induced cytotoxicity by use of a genomewide approach. *Am J Hum Genet*. 2007;81:427–37. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1950832&tool=pmcentrez&rendertype=abstract>.
7. Borghaei H, Langer CJ, Millenson M, Ruth KJ, Litwin S, Tuttle H, et al. Phase II study of paclitaxel, carboplatin, and cetuximab as first line treatment, for patients with advanced non-small cell lung cancer (NSCLC). *J Thorac Oncol*. 2008;3:1286–92. <http://www.ncbi.nlm.nih.gov/pubmed/18978564>.
8. McWhinney SR, Goldberg RM, McLeod HL. Platinum neurotoxicity pharmacogenetics. *Mol Cancer Ther*. 2009;8:10–6. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2651829&tool=pmcentrez&rendertype=abstract>.
9. Rabik CA, Dolan ME. Molecular mechanisms of resistance and toxicity associated with platinating agents. *Cancer Treat Rev*. 2007;33:9–23. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1855222&tool=pmcentrez&rendertype=abstract>.
10. Cassidy J, Saltz L, Twelves C, Van Cutsem E, Hoff P, Kang Y, et al. Efficacy of capecitabine versus 5-fluorouracil in colorectal and gastric cancers: a meta-analysis of individual data from 6171 patients. *Ann Oncol*. 2011;22:2604–9. <http://www.ncbi.nlm.nih.gov/pubmed/21415237>.
11. Kumar CC. Genetic abnormalities and challenges in the treatment of acute myeloid leukemia. *Genes Cancer*. 2011;2:95–107. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3111245&tool=pmcentrez&rendertype=abstract>.
12. Rowinsky EK, Wright M, Monsarrat B, Donehower RC. Clinical pharmacology and metabolism of taxol (paclitaxel): update 1993. *Ann Oncol*. 1994;5:S7–16. <http://www.ncbi.nlm.nih.gov/pubmed/7865438>.
13. Huang RS, Kistner EO, Bleibel WK, Shukla SJ, Dolan ME. Effect of population and gender on chemotherapeutic agent-induced cytotoxicity. *Mol Cancer Ther*. 2007;6:31–6. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2669540&tool=pmcentrez&rendertype=abstract>.
14. Zhang W, Duan S, Kistner EO, Bleibel WK, Huang RS, Clark TA, et al. Evaluation of genetic variation contributing to differences in gene expression between populations. *Am J Hum Genet*. 2008;82:631–40. <http://www.sciencedirect.com/science/article/pii/S0002929708001365>.
15. Gamazon ER, Huang RS, Cox NJ, Dolan ME. Chemotherapeutic drug susceptibility associated SNPs are enriched in expression quantitative trait loci. *Proc Natl Acad Sci USA*. 2010;107:9287–92. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2889115&tool=pmcentrez&rendertype=abstract>.
16. Schork AJ, Thompson WK, Pham P, Torkamani A, Roddey JC, Sullivan PF, et al. All SNPs are not created equal: genome-wide association studies reveal a consistent pattern of enrichment among functionally annotated SNPs. *PLoS Genet*. 2013;9:e1003449.

- <http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1003449>.
17. Hindorf LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci USA*. 2009;106:9362–7. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2687147&tool=pmcentrez&rendertype=abstract>.
  18. O'Donnell PH, Stark AL, Gamazon ER, Wheeler HE, McIlwee BE, Gorsic L, et al. Identification of novel germline polymorphisms governing capecitabine sensitivity. *Cancer*. 2012;118:4063–73. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3413892&tool=pmcentrez&rendertype=abstract>.
  19. Abecasis GR, Auton A, Brooks LD, DePristo Ma, Durbin RM, Handsaker RE, et al. An integrated map of genetic variation from 1,092 human genomes. *Nature*. 2012;491:56–65. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3498066&tool=pmcentrez&rendertype=abstract>.
  20. Lappalainen T, Sammeth M, Friedländer MR, Höfer PaC, Monlong J, Rivas Ma, et al. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*. 2013;501:506–11. <http://www.ncbi.nlm.nih.gov/pubmed/24037378>.
  21. Hartford CM, Duan S, Delaney SM, Mi S, Kistner EO, Lamba JK, et al. Population-specific genetic variants important in susceptibility to cytarabine arabinoside cytotoxicity. *Blood*. 2009;113:2145–53. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2652364&tool=pmcentrez&rendertype=abstract>.
  22. Wen Y, Gorsic LK, Wheeler HE, Ziliak DM, Huang RS, Dolan ME. Chemotherapeutic-induced apoptosis: a phenotype for pharmacogenomics studies. *Pharm Genom*. 2011;21:476–88. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3134538&tool=pmcentrez&rendertype=abstract>.
  23. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernysky A, et al. The genome analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 2010;20:1297–303. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2928508&tool=pmcentrez&rendertype=abstract>.
  24. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*. 2007;81:559–75. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1950838&tool=pmcentrez&rendertype=abstract>.
  25. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet*. 2006;38:904–9. Epub 2006 Jul 23.
  26. Bush WS, Dudek SM, Ritchie MD. Biofilter: a knowledge-integration system for the multi-locus analysis of genome-wide association studies. *Pac Symp Biocomput*. 2009. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2859610&tool=pmcentrez&rendertype=abstract>.
  27. Aken BL, Ayling S, Barrell D, Clarke L, Curwen V, Fairley S, et al. The Ensembl gene annotation system. *Database*. 2016. <http://www.ncbi.nlm.nih.gov/pubmed/27337980>.
  28. Finn RD, Bateman A, Clements J, Coghill P, Eberhardt RY, Eddy SR, et al. Pfam: the protein families database. *Nucleic Acids Res*. 2014;42:D222–30. <https://academic.oup.com/nar/article-lookup/doi.org/10.1093/nar/gkt1223>.
  29. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The gene ontology consortium. *Nat Genet*. 2000;25:25–9. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3037419&tool=pmcentrez&rendertype=abstract>.
  30. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*. 2000;28:27–30. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=102409&tool=pmcentrez&rendertype=abstract>.
  31. Vastrik I, D'Eustachio P, Schmidt E, Joshi-Tope G, Gopinath G, Croft D, et al. Reactome: a knowledge base of biologic pathways and processes. *Genome Biol*. 2007;8:R39. <http://www.ncbi.nlm.nih.gov/pubmed/17367534>.
  32. Kim D, Li R, Dudek SM, Frase AT, Pendergrass SA, Ritchie MD. Knowledge-driven genomic interactions: an application in ovarian cancer. *BioData Min*. 2014;7:20. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4161273&tool=pmcentrez&rendertype=abstract>.
  33. Holzinger ER, Dudek SM, Frase AT, Krauss RM, Medina MW, Ritchie MD. ATHENA: a tool for meta-dimensional analysis applied to genotypes and gene expression data to predict HDL cholesterol levels. *Pac Symp Biocomput*. 2013. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3587764&tool=pmcentrez&rendertype=abstract>.
  34. Holzinger ER, Dudek SM, Frase AT, Pendergrass Sa, Ritchie MD. ATHENA: the analysis tool for heritable and environmental network associations. *Bioinformatics*. 2014;30:698–705. <http://www.ncbi.nlm.nih.gov/pubmed/24149050>.
  35. Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, et al. An integrated encyclopedia of DNA elements in the human genome. *Nature*. 2012;489:57–74. <https://doi.org/10.1038/nature11247>.
  36. Wheeler HE, Gamazon ER, Wing C, Njiaju UO, Njoku C, Baldwin RM, et al. Integration of cell line and clinical trial genome-wide analyses supports a polygenic architecture of paclitaxel-induced sensory peripheral neuropathy. *Clin Cancer Res*. 2013;19:491–9. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3549006&tool=pmcentrez&rendertype=abstract>.
  37. Ricci MS. Chemotherapeutic approaches for targeting cell death pathways. *Oncology*. 2006;11:342–57. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3132471&tool=pmcentrez&rendertype=abstract>.
  38. Brown JM, Attardi LD. The role of apoptosis in cancer development and treatment response. *Nat Rev Cancer*. 2005;5:231–7. <http://www.ncbi.nlm.nih.gov/pubmed/15738985>.
  39. Moen EL, Zhang X, Mu W, Delaney SM, Wing C, McQuade J, et al. Genome-wide variation of cytosine modifications between European and african populations and the implications for complex traits. *Genetics*. 2013;194:987–96. <http://www.ncbi.nlm.nih.gov/pubmed/23792949>.
  40. Quintela-Fandino M, Arpaia E, Brenner D, Goh T, Yeung FA, Blaser H, et al. HUNK suppresses metastasis of basal type breast cancers by disrupting the interaction between PP2A and cofilin-1. *Proc Natl Acad Sci USA*. 2010. 107(6):2622–7. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2823890&tool=pmcentrez&rendertype=abstract>.
  41. Sinilnikova OM, McKay JD, Tavtigian SV, Canzian F, DeSilva D, Biessy C, et al. Haplotype-based analysis of common variation in the acetyl-coA carboxylase alpha gene and breast cancer risk: a case-control study nested within the European prospective investigation into cancer and nutrition. *Cancer Epidemiol Biomark Prev*. 2007;16:409–15. <http://www.ncbi.nlm.nih.gov/pubmed/17372234>.
  42. Amend K, Hicks D, Ambrosone CB. Breast cancer in African-American women: differences in tumor biology from European-American women. *Cancer Res*. 2006;66:8327–30. <http://www.ncbi.nlm.nih.gov/pubmed/16951137>.
  43. Liu H, Yang Y, Xiao J, Yang S, Liu Y, Kang W, et al. Semaphorin 4D expression is associated with a poor clinical outcome in cervical cancer patients. *Microvasc Res*. 2014;93:1–8. <http://www.ncbi.nlm.nih.gov/pubmed/24603190>.
  44. Shen Y-M, He X, Deng H-X, Xie Y-P, Wang C-T, Wei Y-Q, et al. Overexpression of the hBiot2 gene is associated with development of human cervical cancer. *Oncol Rep*. 2011;25:75–80. <http://www.ncbi.nlm.nih.gov/pubmed/21109960>.

45. Rositch AF, Nowak RG, Gravitt PE. Increased age and race-specific incidence of cervical cancer after correction for hysterectomy prevalence in the United States from 2000 to 2009. *Cancer*. 2014;120:2032–8. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4073302&tool=pmcentrez&rendertype=abstract>.
46. Houston KA, Henley SJ, Li J, White MC, Richards TB. Patterns in lung cancer incidence rates and trends by histologic type in the United States, 2004–2009. *Lung Cancer*. 2014;86:22–8. <http://www.ncbi.nlm.nih.gov/pubmed/25172266>.
47. Israël A. The IKK complex, a central regulator of NF-kappaB activation. *Cold Spring Harb Perspect Biol*. 2010;2:a000158. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2829958&tool=pmcentrez&rendertype=abstract>.
48. Karin M. NF- $\kappa$ B as a Critical Link Between Inflammation and Cancer. *Cold Spring Harb Perspect Biol*. 2009;1:a000141. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2773649&tool=pmcentrez&rendertype=abstract>.
49. Oh CS, Toke DA, Mandala S, Martin CE. ELO2 and ELO3, homologues of the *Saccharomyces cerevisiae* ELO1 gene, function in fatty acid elongation and are required for sphingolipid formation. *J Biol Chem*. 1997;272:17376–84. <http://www.ncbi.nlm.nih.gov/pubmed/9211877>.
50. Pizer ES, Wood FD, Pasternack GR, Kuhajda FP. Fatty acid synthase (FAS): a target for cytotoxic antimetabolites in HL60 promyelocytic leukemia cells. *Cancer Res*. 1996;56:745–51. <http://www.ncbi.nlm.nih.gov/pubmed/8631008>.
51. Lupu R, Menendez JA. Pharmacological inhibitors of fatty acid synthase (FASN)—catalyzed endogenous fatty acid biogenesis: a new family of anti-cancer agents? *Curr Pharm Biotechnol*. 2006;7:483–93. <http://www.ncbi.nlm.nih.gov/pubmed/17168665>.
52. Kuhajda FP. Fatty-acid synthase and human cancer: new perspectives on its role in tumor biology. *Nutrition*. 2000;16:202–8. <http://www.ncbi.nlm.nih.gov/pubmed/10705076>.
53. Kuhajda FP, Pizer ES, Li JN, Mani NS, Frehywot GL, Townsend CA. Synthesis and antitumor activity of an inhibitor of fatty acid synthase. *Proc Natl Acad Sci USA*. 2000;97:3450–4. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=16260&tool=pmcentrez&rendertype=abstract>.
54. Angeles TS, Hudkins RL. Recent advances in targeting the fatty acid biosynthetic pathway using fatty acid synthase inhibitors. *Expert Opin Drug Discov*. 2016;11:1187–99. <https://www.tandfonline.com/doi/full/10.1080/17460441.2016.1245286>.
55. Zhang J-S, Lei J-P, Wei G-Q, Chen H, Ma C-Y, Jiang H-Z. Natural fatty acid synthase inhibitors as potent therapeutic agents for cancers: a review. *Pharm Biol*. 2016;209:1–7.
56. Reymond N, d'Água BB, Ridley AJ. Crossing the endothelial barrier during metastasis. *Nat Rev Cancer*. 2013;13:858–70. <https://doi.org/10.1038/nrc3628>.
57. Nicolae DL, Gamazon E, Zhang W, Duan S, Dolan ME, Cox NJ. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet*. 2010;6:e1000888. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2848547&tool=pmcentrez&rendertype=abstract>.
58. Wheeler HE, Aquino-Michaels K, Gamazon ER, Trubetskoy VV, Dolan ME, Huang RS, et al. Poly-omic prediction of complex traits: OmicKriging. *Genet Epidemiol*. 2014;38:402–15. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4072756&tool=pmcentrez&rendertype=abstract>.
59. Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D. Methods of integrating data to uncover genotype–phenotype interactions. *Nat Rev Genet*. 2015;16:85–97. <https://doi.org/10.1038/nrg3868>.
60. Stark AL, Zhang W, Zhou T, O'Donnell PH, Beiswanger CM, Huang RS, et al. Population differences in the rate of proliferation of international HapMap cell lines. *Am J Hum Genet*. 2010;87:829–33. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2997375&tool=pmcentrez&rendertype=abstract>.