Research Article

For reprint orders, please contact: reprints@futuremedicine.com

Pharmacogenomics

Evaluating the role of admixture in cancer therapy via *in vitro* drug response and multivariate genome-wide associations

Aim: We investigate the role of ethnicity and admixture in drug response across a broad group of chemotherapeutic drugs. Also, we generate hypotheses on the genetic variants driving differential drug response through multivariate genome-wide association studies. **Methods:** Immortalized lymphoblastoid cell lines from 589 individuals (Hispanic or non-Hispanic/Caucasian) were used to investigate dose-response for 28 chemotherapeutic compounds. Univariate and multivariate statistical models were used to elucidate associations between genetic variants and differential drug response as well as the role of ethnicity in drug potency and efficacy. **Results & Conclusion:** For many drugs, the variability in drug response appears to correlate with self-reported race and estimates of genetic ancestry. Additionally, multivariate genome-wide association analyses offered interesting hypotheses governing these differential responses.

Keywords: admixture • cancer • Caucasian • cytotoxicity • drug • genome-wide association study • GWAS • hispanic • *in vitro* • lymphoblastoid cell lines

John Jack*[1,2], Tammy M Havener[3], Howard L McLeod[4,5], Alison A Motsinger-Reif[1,2] & Matthew Foster[6]

[1]Department of Statistics, North Carolina State University, 2601 Stinson Drive, Raleigh, NC 27695, USA

[2]Bioinformatics Research Center, North Carolina State University, 2601 Stinson Drive, Raleigh, NC 27695, USA

[3]Center for Pharmacogenomics & Individualized Therapy, University of North Carolina, 120 Mason Farm Road, Chapel Hill, NC 27514, USA

[4]DeBartolo Family Personalized Medicine Institute, Moffitt Cancer Center, 12902 Magnolia Drive, Tampa, FL 33612, USA

[5]Pharmacogenetics for Every Nation Initiative, 1119 Oxbridge Drive, Tampa, FL 33549, USA

[6]Lineberger Comprehensive Cancer Center, University of North Carolina, 101 Manning Drive, Chapel Hill, NC 27514, USA

*Author for correspondence:
Tel.: +1 919 515 1398
Fax: +1 919 515 7315
john.jack@ncsu.edu

Pharmacogenomics research has increased the understanding of genetic variability and its relationship to individual drug response [1]. It is well established that the majority of drugs/doses to combat diseases and disorders do not exhibit uniform effects for individuals. In particular, chemotherapeutic agents often have a narrow therapeutic range, which blurs the line between treatment and inefficacy or drug-related toxicity [2,3]. While the interindividual variation in drug response is undeniable, the proportion of this variation that is due to genetic variation is an open question. With family-based design cohorts available, it is unusual to be able to formally assess the heritability of differential drug response [4]. Pharmacogenomics investigation often relies on indirect evidence of a genetic etiology, such as differential response in animal models, and racial differences in clinical outcomes.

For antineoplastic therapies, there is clear evidence of racial disparities in outcome. While a portion of this variation is undoubtedly due to socioeconomic factors, studies with consistent and high access to therapy and adherence across racial and ethnic subgroups have also demonstrated disparate outcomes [5]. Most pharmacogenomics studies of genetic determinants of cancer disparities have focused on racial differences – between subjects of African versus European ancestry. Genetic determinants of disparities between ethnicities (e.g., Hispanic versus non-Hispanic Caucasian) are less well investigated. One reason for this paucity is that ethnic groups are typically defined by common geography and culture rather than ancestry. Such groups are typically highly admixed, featuring recent ancestry from multiple continental populations. However, with the refinement of high-throughput genomic methods, the potential to understand the relationships between complex diseases and admixed populations is becoming possible.

For example, admixture mapping in African–Americans successfully identified the genetic locus responsible for a form of neu-

Future Medicine part of fsg

tropenia [6]. It has also been implicated in finding important variations linked to prostate cancer [7,8], hypertension [9] and renal disease [10]. However, highly admixed ethnic groups such as the Hispanic population have offered greater challenges, despite well-described disparities in outcome. To this point, the well-described [11] higher rates of relapse and poorer survival of Hispanic children with acute lymphoblastic leukemia (ALL) have recently been linked to percentage of Native American ancestry using principal component analysis (PCA) [12]. Furthermore, genome-wide association studies (GWASs) have linked polymorphisms in the *ARID5B* gene to both elevated risk of developing ALL and risk of relapse after multiagent chemotherapy in this population [13,14]. Despite these advances, whether these genetic associations are responsible for variations in response to particular antineoplastic agents remains unclear. As ALL chemotherapy regimens typically contain seven or more chemotherapeutic agents, the contribution of any individual drug to genetically determined disparities is unlikely to be determined in analyses of clinical trials. For this reason, preclinical models of genetically determined drug susceptibility are critical to inform clinical investigations of these disparities.

While there are a number of approaches for gene mapping in pharmacogenomics, cell line models have emerged as a promising model system. As recently reviewed in [4], Epstein–Barr virus (EBV) immortalized lymphoblastoid cell lines (LCLs) have been used to demonstrate the heritability of dose response [15] and performing association mapping for pharmacogenomics [16,17] and toxicogenomics [18]. Additionally, the association mapping results can be functionally tested using knock-down experiments [19,20] or candidate gene analyses in clinical outcomes [21,22].

In the current study, we use cytotoxicity dose-response data on 28 anticancer therapeutic agents with samples from two distinct cohorts (Hispanic and non-Hispanic Caucasian) to enable understanding of two questions: are there global differences in cytotoxic response between the two ethnically-different cohorts at a high level? Are these ethnic differences robust at finer levels of stratification? Additionally, we used genome-wide association mapping to look for genetic variants that are associated with dose response – both in the Hispanic cohort and in a joint analysis combining the Hispanic and non-Hispanic Caucasian cohorts.

Materials & methods

Study subjects

Genotype data and drug response phenotypes were collected from two sources: (i) non-Hispanic Caucasian patients from the Pharmacogenomics and Risk of Cardiovascular Disease clinical trial study at the Children's Hospital of Oakland Research Institute (CHORI) at Oak Ridge [17] and (ii) Hispanic Mexican–American individuals from the Human Variation Panel (HVP) of the International HapMap consortium [23]. For (i), genotype data were collected for each of the 500 patients using one of two technologies – 314,621 or 620,901 markers using HumanHap300 BeadChip or HumanQuad610 BeadChip platforms, respectively, as previously described in detail in [17]. The phenotypic data on dose-dependent, drug-induced cytotoxicity was measured in LCLs derived from each subject. For (ii), genotype data was collected on 909,623 SNPs from 400 individuals (167 males, 233 females) across four self-reported populations: African–Americans in North America (n = 100), Caucasians in North America (n = 100), Han Chinese in Los Angeles (n = 100) and Mexicans in Los Angeles (n = 100). Of the 100 Mexican individuals, phenotypic data on dose-dependent, drug-induced cytotoxicity was measured in LCLs derived from 93 of those individuals.

Genotyping & quality controls

Genotyping data collection and quality control (QC) from the CHORI dataset was described previously elsewhere [17]. Identical QC methods were applied to the HVP in order to facilitate comparison of the two distinct cohorts as well as subsequent combined analyses for association mapping. These data were processed using the PLINK software package v1.07 [24]. Markers were excluded for the following conditions: significant deviation from Hardy–Weinberg equilibrium (HWE) via exact tests ($\alpha < 10^{-5}$), small minor allele frequencies (MAF <5%), missing data per individual (>5%) or missing data per SNP (>5%). Subsequently, for HVP, 32,697 markers were excluded by the HWE filter, 151,925 markers were excluded for low MAF, 28,658 markers were excluded from SNP missingness and 0 individuals were excluded for missingness. Additionally, mitochondrial, X, Y and unknown chromosome markers were dropped from analysis. After pruning, there were 167 males and 233 females with genotype information for 677,966 SNPs for the HVP dataset.

Next, we removed all samples from HVP except those with phenotypic data. The remaining 93 individuals all reported Hispanic (Mexican) ethnicity. To evaluate cryptic relatedness, identity by state/identity by descent (IBS/IBD) and inbreeding coefficients estimates were calculated with PLINK. To assess population stratification, PCA was performed with Eigenstrat via smartpca v8000 [25]. Four individuals were identified as potentially related, and were removed from the dataset to prevent confounding. Two samples were

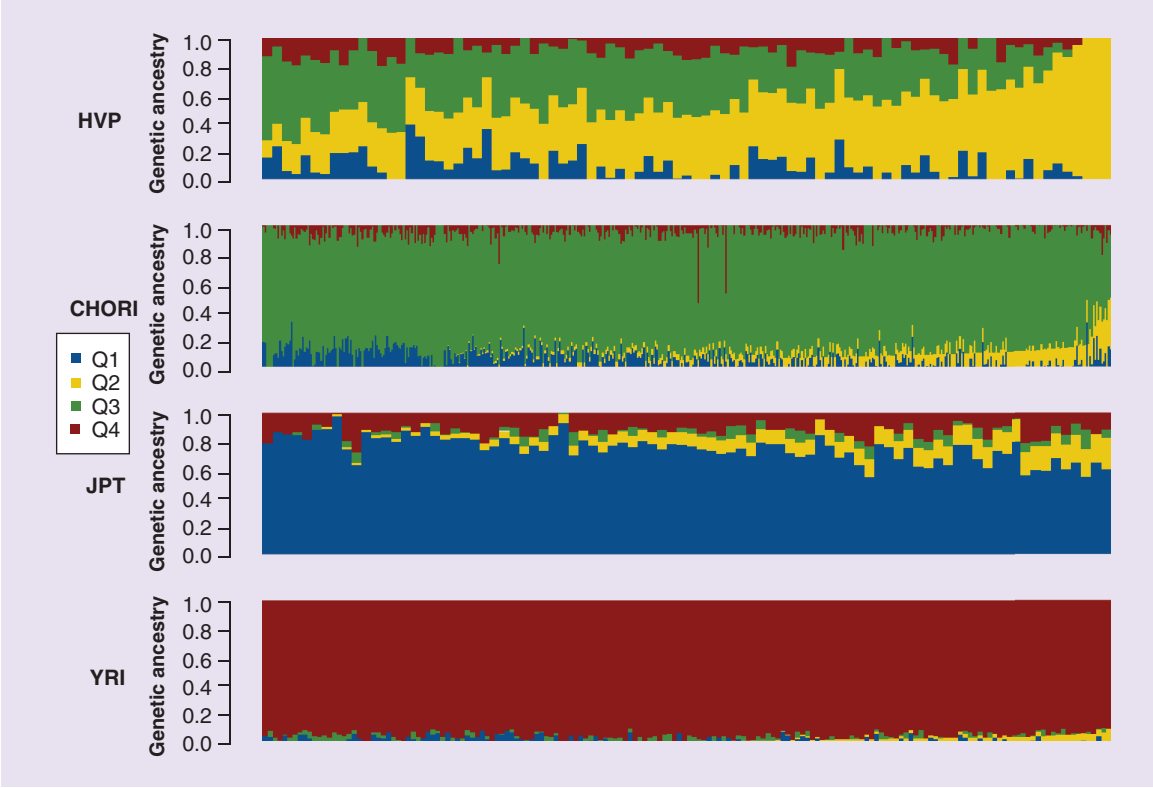


Figure 1. Evaluation of genetic ancestry. Unsupervised analysis performed with ADMIXTURE across the entire set of individuals for 202 Ancestral Informative Markers. The results are split up into one plot per self-reported group. Red indicates African ancestry, blue indicates Asian ancestry, green indicates Caucasian ancestry and yellow indicates Native American ancestry.

identified as outliers in PC3, and they were flagged and removed due to high P(IBD = 1) = 0.2759. An additional sample was flagged and removed as an outlier in PC2. Finally, another sample was flagged and removed for having both a high inbreeding coefficient and a high P (IBD = 1). Ultimately, 89 samples from HVP/Hispanic ethnicity and 500 samples from CHORI/non-Hispanic Caucasian ethnicity passed genotypic QC.

Phenotyping & QC

Epstein–Barr virus immortalized LCLs were obtained from two sources: 93 HVP commercially available cell lines from Coriell Cell Repositories (NJ, USA) and 500 CHORI cell lines were the generous gift from the lab of Ronald Krauss at CHORI. Dose-dependent cytotoxicity data were collected across 28 drugs for each cell line. The study design and QC pipeline for the CHORI phenotypic data were described elsewhere in detail [17]. We will summarize the phenotyping and QC methods for the 93 HVP samples, and refer the reader to [17] for specific details on CHORI.

All LCLs were cultured in RPMI medium 1640 containing 2 mM L-glutamine (Gibco, Life Technologies, NY, USA) and 15% fetal bovine serum (Sigma-Aldrich Corp, MO, USA) at 37°C, 5% CO_2. There were no media antibiotics used. Using 384 well plates, individual cell lines were seeded with approximately 4000 cells/well. Moreover, each 384 well plate contained LCLs from a single individual cell line. Two plate formats were used to capture six concentrations for each of the 28 drugs with replication; 14 drugs on the first plate format and 14 drugs on the second plate format. The list of drugs and concentrations is summarized in Supplementary Table 1. In order to assess the interplate reproducibility and variability, each sample was assayed on the first and second plate formats twice with each replicate plate corresponding to a different laboratory day.

Each plate includes controls for background noise and drug solvation effects. Background noise was estimated from viability measurements for the LCLs treated with a lethal dose of 10% DMSO. The drug solvation effects were determined by the reduction in cell viability due to the treatment with drug solvent (DMSO, in low dosage). A series of viability readings were used for LCLs exposed (only) to DMSO at different concentrations: 0% (H_20), 0.01, 0.1, 1 and 2% DMSO. Every exposure scenario (for all controls and drug concentrations) for each plate were performed

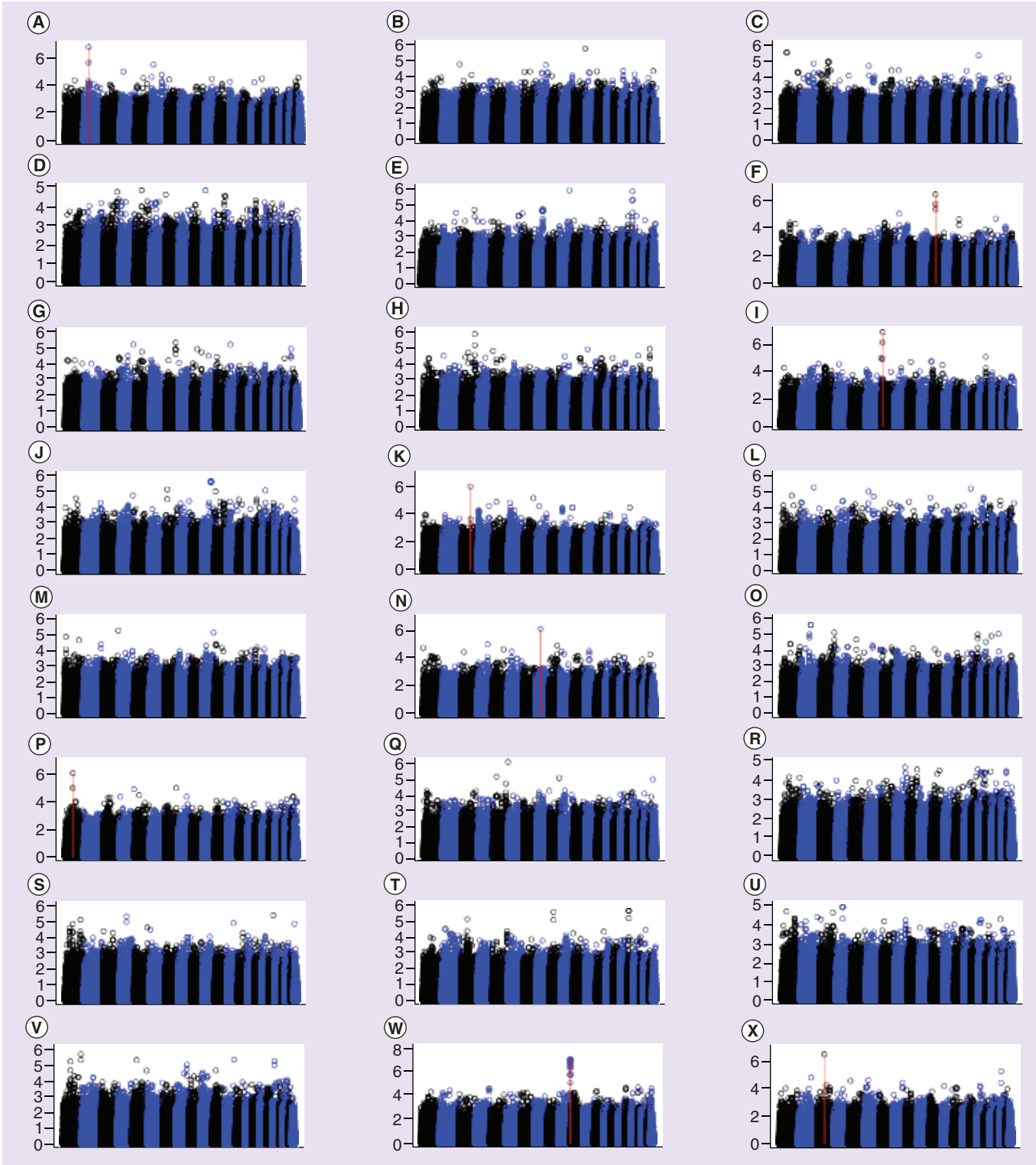
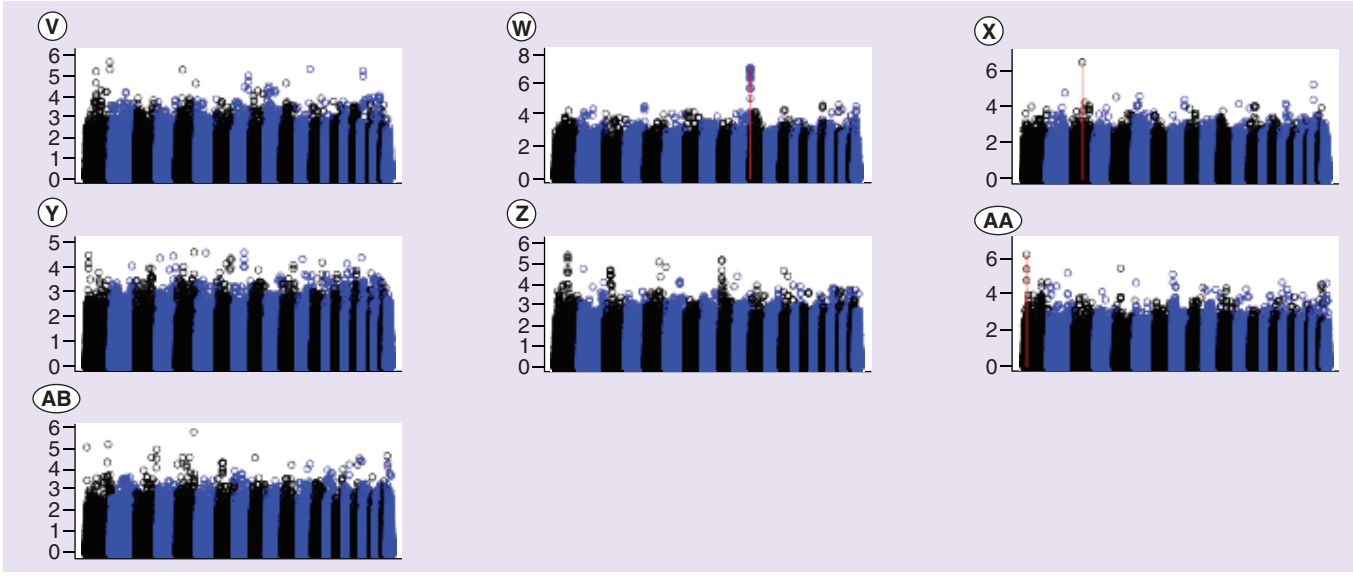


Figure 2. Manhattan plots of 28 chemotherapeutics for hispanic analysis. Manhattan plots are provided for each drug showing the -log10(p) for every SNP. The different chromosomes are represented by alternating colors. Any SNPs with -log10 (p-value) >6 are indicated with red vertical lines. For these analyses, the Human Variation Panel dataset was used. Each plot corresponds to a drug in the following way: **(A)** arsenic trioxide, **(B)** azacitidine, **(C)** carboplatin, **(D)** cladribine, **(E)** cytarabine, **(F)** dasatinib, **(G)** daunorubicin, **(H)** docetaxel, **(I)** doxorubicin, **(J)** epirubicin, **(K)** etoposide, **(L)** 5-fluorouracil, **(M)** floxuridine, **(N)** fludarabine, **(O)** gemcitabine, **(P)** hydroxyurea, **(Q)** idarubicin, **(R)** mitomycin, **(S)** mitoxantrone, **(T)** oxaliplatin, **(U)** paclitaxel, **(V)** sunitinib, **(W)** temozolomide, **(X)** teniposide, **(Y)** topotecan, **(Z)** vinblastine, **(AA)** vincristine, **(AB)** vinorelbine. For plots V–AB, see facing page.



in 2 × 2 quadruplicates. These quadruplicates were used to assess intraplate variability and reproducibility. Handling of the resulting 71,424 wells was automated using a Tecan Freedom EVO150 (Tecan Group Ltd, Seestrasse, Switzerland) with a 96 head MCA. Each plate was incubated for 72 h before the addition of Alamar Blue (Biosource International, Camarillo California) and incubated another 18 h after exposure to Alamar Blue. After the incubation phase, plates were read on an Infinite F200 microplate reader with Connect Stacker (Tecan Group Ltd) and iControl software (Version 1.6) was used to measure fluorescence intensity at EX535nm and EM595nm. The resulting relative fluorescence units are proportional to the concentration of living cells in each well.

Statistical analysis

Associating ancestry/ethnicity with drug response

Univariate and multivariate methods were used to investigate the relationship between reported ethnicity or estimated genetic ancestry and drug response. In univariate approaches, multivariate dose response data are typically fit to a nonlinear model (e.g., a hill function) and a summary metric is used to summarize the behavior of the curve – for example, the potency (AC50: concentration at which 50% of the overall response occurs) or efficacy (Emax: maximum response recorded). This has been a commonly used approach in LCL studies in pharmacogenomics [26–29].

Prior to the univariate analysis, the concentration response data for each individual were fit to a four parameter hill model using nonlinear least squares in the R statistical language. For each individual and drug combination, we used the mean across replicates at each concentration and fit the model across all six concentrations. Then, summary metrics for each individual and drug combination was extracted from the curve fit, the AC50, which provided a measure to test for univariate associations.

An analysis of covariance (ANCOVA) model was used to determine the relationship between self-reported race and drug potency (AC50 values for each individual per drug). Next, a multivariate analysis of covariance (MANCOVA) model was used to determine the relationship between self-reported race and the full matrix of drug concentration-response profiles. The model covariates for both modeling implementations included cellular growth rate, experimental date and time. These covariates were selected after extensive covariate analysis [17]. Hence, the same covariates were used in the current paper and the original CHORI cohort analysis.

Next, admixture refinement was performed on the dataset in an attempt to tease apart the amount of Native American ancestry per individual. Prior to this analysis, HapMap data were added for two global populations, YRI (African) and JPT (Asian) in order to increase the power to differentiate between Native American ancestry and Asian or African ancestry. Using a set of 446 Ancestral Informative Markers (AIMs) [30] (See Supplementary Table 2) which were effective in differentiating the amount of Native American ancestry in the populations, we filtered the four genotypic datasets (CHORI, HVP, CHB, YRI) down to the intersection of directly genotyped data available. The result was 205 AIMs across all populations. Admixture refinement was performed on this subset of genetic markers using the structure-based method in ADMIXTURE [31]. We used cross-validation on $1 \leq K \leq 5$ to determine the best choice for clustering these groups by genetic ancestry. The results indicated that cross-validation error was least pronounced by a selection of $K = 4$, which corroborates the expectation

of ancestral estimates featuring three global and one admixture clusters from these four datasets. An unsupervised analysis was performed with K = 4 groups to distinguish the three global populations from the admixture ancestry. Last, univariate and multivariate models were implemented as described above, replacing self-reported ethnicity with the estimates of Native American ancestry per individual. The analyses were performed on the entire, combined dataset as well as stratified analyses of each each cohort individually.

Genome-wide association mapping

Genome-wide association analyses were performed for each of the 28 drugs for the HVP Mexican samples using the MAGWAS package [32]. Briefly, the approach uses a MANCOVA design to find associations. The rationale for this statistical approach is that modeling the vector of normalized responses jointly captures the concentration response relationship better than the univariate association methods relying on summary metrics like potency or efficacy obtained via fitted hill function parameters [32].

Joint multivariate GWASs were performed on the combined datasets of CHORI and HVP for each of the 28 drugs with MAGWAS. The genotypic data were combined using PLINK. After filtering for minor allele frequency and missingness by SNP(s), there were 519,094 genetic markers for analysis. Imputation was considered for combining across the two datasets, however, confounding from imputation with multiple genotyping platforms (with differential coverage) and from different racial subpopulations limited the validity of this approach. Confounding from imputation across different genotyping platforms was prevented by using only directly genotyped variants. The phenotypic data across all concentrations as well as covariates (cellular growth rate, experimental date and time) were merged. PCA via Eigenstrat was recalculated on the combined set of genotype data and the first three PCs were selected for covariates to account for population stratification between Caucasians and Hispanics as well as capturing batch effects as described in [17].

Results

Associating race/ethnicity with drug response

The results of the univariate and multivariate models showed there are significant variations in drug response across the vast majority of drugs tested (Supplementary Table 3) in relation to self-reported race. For the interested reader, we provide some descriptive statistics on the AC50s (Supplementary Table 4) for each self-reported race. The results of the ADMIXTURE model estimates for the four populations are illustrated in Figure 1. While the data was pooled across all samples/populations for admixture refinement, the results are presented in four distinct stacked bar plots (per self-reported race) in order to easily verify the differences in genetic ancestry per group. Each individual is represented (along the x-axis) by a stacked bar of the percentage genetic ancestry to each of the four clusters. The yellow cluster, Q2, indicates the amount of Native American ancestry, since it is not present in the African population, and minimally represented in the Caucasian and Asian populations. Using the values from the Q2 group, these admixture results were modeled with ANCOVA and MANCOVA across all individual drug responses. The results strongly indicate (Supplementary Table 5) that variation in drug response correlates with Native American ancestry – so admixture at an individual level is strongly associated with differential dose response. Notably, for stratified analyses (HVP only or CHORI only), the multivariate and univariate analysis of covariance models showed only a small subset of drugs correlating variation in drug response with Native American ancestry estimates (Supplementary Table 6).

Table 1. Peak associations by SNP and gene for Human Variation Panel-only analysis.

Drug name	RSID	Chromosome	Gene	-log10(p)
Arsenic trioxide	rs6544994	2	*KCNK12*	6.7
Dasatinib	rs831612	11	*C11orf91*	6.4
Doxorubicin	rs2072167	7	*ETV1*	6.94
Etoposide	rs9657904	3	*CBLB*	6.04
Fludarabine	rs7827050	8	None	6.15
Hydoxyurea	rs6696562	1	*FRRS1*[†]/*AGL*[†]	6.11
Temozolomide	rs503660	10	*MGMT*	7.02
Teniposide	rs11128244	3	*LINC00877*	6.5
Vincristine	rs12749135	1	*WNT4*[†]	6.26

[†]Within 100 kbp upstream of a gene encoding region.

Table 2. Peak associations by SNP and gene for Human Variation Panel and Children's Hospital of Oakland Research Institute combined analysis.

Drug name	RSID	Chromosome	Gene	-log10(p)
Azacitidine	rs795118	4	None	6.13
Carboplatin	rs9819958	3	*LOC646168*‡/*GOLIM4*‡	6.52
Carboplatin	rs522134	11	*VPS26B*	6.38
Daunorubicin	rs4793487	17	*SLC39A11*	6.21
Docetaxel	rs237617	9	*OR1L1*	6.31
Doxorubicin	rs17364596	2	*FSTL4*†	6.18
Doxorubicin	rs6596147	5	*NOL10*†/*ATP6V1C2*†	6.03
5-Fluorouracil	rs8039721	15	*MEGF11*	6.87
Idarubicin	rs7582313	2	None	7.02
Mitomycin	rs10500551	16	*NFAT5*	7.43
Mitomycin	rs1800566	16	*NQO1*	6.55
Mitomycin	rs12596679	16	*WWP2*	7.03
Oxaliplatin	rs10092265	8	*CSMD1*	6.22
Oxaliplatin	rs10821910	10	*C10orf107*	6.08
Paclitaxel	rs2663711	4	*SPATA5*	7.26
Temozolomide	rs4751099	10	*MGMT*	15.95

†Within 100 kbp upstream of a gene encoding region.
‡Within 100 kbp down-steam of a gene encoding region.

Genome-wide association mapping

While the overall impact of ancestral differences may illustrate differences in drug disposition across broad populations, these associations do not elucidate the genes that are associated with differential response. The results of 28 HVP GWAS (one for each drug) are represented in Figure 2. We further investigated SNPs with suggestive associations ($p < 10^{-6}$) or genome-wide significance ($p < 10^{-8}$), and we report on the top associated SNPs for those regions. In all, there were nine peak SNP associations across nine drugs which are summarized in Table 1. Of these associated SNPs, six are located within a gene encoding region, and two additional SNPs are located within 100 kbp of a gene encoding region. The multivariate GWAS results of the stratified analyses, performed on the merged data (589 samples), are summarized in Table 2, where we report the peak associations for each drug. There were 16 peak associations across 11 drugs. Of these SNPs, 11 are located within a gene encoding region and three more SNPs were located within 100 kbp of a gene encoding region. Additionally, the results of the 28 combined data analyses are illustrated as manhattan plots in Figure 3.

In Figure 4, we show the combined analyses across all 84 GWAS – 56 performed in the current study along with the results from the CHORI only analysis [17]. The plot was generated using Synthesis-View [33]. N.B., in the Figure 4, a few SNPs are labeled as suggestive in CHORI and/or combined but missing in the HVP study. Those SNPs were dropped in a *post hoc* filtering; we do not report on association results where the SNP has ≤0.04% rate for any one genotype (aa, Aa, AA), since this study was not powered for rare variant analysis.

Discussion

For this work, we explored the genetic underpinnings of cytotoxic responses for 28 chemotherapeutic drugs in 589 individual LCLs representing two distinct subpopulations: Hispanic (Mexican) and non-Hispanic/Caucasian. We performed multiple analyses to determine the role of self-reported race and genetic ancestry estimates in relation to variations in drug responses. Furthermore, we performed 56 GWASs: 28 analyses involving HVP only samples and 28 analyses spanning HVP/Hispanic and CHORI/non-Hispanic Caucasian samples.

Disease progression and response to medication is believed to be at least partially determined by genetic factors such as race or ancestry. Indeed, we have shown significant variations in drug responses and self-reported race as well as significant variability between drug responses and genetic ancestry estimates. We employed two statistical methodologies to identify this connection for 28 chemotherapeutic agents. Interest-

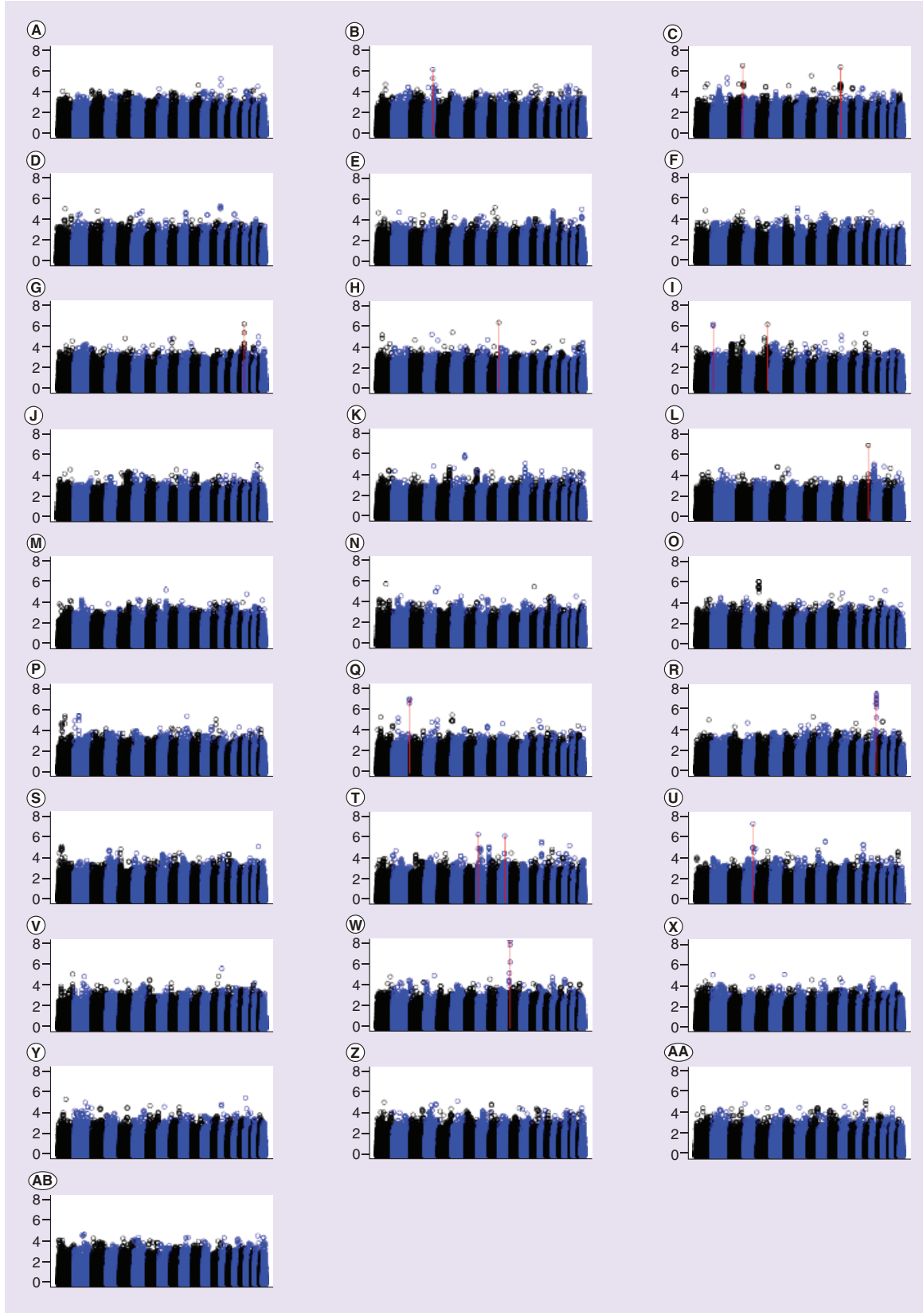

A
B
C
D
E
F
G
H
I
J
K
L
M
N
O
P
Q
R
S
T
U
V
W
X
Y
Z
AA
AB
8
6
4
2
0

Figure 3. Manhattan plots of 28 chemotherapeutics for combined data analysis (see facing page). Manhattan plots are provided for each drug showing the -log10(p) for every SNP. The different chromosomes are represented by alternating colors. Any SNPs with -log10 (p-value) >6 are indicated with red vertical lines. For these analyses, the combined data (CHORI and HVP) were used. Each plot corresponds to a drug in the following way: **(A)** arsenic trioxide, **(B)** azacitidine, **(C)** carboplatin, **(D)** cladribine, **(E)** cytarabine, **(F)** dasatinib, **(G)** daunorubicin, **(H)** docetaxel, **(I)** doxorubicin, **(J)** epirubicin, **(K)** etoposide, **(L)** 5-fluorouracil, **(M)** floxuridine, **(N)** fludarabine, **(O)** gemcitabine, **(P)** hydroxyurea, **(Q)** idarubicin, **(R)** mitomycin, **(S)** mitoxantrone, **(T)** oxaliplatin, **(U)** paclitaxel, **(V)** sunitinib, **(W)** temozolomide, **(X)** teniposide, **(Y)** topotecan, **(Z)** vinblastine, **(AA)** vincristine, **(AB)** vinorelbine. HVP: Human variation panel.

ingly, there were several antineoplastic agents which failed to produce significant results in the ANCOVA model testing self-reported race (cytarabine, dasatinib, fludarabine, oxaliplatin, sunitinub and temozolomide) or genetic admixture (same group sans cytarabine) for variations in drug response. Yet, every drug exhibited significant results rejecting the null hypothesis in the multivariate analyses for self-reported race and genetic admixture. Additionally, when we ran MANCOVA and ANCOVA models of admixture and drug response on HVP or CHORI cohorts alone, we saw only a handful of drugs correlating variation with ancestry. Moreover, while self-reported race and estimated ancestry correlate with variations in drug response on the combined data across most of the drugs, these results are not reproduced exactly on the stratified datasets. This could be due to the fact that the number of samples is significantly reduced when testing the admixed population on its own. Additionally, the dynamic range of ancestry is significantly reduced when testing the Hispanic population and non-Hispanic populations alone.

In the multivariate models, the entire concentration response is utilized (using the same samples and covariates as the univariate models), whereas the univariate model carries the intrinsic assumption that concentration response relationships are sufficiently captured with a single value summarizing the individual curves (in this case, the IC50 value). However, the variability in concentration response might be better identified using a different metric for certain drugs (e.g., efficacy/Emax values). Also, the choice of curve fitting parameterization methods can have a significant impact on univariate models. Given the assumptions that genetic ancestry inherently plays a role in variations in complex phenotypes/diseases and IC50/potency will not always sufficiently elucidate important variability in drug response across different individuals, our results from the multivariate analyses show that, across these 28 drugs, variation in cytoxicity from treatment appears to be related to the genetic ancestry and self-reported race of Caucasian and Hispanic individuals.

Leveraging the data from CHORI [17], we were able to perform 56 independent GWAS representing analyses of the HVP group of individuals across 28 chemotherapeutic agents and combined analyses of the HVP and CHORI cohorts. There were a number of interesting association results. Although the HVP cohort was not very large by GWAS standards, nine of the 28 drugs had at least one suggestive association (-log(p-value) >6). Six of nine of the suggestive associations were for SNPs identified in gene encoding regions (Table 1). For the three remaining SNPs, two were within 100 kbp of a gene encoding region.

For the drug arsenic trioxide, the HVP-only GWAS found a suggestive association for the SNP, rs6544994. This particular SNP was not present on the genotyping platforms for CHORI. As such, there was no way to test for suggestive at this location for the CHORI or combined analyses. The SNP is located in the gene encoding region for *KCNK12*. The protein encoded by *KCNK12* is involved in pore formation for potassium channels. It has been shown to be relevant to neurological disorder predisposition in Hispanic and other populations. It is unclear or even unlikely that *KCNK12* plays a direct role in arsenic trioxide induced cytotoxicity in LCLs. However, at least one SNP from *KCNK12* is highly associated with the *MSH2* gene in LCLs (in particular with HapMap MEX populations) [34]. *MSH2*, an important gene in DNA mismatch repair, modulates apoptosis – which is one of the pathways implicated in the mechanism of action for arsenic trioxides [35].

Three suggestive associations were discovered for the drug doxorubicin: two from the combined analysis and one from the HVP analysis. For the HVP analysis, the SNP rs2072167 was suggestively associated. The SNP is located within the gene encoding region for *ETV1*. This gene is part of the ETS family of transcription regulators, which activate or repress genes in a variety of processes – including cell proliferation, differentiation and apoptosis – and might be important for some tumorigenesis [36]. In the combined analysis, the two SNPs most suggestive association results were for rs17364596 and rs6596147. These SNPs failed to reach suggestive level significance in the CHORI analysis alone, and they were dropped from the HVP analysis due to insufficient genotypes for the homozygous dominant alleles. The SNP rs17364596 is upstream from *ATP6V1C2* and *NOL10* while rs6596147 is upstream from *FSTL4*.

Another intriguing finding occurred in the HVP analysis for the drug etoposide. The peak suggestive

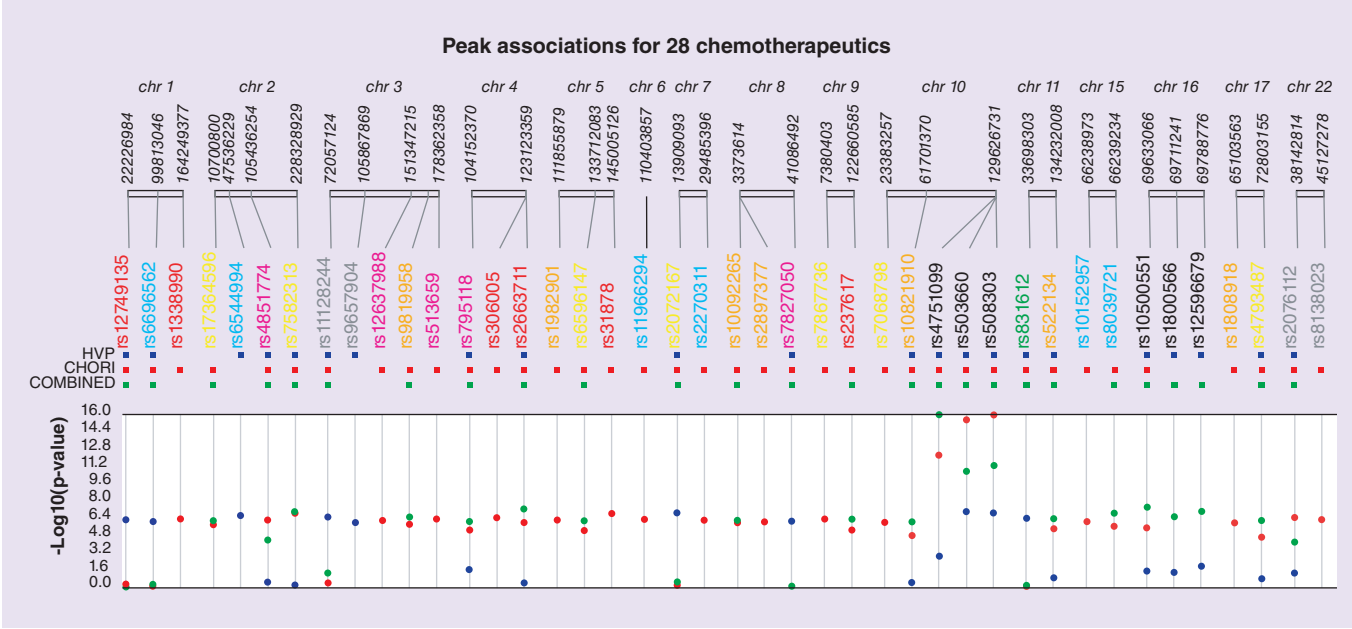


Figure 4. Strongest associations by SNP and cohort. All significant (-log10(p)>6) associations across 84 genome-wide association analyses (i.e., 28 HVP only, 28 CHORI only, and 28 combined analyses). The SNPs are colored by the association drug family. The human genome build hg19 was used for SNP location information.

association for SNP rs9657904 was unavailable for the CHORI or combined analysis, since it was not genotyped with the technology available to CHORI. The SNP is located on chromosome 3 in the region encoding for *CBLB*. This protein affects a wide variety of signaling pathways, as it is generally an important component for signal transduction. Furthermore, the protein encoded by *CBLB* is a negative regulator of the PI3K/Akt pathways, a survival pathway linked to chemotherapy resistance [37].

Notably, for the drug temozolomide, there were highly significant associations ($p < 10^{-15}$) across all three analyses (HVP, CHORI and combined) occurring in the gene encoding region for *MGMT*, a gene whose expression is well known to be inversely associated with clinical response to temozolomide [38]. This particular association result has been followed up, and differential gene expression correlated significantly with the primary SNP in this region [39]. The fact that the current study was able to independently confirm and recapitulate this association result in both the HVP only and combined analyses is an important finding, furthering the credibility of the LCL model system and MAGWAS package for finding clinically relevant genetic variants across different ethnicities.

A surprising result occurred with the drug mitomycin. Both the HVP and CHORI analyses alone did not produce any suggestive associations. However, the combined analysis showed a peak significant association at rs10500551 ($p < 10^{-8}$). The SNP are found in the region encoding the *NFAT5* gene. *NFAT5* dependent gene regulation has a profound role in osmotic stress response [40]. Osmotic pressure has been shown to play a role in the uptake of chemotherapeutic agents *in vitro* [41]. This is a potentially interesting finding which warrants further investigation. Additionally, two more peak associations for SNPs were shown to be suggestive, one found in the region encoding for *NQ01* (relevant in oxidative stress) and the other found in the region encoding for *WWP2* (relevant in the TGF-B pathway). The *NQ01* association is particularly interesting given the clinical significance of this gene for mitomycin treatment outcomes [42].

While these results are promising, they should be viewed as hypothesis generation. The exception is temozolomide; the strong association result on temozolomide from CHORI was followed through with an analysis of gene expression correlations with genetic variation on *MGMT* markers [39]. To understand the implications of the other results reported herein, additional follow up experiments are required. Additionally, it is important to note that while LCLs have proven to be a useful model system, they are not without limitations. For example, LCLs do not express certain proteins implicated in drug metabolism – for example, cytochrome P450s – as such, they cannot predict cytotoxicity profiles from reactive metabolites of many drugs. Furthermore, the nature and conditions for *in vitro* assays (e.g., number of passages) can potentially lead to genotypic and phenotypic changes that modify the model's response to drugs or other stimuli [43].

Conclusion

There is considerable interest in understanding the relationship between ethnicity and drug response outcomes. We generated data and analyses that address this complex relationship. Using self-reported race and estimations of admixture, we have shown strong associations between ethnic differences and drug response. Additionally, we have used the LCL model to performed 56 multivariate GWASs, which offers hypotheses on the genetic differences underlying changes in concentration response seen across Hispanic and non-Hispanic samples. These findings warrant further investigation to ascertain the potential clinical relevance of the genes implicated in our results.

Financial & competing interests disclosure

This work was supported by two grants: T32 GM081057 and R01CA161608 from the National Institute of General Medicine and the National Cancer Institute, respectively. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

Executive summary

- High-throughput data capturing cell viability measurements *in vitro* for 28 chemotherapeutic drugs at six different concentrations per drug using lymphoblastoid cell lines from 589 individuals.
- Genetic ancestry estimated for each sample to gauge the level of admixture across these individuals from two distinct racial groups (Hispanic and Caucasian).
- Analysis of variance and multivariate analysis of variance models suggest significant differences across 28 chemotherapeutic agents by self-reported ethnicity.
- Significant variation in drug response exists across individuals from different subpopulations based on estimates of Native American ancestry.
- In total, 56 multivariate genome-wide association analyses were performed which includes 28 Hispanic only analyses and 28 stratified (Hispanic and non-Hispanic Caucasian).
- Hispanic and non-Hispanic Caucasian samples exhibit unique association results indicating complex relationship between genome and drug response and treatment outcomes.
- Recapitulation of known variants strongly associated with drug response of temozolomide in two distinct populations of very different genetic ancestry.
- Significant associations between polymorphisms and drug response in a lymphoblastoid cell line model from multivariate genome-wide association study were found for nine of 28 chemotherapeutic agents among the Hispanic cohort.
- Significant associations between polymorphisms and drug response in a lymphoblastoid cell line model from multivariate GWAS were found for 11 of 28 chemotherapeutic agents for the combined Hispanic and non-Hispanic dataset.
- Combined association analysis spanning Hispanic and non-Hispanic Caucasian cohorts suggests significant genetic variants of interest for certain chemotherapeutic drugs where independent stratified analyses lacked strong associations.

References

References of special note have been highlighted as:
• Of interest; •• of considerable interest.

1 Crews KR, Hicks JK, Pui C-H, Relling MV, Evans WE. Pharmacogenomics and individualized medicine: translating science into practice. *Clin. Pharmacol. Ther.* 92(4), 467–475 (2012).

2 Low SK, Chung S, Takahashi A *et al.* Genome-wide association study of chemotherapeutic agent-induced severe neutropenia/leucopenia for patients in Biobank Japan. *Cancer Sci.* 104(8), 1074–1082 (2013).

3 Wheeler HE, Maitland ML, Dolan ME, Cox NJ, Ratain MJ. Cancer pharmacogenomics: strategies and challenges. *Nat. Rev. Genet.* 14(1), 23–34 (2012).

4 Jack J, Rotroff D, Motsinger-Reif A. Lymphoblastoid cell lines models of drug response: successes and lessons from this pharmacogenomic model. *Curr. Mol. Med.* 14(7), 833–840 (2014).

•• **Comprehensive review of the lymphoblastoid cell model as a tool/method for generating hypotheses on biological response to drug treatment.**

5 Pollock BH, Debaun MR, Camitta BM *et al.* Racial differences in the survival of childhood b-precursor acute lymphoblastic leukemia: a pediatric oncology group study. *J. Clin. Oncol.* 18(4), 813–813 (2000).

6 Nalls MA, Wilson JG, Patterson NJ *et al.* Admixture mapping of white cell count: genetic locus responsible for lower white blood cell count in the Health ABC and Jackson Heart studies. *Am. J. Hum. Genet.* 82(1), 81–87 (2008).

7 Haiman CA, Patterson N, Freedman ML *et al.* Multiple regions within 8q24 independently affect risk for prostate cancer. *Nat. Genet.* 39(5), 638–644 (2007).

8 Freedman ML, Haiman CA, Patterson N *et al.* Admixture mapping identifies 8q24 as a prostate cancer risk locus in African–American men. *Proc. Nat. Acad. Sci.* 103(38), 14068–14073 (2006).

9 Zhu X, Luke A, Cooper RS *et al.* Admixture mapping for hypertension loci with genome-scan markers. *Nat. Genet.* 37(2), 177–181 (2005).

10 Kao WL, Klag MJ, Meoni LA *et al. MYH9* is associated with nondiabetic end-stage renal disease in African Americans. *Nat. Genet.* 40(10), 1185–1192 (2008).

11 Kadan-Lottick NS, Ness KK, Bhatia S, Gurney JG. Survival variability by race and ethnicity in childhood acute lymphoblastic leukemia. *JAMA* 290(15), 2008–2014 (2003).

12 Yang JJ, Cheng C, Devidas M *et al.* Ancestry and pharmacogenomics of relapse in acute lymphoblastic leukemia. *Nat. Genet.* 43(3), 237–241 (2011).

13 Xu H, Cheng C, Devidas M *et al. ARID5B* genetic polymorphisms contribute to racial disparities in the incidence and treatment outcome of childhood acute lymphoblastic leukemia. *J. Clin. Oncol.* 30(7), 751–757 (2012).

14 Chokkalingam A, Hsu L-I, Metayer C *et al.* Genetic variants in *ARID5B* and *CEBPE* are childhood ALL susceptibility loci in Hispanics. *Cancer Causes Control* 24(10), 1789–1795 (2013).

15 Peters EJ, Motsinger-Reif A, Havener TM *et al.* Pharmacogenomic characterization of US FDA-approved cytotoxic drugs. *Pharmacogenomics* 12(10), 1407–1415 (2011).

• **First report for *in vitro* responses to the anticancer drugs in lymphoblastoid cell lines derived from of CHORI (non-Hispanic Caucasian) patients.**

16 Chen SH, Yang W, Fan Y *et al.* A genome-wide approach identifies that the aspartate metabolism pathway contributes to asparaginase sensitivity. *Leukemia* 25(1), 66–74 (2011).

17 Brown CC, Havener TM, Medina MW *et al.* Genome-wide association and pharmacological profiling of 29 anticancer agents using lymphoblastoid cell lines. *Pharmacogenomics* 15(2), 137–146 (2014).

• **First report for *in vitro* responses to the anticancer drugs in lymphoblastoid cell lines derived from of CHORI (non-Hispanic Caucasian) patients.**

18 Abdo N, Xia M, Brown CC *et al.* Population-based *in vitro* hazard and concentration–response assessment of chemicals: the 1000 genomes high-throughput screening study. *Environ. Health Perspect.* 123(5), 458–466 (2015).

19 Li L, Fridley BL, Kalari K *et al.* Gemcitabine and arabinosylcytosin pharmacogenomics: genome-wide association and drug response biomarkers. *PLoS ONE* 4(11), e7765 (2009).

20 Niu N, Qin Y, Fridley BL *et al.* Radiation pharmacogenomics: a genome-wide association approach to identify radiation response biomarkers using human lymphoblastoid cell lines. *Genome Res.* 20(11), 1482–1492 (2010).

21 Li L, Fridley B, Kalari K *et al.* Gemcitabine and cytosine arabinoside cytotoxicity: association with lymphoblastoid cell expression. *Cancer Res.* 68(17), 7050–7058 (2008).

22 Mitra AK, Crews K, Pounds S *et al.* Impact of genetic variation in FKBP5 on clinical response in pediatric acute myeloid leukemia patients: a pilot study. *Leukemia* 25(8), 1354–1356 (2011).

• **First report for *in vitro* responses to the anticancer drugs in lymphoblastoid cell lines derived from of CHORI (non-Hispanic Caucasian) patients.**

23 International Hapmap Consortium. A haplotype map of the human genome. *Nature* 437(7063), 1299–1320 (2005).

24 Purcell S, Neale B, Todd-Brown K *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81(3), 559–575 (2007).

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.* 38(8), 904–909 (2006).

26 Niu N, Schaid D, Abo R *et al.* Genetic association with overall survival of taxane-treated lung cancer patients – a genome-wide association study in human lymphoblastoid cell lines followed by a clinical association study. *BMC Cancer* 12(1), 1–13 (2012).

27 Tan X-L, Moyer AM, Fridley BL *et al.* Genetic variation predicting cisplatin cytotoxicity associated with overall survival in lung cancer patients receiving platinum-based chemotherapy. *Clin. Cancer Res.* 17(17), 5801–5811 (2011).

28 Niu N, Tan X, Fridley BL *et al.* Abstract 2271: metformin pharmacogenomics: a genome-wide associate study to identify genetic and epigenetic biomarkers involved in metformin response. *Cancer Res.* 73(8 Suppl.), 2271–2271 (2013).

29 Li L, Fridley B, Kalari K *et al.* Discovery of genetic biomarkers contributing to variation in drug response of cytidine analogues using human lymphoblastoid cell lines. *BMC Genomics* 15(1), 93 (2014).

30 Galanter JM, Fernandez-Lopez JC, Gignoux CR *et al.* Development of a panel of genome-wide ancestry informative markers to study admixture throughout the Americas. *PLoS Genet.* 8(3), e1002554 (2012).

31 Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* 19(9), 1655–1664 (2009).

32 Brown C, Motsinger-Reif A. Software for genome-wide association studies having multivariate responses: introducing MAGWAS. *NC State Department Stat. Tech. R* 2641, 1–9 (2012).

•• **Software for the multivariate genome-wide association analysis methods used in this manuscript.**

33 Pendergrass SA, Dudek SM, Crawford DC, Ritchie MD. Synthesis-View: visualization and interpretation of SNP association results for multi-cohort, multi-phenotype data and meta-analysis. *BioData Min.* 3(10) (2010).

34 Abduljaleel Z, Al-Allaf F, Khan W *et al.* DNA mismatch repair *MSH2* gene-based SNP associated with different populations. *Mol. Genet. Genomics* 289(3), 469–487 (2014).

35 Miller WH, Schipper HM, Lee JS, Singer J, Waxman S. Mechanisms of action of arsenic trioxide. *Cancer Res.* 62(14), 3893–3903 (2002).

36 Croce CM. Oncogenes and cancer. *N. Engl. J. Med.* 358(5), 502–511 (2008).

37 Li Y, Qu X, Qu J *et al.* Arsenic trioxide induces apoptosis and G2/M phase arrest by inducing Cbl to inhibit PI3K/Akt signaling and thereby regulate p53 activation. *Cancer Lett.* 284(2), 208–215

38 Hegi ME, Diserens A-C, Gorlia T *et al. MGMT* gene silencing and benefit from temozolomide in glioblastoma. *N. Engl. J. Med.* 352(10), 997–1003 (2005).

39 Brown CC, Havener TM, Medina MW *et al.* A genome-wide association analysis of temozolomide response using lymphoblastoid cell lines reveals a clinically relevant association with MGMT. *Pharmacogenet. Genomics* 22(11), 796–802 (2012).

40 Neuhofer W. Role of NFAT5 in inflammatory disorders associated with osmotic stress. *Curr. Genomics* 11(8), 584–590 (2010).

41 Stephen RL, Novak JM, Jensen EM, Kablitz C, Buys SS. Effect of osmotic pressure on uptake of chemotherapeutic agents by carcinoma cells. *Cancer Res.* 50(15), 4704–4708 (1990).

42 Fleming RA, Drees J, Loggie BW *et al.* Clinical significance of a NAD (P) H: quinone oxidoreductase 1 polymorphism in patients with disseminated peritoneal cancer receiving intraperitoneal hyperthermic chemotherapy with mitomycin C. *Pharmacogenet. Genomics* 12(1), 31–37 (2002).

43 Reid Y, Mintzer J. The current state of cell contamination and authentication – and what it means for biobanks. *Biopreserv. Biobank.* 10(3), 236–238 (2012).