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

Ruowang Li, Dokyoon Kim, Heather Wheeler, Scott Dudek, Eileen Dolan, Marylyn Ritchie

Abstract

Identifying genetic variants associated with chemotherapeutic induced cytotoxicity is an important step towards personalized treatment of cancer patients. However, annotating and interpreting the associated genetic variants remain 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 in multiple drugs. In this study, we used biological knowledge to annotate and compare genetic variants associated with chemotherapeutic drugs cisplatin, carboplatin, capecitabine, cytarabine, and paclitaxel's cytotoxicity in HapMap CEU and YRI lymphoblastoid cell lines. Using the higher-level annotations, we observed distinct biological modules that are associated with cell line populations as well as types of chemotherapeutic drugs. We also integrated genetic variants and gene expression variables to build predictive models for chemotherapeutic drug cytotoxicity and we prioritized the network models based on the enrichment of DNA regulatory data. By using biological knowledge and DNA regulatory information, we propose a novel approach for jointly analyzing associated genetic variants in multiple chemotherapeutic drugs.

Introduction

A better understanding of cytotoxicity associated with chemotherapeutic drugs can lead to more precise and personalized treatment of cancer patients. As genomic sequencing becoming more prevalent, understanding the genetic contribution of cytotoxicity could help patients to receive the most beneficial treatments while minimizing toxic effects. Lymphoblastoid cell lines (LCLs) have been established as a model system to study the genetic components of drug-induced cytotoxicity by measuring cytotoxicity through cell growth inhibition[1]. Previous genome-wide association analyses have identified numerous genetic variants and gene expression variables associated with drug cytotoxicity[2–5]. 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 drug cytotoxicity in five chemotherapeutic drugs: Cisplatin, Carboplatin, Capecitabine, Cytarabine, and Paclitaxel in two HapMap populations: Utah Residents with European ancestry (CEU) and African individuals from Yoruba (YRI). Cisplatin and carboplatin has been commonly used to treat head and neck, lung, and ovarian cancers[6–8]. Capecitabine is mainly used for colon and breast cancers[9]. Patients with leukemia have long been treated with Cytarabine[10]. Paclitaxel treats a wide range of cancers including lung, breast, and ovarian cancers[11]. Previous studies have shown that drugs in the same class have common genetic loci associated with drug induced cytotoxicity, for example, between cisplatin and carboplatin[3]. An individual's ancestral background has also been linked to differential risks for cytotoxicity[12]. A better 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: 1. There are a large number of genetic variations that can be associated with drug cytotoxicity. 2. Each variant is a surrogate for many other variants in the same region. To address these issues, we annotated genetic variants using higher-level biological annotations so that they can be grouped into more interpretable biological modules. Between CEU and YRI, we found population specific annotations across drugs. 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 Genome-Wide Association Studies (GWAS) to identify genetic variants that have the strongest independent genetic effects on cytotoxicity. Differences in gene expression levels have also been shown to affect cell's cytotoxicity response[13]. Combining information from SNPs and gene expression has been mainly through eQTL analysis[4]. While the eQTL method can capture a linear relationship between SNPs and gene expression, it omits 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 a 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[14,15]. To this end, we used DNAseI and genome segmentation data published by the ENCODE consortium to prioritize the network models. Our studies suggest that combining genetic and functional genomics information could be a useful approach for interpreting genetic factors susceptible for chemotherapeutic drug responses.

Results

Chemotherapeutic drug genetic associations

Cell growth inhibition was measured on unrelated CEU and YRI LCLs following the treatments of each chemotherapeutic drug: Cisplatin, Carboplatin, Cytarabine, Capecitabine, and Paclitaxel. Increasing concentrations of each drug were applied to between 29 and 77 LCLs and their dose-dependent inhibition was calculated as IC_{50}, concentration required to stop 50% of cell growth, or AUC, area under the curve.

Genome-wide SNP data for the LCLs were obtained from the 1000 Genomes Project and were evaluated for their association with each drug's cytotoxicity. Gender and significant principal components of ancestry (2 or 3) were adjusted for in the linear regression model. We identified between 1,230 and 2,749 SNPs significantly associated with each drug response, respectively ($P < 0.0005$) (Table 1). Gene expressions of the LCLs measured by RNA-Seq were downloaded from the gEUVADIS consortium (http://www.geuvadis.org/). Around 20,000 genes' normalized RPKM (reads per kilobase per million) values were tested for association with each drug's IC_{50} or AUC. To keep the number of associated genes were similar across drugs, we used $P < 0.005$ or $P < 0.0005$ to select candidate genes. We identified between 65 and 295 genes whose expression levels were associated with drug outcome (Table 1). A list of associated SNPs and gene expression can be found in supplement materials (Table S1).

Table 1. Genotype and gene expression associations with chemotherapeutic drugs

Drugs	Population	Discovery LCLs	Discovery Associated SNPs	Discovery Associated Expression	Hapmap 3 Replication LCLs	Replicated SNPs
Cisplatin	CEU	72	1945	121	40	324
	YRI	77	2157	76*	46	270
Carboplatin	CEU	72	2530	169*	40	304
	YRI	75	2364	194	44	248
Cytarabine	CEU	72	2156	126	40	276
	YRI	77	2749	106*	46	725
Capecitabine	CEU	73	2014	65*	40	137
	YRI	76	2485	295	46	306
Paclitaxel	CEU	29	1230	94	NA	NA
	YRI	29	1466	80	NA	NA

* denotes p<0.0005

To replicate the SNP associations, we applied chemotherapeutic drugs on an independent set of HapMap phase 3 LCLs. Following the same protocol, LCLs were treated with four of the five drugs: cisplatin, carboplatin, cytarabine, and capecitabine. We performed an association analysis on the independent LCLs and we replicated between 137 and 725 SNPs that were associated in the original samples (Table 1).

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 the SNPs that are associated with each drug response using gene regions, KEGG pathways, GO terms, REACTOME, and protein families (Pfam) using Biofilter[16]. 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 (1000x) for each drug's IC_{50} or AUC. Using the permuted IC_{50} 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 ($P < 0.005$).

A related measurement of chemotherapeutic-induced cytotoxicity, cell lines' apoptosis, is one of the manifestations of cell growth inhibition[17]. We treated LCLs with a subset of the drugs (cytarabine, cisplatin, and paclitaxel) and measured the cell lines' apoptosis (Table 2). We identified SNPs that are associated with apoptosis (results not shown) 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 (Figure 1).

Table 2. Apoptosis phenotype measured on LCLs

Drug	**Population**	**Sample size**
Cytarabine 5uM	CEU	30
	YRI	35
Cytarabine 40uM	CEU	30
	YRI	35
Cisplatin 5uM	CEU	30
	YRI	35
Paclitaxel 12.5nM	CEU	30
	YRI	35

When we compared the associated functional annotations across different populations, we observed some annotations are population specific. For gene annotations, a group of genes including *HUNK*, *MTMR9*, P*RAMEF4*, and *ACACA* were only associated in the CEU population (Figure 1a). Meanwhile, *Spermatogenesis family BioT2*, *GNS1/SUR4 family*, *Translin family*, and *Leukotriene A4 hydrolase C-terminal* in pfam, *IKK* related terms in REACTOME, and several neuronal development and leukocytes GO terms 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.

Within each population, we observed that some drugs have similar associated annotation patterns. In particular, cisplatin and carboplatin have many functional annotations in common. Cytarabine and capcitabine have a number of overlapped annotations (Figure 1).

We also observed overlapping annotations between drug cytotoxicity and apoptosis. *TSNAX-DISC1* and *DISC1* 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.

Figure 1. Pan-drug analysis of functional annotations

a. Gene

b. GO term

c. KEGG pathway

d. REACTOME

e. Pfam

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). Functional terms were grouped using hierarchical clustering according to its enrichment across drugs and populations.

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

Starting with the SNPs and gene expression 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. Figure 2 shows an example genetic network for drug cytotoxicity.

Figure 2. An example GENN network model. W is a weight node, PADD is an addition activation node.

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 more functionally important variables. Previous studies suggest that SNPs that lie in the open chromatin and regulatory regions are more likely to be functional[18]. 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 chromatins 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 individual score for each feature normalized by network size (Figure 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).

Figure 3. Schematic for functional score calculation

Table 3. Network model identified by GENN

Drugs	Population	R^2			SNPs (LD)	DNAsel	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	
					rs2570317	NA	R, R, R, R, R, R	
	YRI	64.3	64.3	NA	rs11204113	NA	R, R, R, R, R, R	NA
					rs10760086	NA	R, R, R, R, R, R	
					rs9303059	NA	R, R, R, R, R, R	
					rs9661131	NA	R, R, R, R, R, R	
					rs6671214	NA	CTCF, CTCF, CTCF, CTCF, CTCF, R	
Carboplatin	CEU	60.4		23.1	rs11233413	9	T, E	TMEM14E
					rs12816395	NA	T, T, R, R, R, R	
					rs79062064	NA	T, T, R, R, R, R	
	YRI	66.2	66.2	NA	rs16823342	NA	R, R, R, R, R, R	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	
Cisplatin	CEU	66.6	46.3	14.0	rs11715866	NA	T, R, R, R, R, R	FABP6 HCFC1 TAS2R30 ZNF192P1
					rs344946	NA	R, R, R, R, R, R	
					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	
					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	
	YRI	52.4	36.8	12.7	rs12255911	NA	T, T, T, T, R, R	IL27
					rs6814234	9	WE, T, T, R, R	
					rs10426529	NA	E, R, R, R, R, R	
Cytarabine	CEU	47.7	42.9	0	rs1281461	NA	R, R, R, R, R, R	RP11-463J10.3 IL11RA
					rs2780788	NA	T, R, R, R, R, R	
					rs593525	11	T, T, T	
					rs4910512	2	T, R, R, R	
					rs7962806	NA	R, R, R, R, R, R	
	YRI	72.2	28.2	45.4	rs7666224,	NA	R, R, R, R, R, R	MAB21L3 RP11-134G8.8
					rs9564627	NA	R, R, R, R, R, R	
					rs2216926	NA	R, R, R, R, R, R	
					rs10913404	NA	R, R, R, R, R, R	
Paclitaxel	CEU	67.1	67.1	NA	rs2116796	NA	R, R, R, R, R, R	NA
					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	MAPKBP1 LPP
					rs10094960	NA	R, R, R, R, R, R	
					rs446139	NA	R, R, R, R, R, R	
					rs9905351	8	T, T, T, R, R, R	
					rs28570663	NA	R, R, R, R, R, R	
					rs10948390	NA	T, T, R, R, R, R	

For each drug and population, we listed R^2 for integration, snp, and gene expression model. Genome segmentations abbreviations are: Enhancer (E), weak Enhancer (WE), CTCF binding (CTCF), transcribed region (T), repressed region (R), transcription start site (TSS)

Discussion

An important contribution of understanding the genetic susceptibility of chemotherapeutic drugs is that we can more precisely utilize the drugs based on patients' genetic information. Most previous studies have evaluated the genotype associations to an individual chemotherapeutic drug; however, a comparative study of multiple drugs in multiple populations could reveal different mechanisms in drug-induced response. Here, we have jointly analyzed the genetic associations of chemotherapeutic drugs induced cytotoxicity for Cisplatin, Carboplatin, Capecitabine, Cytarabine, and Paclitaxel in CEU and YRI populations.

We performed genome-wide SNP association analysis for each drug in both populations to identify the significant genetic associations. A major challenge to interpret the 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. To this end, we 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 individuals of European and African ancestry. Interestingly, ancestry has also been reported to affect gene expression[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 (Figure 1a). Previous report has shown that differences exist between African Americans and European American women in the nature of breast cancer[19]. *SEMA4D* and *CCDC7* genes were associated in the YRI population (Figure 1a). Expressions of the genes have been reported to correlate with poor outcome in cervical and lung cancer. In addition, a recent survey has found that African Americans are more likely to develop cervical and lung cancer[20]. These candidate genes could be further validated in their respective population. Several *IKK* related REACTOME pathways were associated with YRI population (Figure 1d). *IKK* is a central regulator of NF-kB pathway[21] and activation of NF-kB pathway has been observed in many solid tumors[22]. Interestingly, NF-kB pathway is associated in both CEU and YRI population (Figure 1c), but *IKK* is only associated with the YRI 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 (Figure 1b,c, d). In Pfam, GNS1/SUR4 family is also involved in fatty acid elongation systems[23]. Fatty acid synthase is an important process for cancer cells to expand and proliferate. High expression of fatty acid synthase was observed in colon, prostate, ovary, breast and endometrium cancers[24,25]. 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 5 drugs. In addition, it was reported that during metastasis, cancer cells extravasate metastasis sites by attaching to endothelial cells[26]. 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, and ovarian cancer[3]. It can be seen that cisplatin and carboplatin have high overlap in all annotations, especially in the YRI population (Figure 1). *IRF4* gene, a known factor in hematological malignancies[27], is associated for cisplatin and carboplatin in YRI population. Previous reports have shown that both cisplatin and carboplatin are effective treatments for hematological malignancies[28,29].

LCLs treated with chemotherapeutic drugs can also result in apoptosis. We found that many functional terms enriched for cell cytotoxicity are also associated with cell apoptosis (Figure 1), indicating shared biological mechanism for the two responses.

The integration of SNP and gene expressions data yielded higher predictive R^2 than SNP or gene expression data alone (Table 3), indicating potential value for combining multiple types of genomics data. 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 add additional interpretability to our models compared with using R^2 alone.

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. 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. We also built predictive network models for drug cytotoxicity that are also functionally relevant. Future work can include additional types of functional data to better reflect the functional relevance of the models.

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)[30]. RNAseq gene expressions on the same individuals were downloaded from the gEUVADIS project[31].The gene expression data were 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[31].

Cytotoxicity data

Lymphoblastoid cell lines from HapMap phase 1 YRI and CEU populations were treated with capecitabine[32], carboplatin[2], cisplatin[5], cytarabine[33], and paclitaxel[17] as previously reported. For carboplatin and cisplatin, their IC_{50}, concentration required to stop 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. All AUC values were also log2 transformed to follow the normal distribution. For replication studies, HapMap phase 3 YRI and CEU cell lines were treated with four of the drugs: capecitabine, carboplatin, cisplatin, and cytarabine.

Quality control for genetic variants and gene expression data

SNP data were first transformed into a variant call format (VCF) 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[34]. 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 2 representing samples. 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 4).

Table 4. SNP and gene expression quality control (QC)

Drugs	Population	SNP QCed (million)	Expression QCed
Cisplatin	CEU	3.87	19,919
	YRI	4.69	20,380
Carboplatin	CEU	3.87	19,923
	YRI	4.64	20,427
Cytarabine	CEU	3.87	19,911
	YRI	4.68	20,380
Capecitabine	CEU	3.88	19,859
	YRI	4.66	20,421
Paclitaxel	CEU	2.71	19,683
	YRI	2.99	20,045

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 variants 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 pruned using PLINK software[35]. The principal components of the pruned SNP data were estimated using Eigenstrat[36]. Along with individual's gender, 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, we used Biofilter (v2.2)[16] to separately map the associated SNPs of each cytotoxicity phenotype to functional groups including genes regions, Pfam, GO term, KEGG pathway, and Reactome. 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 one thousand permutation tests, where we permuted each drug's cytotoxicity and performed GWA 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[37,38]. The software and its modeling processes have been described previously[39]. 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

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

variations explained by SNPs and gene expressions. We scaled the cytotoxicity to be between 0 and 1 using min-max scaling so that it matches the output of neural networks, where

And the R^2 is calculated as:

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

* D_i is the value of cytotoxicity for the i_{th} sample

Linkage disequilibrium patterns exist in the associated SNPs because many are proximately located. Even though they 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 resulted from LD, for each cytotoxicity phenotype, pairwise LD among all associated SNPs were estimated. r2 > 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. One gene from each cluster was selected as the tag gene for the cluster.

We first used ATHENA to perform variable selections 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 included in a minimum of 2 out of 5 models built from different cross validations. The variable selection step did not take into consideration of the testing R squared to avoid over-fitting. Using the selected SNPs and gene expressions, we used ATHENA to build five models, one for each cross validation, for each cytotoxicity phenotype.

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[18] 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/). The data contains merged DNAse-I peaks from UW and Duke that passed FDR 1% cutoff. Genome segmentations of six ENCODE cell lines was obtained from (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 splits 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 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 case when SNPs in LD with the tagging SNP has 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 prediction accuracy (R^2). Of those, we selected the model with highest functional score as the final model. Once we have the final model, we used SNPs and gene expressions to separately built SNP and gene expression only models. In case the models have negative R^2 value, the R^2 value was replaced with 0. The mean of testing R^2s for snp and gene expressions models were shown in Table 3.

Reference

1. Wheeler, H. E. & Dolan, M. E. Lymphoblastoid cell lines in pharmacogenomic discovery and clinical translation. **13,** 55–70 (2012).

2. Huang, R. S., Duan, S., Kistner, E. O., Hartford, C. M. & Dolan, M. E. Genetic variants associated with carboplatin-induced cytotoxicity in cell lines derived from Africans. *Mol. Cancer Ther.* **7,** 3038–46 (2008).

3. Wheeler, H. E. *et al.* Genome-wide meta-analysis identifies variants associated with platinating agent susceptibility across populations. *Pharmacogenomics J.* **13,** 35–43 (2011).

4. Huang, R. S. *et al.* A genome-wide approach to identify genetic variants that contribute to etoposide-induced cytotoxicity. *Proc. Natl. Acad. Sci. U. S. A.* **104,** 9758–63 (2007).

5. Huang, R. S. *et al.* Identification of genetic variants contributing to cisplatin-induced cytotoxicity by use of a genomewide approach. *Am. J. Hum. Genet.* **81,** 427–37 (2007).

6. Borghaei, 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.* **3,** 1286–1292 (2008).

7. McWhinney, S. R., Goldberg, R. M. & McLeod, H. L. Platinum neurotoxicity pharmacogenetics. *Mol. Cancer Ther.* **8,** 10–6 (2009).

8. Rabik, C. A. & Dolan, M. E. Molecular mechanisms of resistance and toxicity associated with platinating agents. *Cancer Treat. Rev.* **33,** 9–23 (2007).

9. Cassidy, J. *et al.* Efficacy of capecitabine versus 5-fluorouracil in colorectal and gastric cancers: a meta-analysis of individual data from 6171 patients. *Ann. Oncol.* **22,** 2604–9 (2011).

10. Kumar, C. C. Genetic abnormalities and challenges in the treatment of acute myeloid leukemia. *Genes Cancer* **2,** 95–107 (2011).

11. Rowinsky, E. K., Wright, M., Monsarrat, B. & Donehower, R. C. Clinical pharmacology and metabolism of Taxol (paclitaxel): update 1993. *Ann. Oncol.* **5 Suppl 6,** S7–16 (1994).

12. Huang, R. S., Kistner, E. O., Bleibel, W. K., Shukla, S. J. & Dolan, M. E. Effect of population and gender on chemotherapeutic agent-induced cytotoxicity. *Mol. Cancer Ther.* **6,** 31–6 (2007).

13. Zhang, W. *et al.* Evaluation of genetic variation contributing to differences in gene expression between populations. *Am. J. Hum. Genet.* **82,** 631–40 (2008).

14. Schork, A. J. *et al.* All SNPs are not created equal: genome-wide association studies reveal a consistent pattern of enrichment among functionally annotated SNPs. *PLoS Genet.* **9,** e1003449 (2013).

15. Hindorff, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. U. S. A.* **106,** 9362–7 (2009).

16. Bush, W. S., Dudek, S. M. & Ritchie, M. D. Biofilter: a knowledge-integration system for the multi-locus analysis of genome-wide association studies. *Pac. Symp. Biocomput.* 368–79 (2009). at <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2859610&tool=pmcentrez&rendertype=abstract>

17. Wen, Y. *et al.* Chemotherapeutic-induced apoptosis: a phenotype for pharmacogenomics studies. *Pharmacogenet. Genomics* **21,** 476–88 (2011).

18. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).

19. Amend, K., Hicks, D. & Ambrosone, C. B. Breast cancer in African-American women: differences in tumor biology from European-American women. *Cancer Res.* **66,** 8327–30 (2006).

20. Rositch, A. F., Nowak, R. G. & Gravitt, P. E. Increased age and race-specific incidence of cervical cancer after correction for hysterectomy prevalence in the United States from 2000 to 2009. *Cancer* **120,** 2032–8 (2014).

21. Israël, A. The IKK complex, a central regulator of NF-kappaB activation. *Cold Spring Harb. Perspect. Biol.* **2,** a000158 (2010).

22. Karin, M. NF- B as a Critical Link Between Inflammation and Cancer. *Cold Spring Harb. Perspect. Biol.* **1,** a000141–a000141 (2009).

23. Oh, C. S., Toke, D. A., Mandala, S. & Martin, C. E. ELO2 and ELO3, homologues of the Saccharomyces cerevisiae ELO1 gene, function in fatty acid elongation and are required for sphingolipid formation. *J. Biol. Chem.* **272,** 17376–84 (1997).

24. Kuhajda, F. P. Fatty-acid synthase and human cancer: new perspectives on its role in tumor biology. *Nutrition* **16,** 202–8 (2000).

25. Kuhajda, F. P. *et al.* Synthesis and antitumor activity of an inhibitor of fatty acid synthase. *Proc. Natl. Acad. Sci. U. S. A.* **97,** 3450–4 (2000).

26. Reymond, N., d'Água, B. B. & Ridley, A. J. Crossing the endothelial barrier during metastasis. *Nat. Rev. Cancer* **13,** 858–70 (2013).

27. Wang, L., Yao, Z. Q., Moorman, J. P., Xu, Y. & Ning, S. Gene expression profiling identifies IRF4-associated molecular signatures in hematological malignancies. *PLoS One* **9,** e106788 (2014).

28. Vogler, W. R. High-dose carboplatin in the treatment of hematologic malignancies. *Oncology* **50 Suppl 2,** 42–6 (1993).

29. Velasquez, W. S. *et al.* Effective salvage therapy for lymphoma with cisplatin in combination with high-dose Ara-C and dexamethasone (DHAP). *Blood* **71,** 117–22 (1988).

30. Abecasis, G. R. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).

31. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501,** 506–11 (2013).

32. O'Donnell, P. H. *et al.* Identification of novel germline polymorphisms governing capecitabine sensitivity. *Cancer* **118,** 4063–73 (2012).

33. Hartford, C. M. *et al.* Population-specific genetic variants important in susceptibility to cytarabine arabinoside cytotoxicity. *Blood* **113,** 2145–53 (2009).

34. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20,** 1297–303 (2010).

35. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81,** 559–75 (2007).

36. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. **38,** 904–909 (2006).

37. Kim, D. *et al.* Knowledge-driven genomic interactions: an application in ovarian cancer. *BioData Min.* **7,** 20 (2014).

38. Holzinger, E. R. *et al.* ATHENA: a tool for meta-dimensional analysis applied to genotypes and gene expression data to predict HDL cholesterol levels. *Pac. Symp. Biocomput.* 385–96 (2013). at <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3587764&tool=pmcentrez&rendertype=abstract>

39. Holzinger, E. R., Dudek, S. M., Frase, A. T., Pendergrass, S. a & Ritchie, M. D. ATHENA: the analysis tool for heritable and environmental network associations. *Bioinformatics* **30,** 698–705 (2014).