Pharmacogenomic characterization of gemcitabine response – a framework for data integration to enable personalized medicine

Michael Harris\textsuperscript{a,b}, Krithika Bhuvaneshwar\textsuperscript{a}, Thanemozhi Natarajan\textsuperscript{a}, Laura Sheahan\textsuperscript{a,e}, Difei Wang\textsuperscript{a}, Mahlet G. Tadesse\textsuperscript{d}, Ira Shoulson\textsuperscript{c}, Ross Filice\textsuperscript{f}, Kenneth Steadman\textsuperscript{b}, Michael J. Pishvaian\textsuperscript{b}, Subha Madhavan\textsuperscript{a,b} and John Deeken\textsuperscript{b}

Objectives Response to the oncology drug gemcitabine may be variable in part due to genetic differences in the enzymes and transporters responsible for its metabolism and disposition. The aim of our in-silico study was to identify gene variants significantly associated with gemcitabine response that may help to personalize treatment in the clinic.

Methods We analyzed two independent data sets: (a) genotype data from NCI-60 cell lines using the Affymetrix DMET 1.0 platform combined with gemcitabine cytotoxicity data in those cell lines, and (b) genome-wide association studies (GWAS) data from 351 pancreatic cancer patients treated on an NCI-sponsored phase III clinical trial. We also performed a subset analysis on the GWAS data set for 135 patients who were given gemcitabine + placebo. Statistical and systems biology analyses were performed on each individual data set to identify biomarkers significantly associated with gemcitabine response.

Results Genetic variants in the ABC transporters (\textit{ABCC1}, \textit{ABCC4}) and the CYP4 family members \textit{CYP4F8} and \textit{CYP4F12}, \textit{CHST3}, and \textit{PPARD} were found to be significant in both the NCI-60 and GWAS data sets. We report significant association between drug response and variants within members of the chondroitin sulfotransferase family (CHST) whose role in gemcitabine response is yet to be delineated.

Conclusion Biomarkers identified in this integrative analysis may contribute insights into gemcitabine response variability. As genotype data become more readily available, similar studies can be conducted to gain insights into drug response mechanisms and to facilitate clinical trial design and regulatory reviews. Pharmacogenetics and Genomics 24:81–93 © 2014 Wolters Kluwer Health | Lippincott Williams & Wilkins.

Keywords: DMET, gemcitabine, NCI-60, pancreatic cancer, probabilistic networks

*R* Innovation Center for Biomedical Informatics, \textsuperscript{a}Lombardi Comprehensive Cancer Center, Developmental Therapeutics Program, \textsuperscript{b}Department of Neurology, Georgetown University Medical Center, \textsuperscript{c}Department of Mathematics and Statistics, Georgetown University, Washington, District of Columbia, \textsuperscript{d}ESAC Inc., Rockville and \textsuperscript{e}US Food and Drug Administration, Silver Spring, Maryland, USA

Correspondence to Subha Madhavan, PhD, Innovation Center for Biomedical Informatics, Georgetown University Medical Center, 2115 Wisconsin Ave NW, Suite 110, Washington, DC 20007, USA
Tel: +1 202 687 3294; fax: +1 202 687 5011; e-mail: sm696@georgetown.edu

Received 13 March 2013 Accepted 30 September 2013

Introduction

Gemcitabine (2’-deoxy-2’,2’-difluorocytidine, dFdC), an analog of deoxycytidine with proven anti-tumorigenic effects, is used in the treatment of solid tumors including pancreatic cancer. Gemcitabine is active toward many solid tumor types but has a narrow therapeutic index and variable responses ranging from lack of efficacy to severe cytotoxicity, which may be attributed to variability in drug exposure and metabolism [1]. Significant variations in individual response to gemcitabine therapy are common among pancreatic cancer patients. Earlier studies based on cell lines as well as patient–control populations have demonstrated that interindividual variations in germline DNA can impact cellular response to oncology drugs. Identification of such variants with functional/regulatory impact can serve to predict toxicity and efficacy of chemotherapeutic agents [2]. In this study, we have focused on genes encoding drug-metabolizing enzymes and transporters (DMETs), and their association with response to gemcitabine. The products of DMET genes play a substantial role in drug pharmacokinetics, and may have a role in predicting response and clinical outcomes in cancer patients. Some of these variants can not only

Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal’s website (www.pharmacogeneticsandgenomics.com).

This is an open-access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives 3.0 License, where it is permissible to download and share the work provided it is properly cited. The work cannot be changed in any way or used commercially.

1744–8872 © 2014 Wolters Kluwer Health | Lippincott Williams & Wilkins

DOI: 10.1097/FPC.0000000000000015

Copyright © Lippincott Williams & Wilkins. Unauthorized reproduction of this article is prohibited.
impact drug metabolism and transport, but also affect the expression of cancer-related signaling proteins in downstream pathways.

Several potentially actionable variants have been reported in DMET genes that affect drug toxicity and efficacy among individuals, yet few have been tested in the clinic. Several variants in genes directly involved in gemcitabine metabolism have been reported to impact gemcitabine response (e.g., deoxycytidine kinase, DCK; DNA polymerase epsilon, POLE; cytidine deaminase, CDA; and transporters: SLC28A1, SLC28A2, SLC28A3, SLC29A1, SLC29A2, ABCC1, ABCC2, and ABCC10). A significant association has been demonstrated between gemcitabine sensitivity and variants in one or more of the above genes [3–5] including associations between single-nucleotide polymorphism (SNP) haplotypes and gemcitabine treatment outcome in pancreatic cancer patients [6–8].

**Regulatory viewpoint**

Developing innovative clinical evaluation tools and advancing personalized medicine has been identified by the FDA as a core priority area in advancing regulatory science [9]. Understanding the relationship between genetic markers and response to medical products, in terms of both efficacy and toxicity, is a critical component of this priority. First steps toward this understanding include modeling this relationship with a combination of in-vitro and in-vivo genomic and phenotypic markers as described in this paper. Prospective validation of such models along with consideration for regulatory approval and clinical adoption of resultant tools will be required, but work such as this helps lay a strong foundation for more personalized, effective, and safe medical product use.

**Importance of data integration to determine clinically actionable variants**

Understanding the genetic and molecular mechanisms underlyng complex diseases such as cancer is extremely challenging. Genome-wide association studies (GWAS) have been extensively used in the past decade to discover important genetic variants. However, the identified SNPs explain only a small proportion of the phenotypic variation, and the predictive power of these SNPs remains low for many complex diseases [10]. To fully elucidate genetic underpinnings of disease a systems biology approach is necessary to characterize variants, mRNA, copy number, proteins, and metabolites, as well as their cellular interactions [11]. Gene set and pathway association analyses are playing an increasingly important role in explaining disease mechanisms through the identification of functional genetic interactions [12]. Many gene–disease association analyses are based on SNP genotype profiling or gene expression studies. However, SNPs can influence many downstream processes including the expression levels of multiple genes and/or protein levels, and variations in expression levels can directly or indirectly impact disease progression and even drug response [13]. An integrative approach combining multiple data types can more accurately capture pathway associations [12] for discovery of clinically actionable variants.

**Statistical approaches commonly used to associate variants with disease and/or drug response**

Fisher’s exact test (FET) is commonly used in the association of germline polymorphisms with drug response [14]. The use of probabilistic networks in conjunction with traditional statistical models for mining relationships and associations from genotype–phenotype data is well established [15]. Probabilistic network methods for pharmacogenomics and newer methods such as the Markov Blanket concept may be helpful to better analyze these complex genomics–pharmacogenomics associations [16]. Considering the complexity of both cancer diagnosis and individual drug response to chemotherapeutics, application of these association methods in conjunction with novel informatics and data integration approaches is necessary to identify clinically relevant variants for validation studies and ultimately testing in the clinic for pharmacogenomics applications.

**Methods**

**Data sets**

We analyzed SNPs in DMET genes from gemcitabine-focused studies on cell lines and patients to identify associations with drug response. The analysis workflow is summarized in Fig. 1.

1. **NCI-60 data sets:**
   - **SNP data:** DNA from the NCI-60 cell lines was provided by the Developmental Therapeutics Program (DTP) of the National Cancer Institute (NCI). The NCI-60 are well-characterized tumor cell lines. DNA was analyzed using the Affymetrix Targeted Human Drug Metabolizing Enzymes and Transporters (DMET) 1.0 chip [17], which determines the genotype for 1256 variants in 170 genes involved in drug disposition. An additional 14 variants in DNA repair enzymes of interest outside of the DMET chip were also genotyped due to their potential role in a number of anticancer drug pathways.
   - **Gene expression data:** This published data set consists of mRNA expression of the NCI-60 cell lines. Raw data on the Affymetrix U133A gene chip, obtained from the Gene Expression Omnibus (GEO, accession number GSE5720), were used in our analysis [18].
   - **Drug sensitivity data:** Drug sensitivity information is denoted by its GI50 value. The GI50 concentration is defined as the concentration required to achieve 50% growth inhibition. A data set containing the
log 10 of mean GI50 values for the NCI-60 cell lines over seven experiments at a concentration of 10^{-6} mol/l for the drug of interest (gemcitabine, NSC613327) was queried and downloaded from the NCI DTP website [http://dtp.nci.nih.gov]. These data have been provided as Supplementary File S1 (Supplemental digital content 1, http://links.lww.com/FPC/A660). These drug sensitivity data for cell lines are analogous to patient outcome data in response to the same drug.

(2) **GWAS: all patients data set**: GWAS data from 351 patients with advanced pancreatic cancer treated on a CALGB phase III clinical trial with gemcitabine with or without bevacizumab (CALGB 80303) were downloaded from dbGaP [dbGaP: phs000250.v1.p1]. Patient outcomes were classified based on RECIST criteria, including CR (complete response), PR (partial response), SD (stable disease), PD (progressive disease), and unevaluated or NA (not applicable). Patients with clinical response reported as ‘unevaluated’ or NA were excluded from further analysis. Of the 561,466 SNPs in the data, we extracted 2847 SNPs located in genes contained on the Affymetrix DMET platform (Affymetrix Inc., Santa Clara, California, USA) for analysis. Results from the GWAS study showed no significant association between bevacizumab and overall survival [19]. We therefore used the entire cohort for analysis to maximize the input data set.

(3) **GWAS: gemcitabine + placebo data set**: We analyzed the GWAS data set on the subgroup of 135 patients who were given gemcitabine and placebo to remove the effects of bevacizumab on the results.

**Filtering and preprocessing of data sets**

The NCI-60 gene expression data were normalized using the Robust Multichip Average (RMA) method in the R Bioconductor package (Affymetrix Inc.) [20]. All SNP data sets were uniformly processed. Samples with more than 60% missing genotype calls were removed. SNPs with the same genotype across all samples were filtered out since they did not contribute new information. Missing calls, denoted as NoCall (NC), PossibleRareAllele (PRA), or NotAvailable (NA) for the NCI-60 SNP data and 0/0 for the GWAS all patients data set were ignored. SNPs with only reference genotype and missing calls across all samples were also filtered out. After filtering, there were 59 cell lines and 432 variants remaining in the NCI-60 SNP genotype data.

For the GWAS: all patients data set, patients with responses CR and PR were aggregated as they were assumed to be more ‘sensitive’ to drug therapy; similarly, patients with PD and SD were aggregated as they were assumed to be more ‘resistant’ to drug therapy. The GWAS: all patients data set had 293 patients with 2846 variants following this aggregation and filtering. After preprocessing of the GWAS: gemcitabine + placebo data set, there were 2837 variants and 135 patients for analysis.
We chose two methods to analyze these SNP data sets: FET and Probabilistic Network Analysis (PNA) using the software BayesiaLab 5.0.6 [http://www.bayesia.com/en/products/bayesialab.php](http://www.bayesia.com/en/products/bayesialab.php). FET is a well-known nonparametric test that evaluates the association between categorical variables [21]. FET was implemented in the PLINK Whole Genome Association Analysis Toolset [22], which conducts an allele-based test to identify genotype–phenotype association.

A probabilistic network (PN) is a representation of a joint probability distribution [23]. PNA was conducted using the genotype data and gemcitabine response values for each SNP data set. PNs can learn complex models and therefore offer an attractive option to analyze pharmacogenomic SNP data for discovery and prediction [24,25]. They also allow for the opportunity to build clinical decision support systems in the future. Each SNP was represented as a node in the network, whose states are the genotype values observed for that SNP. Gemcitabine response was represented as a special node called the target node, whose states were the discrete gemcitabine response values (sensitive or resistant).

Probabilistic network analysis requires discrete outcome variables, and hence the GI50 values from the NCI-60 gemcitabine response were dichotomized into two groups that correspond to resistant and sensitive phenotypes. We decided to use two levels to ensure adequate sample size in each group. To understand the distribution of the data, a histogram of the log 10 (GI50) values with the corresponding kernel density estimate were plotted (Fig. 2). The smallest antimode, observed at -6.875, was used as the cutoff value [26] between Resistant and Sensitive groups, as displayed by the vertical red line. After the above filtering steps, we produced a consolidated file containing the variant data and associated drug response values. We then used this as input for the SNP analysis described below.

### Estimation of haploblocks

We estimated haploblocks using PLINK to estimate the functional impact of the SNPs and to annotate SNPs located in intergenic regions, which can be associated with SNPs in gene or coding regions if they fall in the same haplotype block. SNPs that fall outside of coding regions (e.g., intronic or UTR regions) are more likely to have a significant impact on gene function if they fall in the same haplotype block as other high-impact coding SNPs. Haplotype blocks were not used for the association analysis and were solely used to determine SNP–gene associations and for functional annotation.

### SNP comparative analysis of gemcitabine response

We investigated the association between genotype data and the gemcitabine response in each data set. Genes found to be significant were considered for further downstream analysis. We also identified genes found in more than one data set as these are likely to play a critical role in gemcitabine response.

For the FET and haplotype block estimation, a PLINK input file was created for each of the SNP data sets. SNPs that were not biallelic were not considered because PLINK only works on biallelic SNPs. FET results with a P-value less than 0.05 were considered significant.

Of the numerous network learning algorithms in PNA, we applied the Augmented Markov Blanket (AMB) algorithm because it has the ability to subset a limited number of SNPs that best predict the outcome [16]. A Markov Blanket corresponds to a set of nodes in the network that make the target independent of all the other nodes conditional on this subset of nodes. The AMB algorithm attempts to find a Markov Blanket for the target node (in this case gemcitabine response) and then uses an unsupervised search to find direct dependencies between each variable belonging to this Markov Blanket. The AMB algorithm implemented in BayesiaLab uses some of the concepts from the Smart-Greedy + algorithm [27]. Specifically, it combines a score-based Augmented Markov Blanket learning algorithm with principles inspired by the constraint-based approaches. The AMB is a nonstochastic learning algorithm that always returns the same result given the same input data and parameters. The AMB algorithm hence created an optimal network with SNPs that were best associated with gemcitabine response [24,28,29].

The genotype data were input ‘as is’ and no further coding was carried out, nor was any genetic model assumed. The parameter we controlled was the ‘structural coefficient,’ which allowed for controlling the extent of the blanket. We first found the optimal value of this
placebo data set. This optimal structural coefficient was used as the reference to avoid making variables and the gene expression data levels were the considered.

Next, to understand the association between SNPs and drug response by applying Student’s two sample t-tests results were selected [7], resulting in a consolidated list of SNPs that affect expression of a gene that in turn affects gemcitabine response.

This integrative analysis was performed in R (http://www.R-project.org), and can be repeated on any drug of interest for which G150 data are available and can help identify SNPs and probe sets best associated with drug response. For each analysis we applied multiple testing correction using Benjamini and Hochberg false discovery rates (FDR) [31]. Because of the discovery nature of this analysis, and hence to maximize the inclusion criteria, the adjusted p-values were not used as a filtering threshold for integration.

Systems biology analysis

The significant genes found in each data set were validated against the literature, and pathway analysis was performed to determine their potential role in one or more of the following processes: gemcitabine metabolism, pancreatic cancer, or interaction with other proteins implicated in the response to gemcitabine or other drugs used to treat pancreatic cancer. We also used the results from the NCI-60 integrated analysis to see whether such associations could be explained at the molecular level using pathway analysis. Pathway analysis was performed using IPA (Ingenuity Systems Inc., Redwood City, California, USA; http://www.ingenuity.com) and Ariadne’s Pathway Studio (Rockville, Maryland, USA).

Results

SNP comparative analysis of gemcitabine response

The PLINK-based FET on the NCI-60 data set resulted in 25 significant SNPs (P < 0.05). PNA analysis identified three SNPs and two of these overlapped with the FET results. In total there were 26 unique SNPs corresponding to 14 unique genes. The genes ABCCI (rs8187858, FET, P = 0.001 [32], cds-synon), CHST3 (rs4148943, FET, P = 0.002, UTR-3), ABCC4 (rs4148551, FET, P = 0.01, UTR-3), CYP2E1 (rs3813867, FET, P = 0.01 [33,34], nearGene-5), and ALDH3A1 (rs887241, FET, P = 0.01, missense [35]) were among the most significant hits. Of the 14 genes identified, three are from the ABCC family (ABCCI, ABCC4, ABCC6), and two are from the CHST family (CHST3, CHST13). CYP genes identified include: CYP2D6, CYP2E1 and CYP4F2, CYP4F8, and CYP4F12. ALDH3A1, PPARD, SLC01B1, and VKORC1 were also identified as significant. Functional analysis of the significant SNPs identified two as damaging: rs2108622,
a missense mutation in the CYP4F2; and rs1056522, a missense mutation in the CHST13 gene.

The FET on the GWAS: gemcitabine + Placebo data set resulted in 123 significant SNPs (P < 0.05). PNA analysis identified 19 significant SNPs, of which 12 overlapped with the FET results. In total there were 130 unique SNPs corresponding to 50 unique genes. Top hits include: CYP39A1 (rs3799884 FET, P = 2.0E−4, intron), ABCC4 (rs7993878, FET, P = 2.3E−4, intron), SLC6A6 (rs2327896, FET, P = 3.1E−4, UTR-3), SLC29A2 (rs2279861, FET, P = 1.1E−3, intron), and ABCB1 (rs17327624, FET, P = 1.3E−3, intron). Among the 50 genes identified were three members of the ABCB family (ABCC1, ABCC4, ABCC8) as well as two CHST family members (CHST8, CHST11). Twelve CYP genes were identified including two members of the CYP2 family (CYP2B6, CYP2J2) and four members of the CYP4 family (CYP4B1, CYP4F3, CYP4F8, CYP4F12). Functional analysis of the significant SNPS identified rs4646487, a missense mutation in the CYP4B1 gene, as damaging.

The FET on the GWAS: all patients data set resulted in 121 significant SNPs (P < 0.05). PNA analysis identified 18 SNPs of which nine overlapped with the FET results. In total there were 130 unique SNPs corresponding to 54 unique genes. Top hits include: CYP4F10P pseudogene (rs1543284, FET, P = 5.0E−5), CYP4F10P pseudogene (rs1543285, FET, P = 2.2E−4), CYP39A1 (rs1063803) (no direct gene association but in the same LD block as SNPs in the CYP4F3 gene, FET, P = 3.6E−4), rs3794989 (no direct gene association but in the same LD block as SNPs in the CYP4F8 gene, FET, P = 4.9E−4), CHST8 (rs17325358, FET, P = 1.5E−3, intron), SLC6A6 (rs2341985, FET, P = 1.5E−3, UTR-3), CYP39A1 (rs3799884, FET, P = 1.8E−3, intron), CHST11 (rs312172, FET, P = 2.1E−3, intron), and SLC29A2 (rs2279861, FET, P = 2.1E−3, intron). The 54 genes identified contained three ABCB family members (ABCC1, ABCC4, ABCC8), five members of the CHST family (CHST3, CHST6, CHST8, CHST9, CHST11), and 10 CYP genes including three members from the CYP4 family (CYP4B1, CYP4F3, CYP4F8), and 10 SLC genes were identified including two members of the SLC7A subfamily (SLC7A7, SLC7A8) and two members of the SLC22A subfamily (SLC22A3, SLC22A6).

Comparison of the significant genes from each data set resulted in the genes ABCC1, ABCC4, and CYP4F8 common to all three data sets. The genes CYP4F12 and PPARD were found in both the NCI-60 and the GWAS: gemcitabine + placebo data sets. CHST3 was found in both the NCI-60 and the GWAS: all patients data set. Thirty-four genes were found in common between the GWAS: all patients and GWAS: gemcitabine + placebo data sets including three members of the UGT2 family (UGT2A1, UGT2A2, UGT2B4).

Table 1 lists the top genes found in each set. Cells in the table are marked ‘Yes’ if an SNP associated with the gene was found to be significant by either FET or PNA. Detailed results of the SNP comparative analysis using PNA and FET can be found in Supplementary File S3 (Supplemental digital content 3, http://links.lww.com/FPC/A662) and Supplementary File S4 (Supplemental digital content 4, http://links.lww.com/FPC/A663), respectively.

Haploblock analysis
The haplotype block estimation (Supplemental File S5, Supplemental digital content 5, http://links.lww.com/FPC/A664) revealed that SNP rs4148943 in the CHST3 gene, which was a top hit in the NCI-60 analysis, is found in the same haplotype block as rs4148946, which was previously reported as a top hit in an analysis of long-term survival in stage III myeloma patients (Table 2) [21]. In addition, rs3842 in the ABCB1 gene has been associated with genetic susceptibility of lung cancer [36] and response to antidepressant treatment [37].

NCI-60 cell line integrative analysis of SNPs, gene expression, and drug response

SNPs in genes that effect their own expression
From the linear regression analysis of NCI-60 SNP and gene expression data, we focused on SNPs that were
associated with expression of their own genes. The results, as shown in Table 3, contain 10 unique SNPs in PPARD, five in ABCC4, two in CYP2E1, and one in SLC01B1. We annotated all variants using the SNPnexus annotation tool [38,39].

**SNPs that affect gene expression and drug response**

The integrative analysis on SNPs, gene expression, and GI50 response revealed significant associations between several variants, and expression of multiple genes that were also associated with GI50 response. Supplementary File S6 (Supplemental digital content 6, http://links.lww.com/FPC/A665) shows 25 unique SNPs (14 unique genes) associated with 51 unique probes (39 unique genes), with a total of 129 combinations. Several of the genes include CHST3 variants affecting SFPQ, SLO4C1, and SLC24A3; PPARD variant affecting SFPQ; CYP2D6 variant affecting FMO5; and CYP4F12 variant affecting FETUB, HNRNPU, and HNRNPAB1.

**Pathway analysis**

We performed pathway analysis to understand the biological role of genes containing significant variants in a comparative study of the NCI-60 cell lines and GWAS data sets. Our study identified five significant variant genes (ABCC1, ABCC4, CYP4F8, CYP4F12, PPARD) common to both the NCI-60 and the GWAS: gemcitabine + placebo data sets with respect to gemcitabine response. Pathway analysis identified shared molecular mechanisms for some genes at both the downstream target level and regulatory level (Fig. 3a). Downstream, ABCC1 and ABCC4 exert positive regulation on other ABC transporters via shared ligands [40] or inducing agents [41]. At the regulatory level, transcription factors MYCN and NFE2L2 [42–45] and cytokines (IL1B) [46,47] play a major role in enhancing the expression of ABCC1 and ABCC4. Pathway analysis of significant genes from NCI-60 (Fig. 3b) and the GWAS: gemcitabine + placebo (Fig. 3c) studies pointed to genes involved in tumor cell proliferation, apoptotic, and inflammatory pathways.

To understand the biological associations between SNPs that affect the expression of target genes as well as drug response, we looked for molecular interactions between genes in which significant SNPs have been identified and target genes that might in-turn affect drug response. The chondroitin sulfotransferase gene, CHST3, was strongly associated with multiple differentially expressed genes in the integrated data set (Fig. 4a). Pathway analysis showed that CHST3 may indirectly influence the RNA-processing gene SFPQ (PTB-associated splicing factor) via other matrix proteins and cell adhesion molecules such as versican and vitronectin, whereas SFPQ shares common pathways with genes whose expression is associated with CHST3 variants. In the PPARD network (Fig. 4b) PPARD may regulate the PTB-associated splicing factor via the
<table>
<thead>
<tr>
<th>SNP (predictor variables)</th>
<th>Chr</th>
<th>dbSNP ID</th>
<th>Annotation</th>
<th>MAF</th>
<th>Probe ID (dependent variables)</th>
<th>Gene</th>
<th>Reference genotype</th>
<th>Coefficients</th>
<th>Estimate</th>
<th>SE</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABCC4_280619 &gt; (rs1059751)</td>
<td>13</td>
<td>rs1059751</td>
<td>mRNA-UTR</td>
<td>0.4237</td>
<td>203196_at</td>
<td>ABCC4</td>
<td>C/C</td>
<td>0.5630</td>
<td>0.2133</td>
<td>0.0257</td>
<td></td>
</tr>
<tr>
<td>ABCC4_280051 &gt; (rs4148551)</td>
<td>13</td>
<td>rs4148551</td>
<td>mRNA-UTR</td>
<td>0.3051</td>
<td>203196_at</td>
<td>ABCC4</td>
<td>A/A</td>
<td>0.5140</td>
<td>0.2180</td>
<td>0.0419</td>
<td></td>
</tr>
<tr>
<td>ABCC4_238593G &gt; (K1116K)</td>
<td>13</td>
<td>rs1751034</td>
<td>coding-synon</td>
<td>0.1441</td>
<td>203196_at</td>
<td>ABCC4</td>
<td>A/A</td>
<td>0.3614</td>
<td>0.2180</td>
<td>0.0370</td>
<td></td>
</tr>
<tr>
<td>ABCC4_279778 &gt; (rs3742106)</td>
<td>13</td>
<td>rs3742106</td>
<td>mRNA-UTR</td>
<td>0.2872</td>
<td>203196_at</td>
<td>ABCC4</td>
<td>T/T</td>
<td>0.4043</td>
<td>0.2333</td>
<td>0.0354</td>
<td></td>
</tr>
<tr>
<td>ABCC4_280434 &gt; (rs4148553)</td>
<td>13</td>
<td>rs4148553</td>
<td>mRNA-UTR</td>
<td>0.4211</td>
<td>203196_at</td>
<td>ABCC4</td>
<td>A/A</td>
<td>0.5243</td>
<td>0.2267</td>
<td>0.0461</td>
<td></td>
</tr>
<tr>
<td>CYP2E1_-1295G &gt; C &gt; (rs3813867)</td>
<td>10</td>
<td>rs3813867</td>
<td>Locus-region</td>
<td>0.0339</td>
<td>222100_at</td>
<td>CYP2E1</td>
<td>G/G</td>
<td>0.3236</td>
<td>0.1467</td>
<td>0.0315</td>
<td></td>
</tr>
<tr>
<td>CYP2E1_-1055C &gt; T &gt; (rs2031920)</td>
<td>10</td>
<td>rs2031920</td>
<td>Locus-region</td>
<td>0.0339</td>
<td>222100_at</td>
<td>CYP2E1</td>
<td>C/C</td>
<td>0.3236</td>
<td>0.1467</td>
<td>0.0315</td>
<td></td>
</tr>
<tr>
<td>PPARD_-2578 &gt; (rs691817)</td>
<td>6</td>
<td>rs691817</td>
<td>Intron</td>
<td>0.0254</td>
<td>208044_s_at</td>
<td>PPARD</td>
<td>C/C</td>
<td>0.6551</td>
<td>0.2031</td>
<td>0.0021</td>
<td></td>
</tr>
<tr>
<td>PPARD_-66611 &gt; (rs2267664)</td>
<td>6</td>
<td>rs2267664</td>
<td>Intron</td>
<td>0.0254</td>
<td>208044_s_at</td>
<td>PPARD</td>
<td>G/G</td>
<td>0.8091</td>
<td>0.3461</td>
<td>0.0231</td>
<td></td>
</tr>
<tr>
<td>PPARD_-66877 &gt; (rs6919734)</td>
<td>6</td>
<td>rs6919734</td>
<td>Intron</td>
<td>0.0172</td>
<td>208044_s_at</td>
<td>PPARD</td>
<td>G/G</td>
<td>0.5693</td>
<td>0.2574</td>
<td>0.0311</td>
<td></td>
</tr>
<tr>
<td>PPARD_-58418 &gt; (rs7771474)</td>
<td>6</td>
<td>rs7771474</td>
<td>Intron</td>
<td>0.0170</td>
<td>208044_s_at</td>
<td>PPARD</td>
<td>C/C</td>
<td>0.1324</td>
<td>0.0553</td>
<td>0.0008</td>
<td></td>
</tr>
<tr>
<td>PPARD_-21172 &gt; (rs3798343)</td>
<td>6</td>
<td>rs3798343</td>
<td>Intron</td>
<td>0.0078</td>
<td>210636_at</td>
<td>PPARD</td>
<td>G/G</td>
<td>0.4381</td>
<td>0.1298</td>
<td>0.0025</td>
<td></td>
</tr>
<tr>
<td>PPARD_-12923C &gt; T (N163N)</td>
<td>6</td>
<td>rs2076167</td>
<td>Coding-synon</td>
<td>0.2500</td>
<td>210636_at</td>
<td>PPARD</td>
<td>T/T</td>
<td>0.2193</td>
<td>0.0840</td>
<td>0.0396</td>
<td></td>
</tr>
<tr>
<td>PPARD_-7112 &gt; (rs7751481)</td>
<td>6</td>
<td>rs7751481</td>
<td>Intron</td>
<td>0.0267</td>
<td>210636_at</td>
<td>PPARD</td>
<td>C/C</td>
<td>0.0991</td>
<td>0.0397</td>
<td>0.0097</td>
<td></td>
</tr>
<tr>
<td>PPARD_-16017 &gt; (rs6457816)</td>
<td>6</td>
<td>rs6457816</td>
<td>Intron</td>
<td>0.1186</td>
<td>210636_at</td>
<td>PPARD</td>
<td>G/G</td>
<td>0.2217</td>
<td>0.0840</td>
<td>0.0368</td>
<td></td>
</tr>
<tr>
<td>PPARD_-9059 &gt; (rs1883322)</td>
<td>6</td>
<td>rs1883322</td>
<td>Intron</td>
<td>0.0267</td>
<td>210636_at</td>
<td>PPARD</td>
<td>T/T</td>
<td>0.2193</td>
<td>0.0840</td>
<td>0.0396</td>
<td></td>
</tr>
<tr>
<td>PPARD_-6457816</td>
<td>6</td>
<td>rs6457816</td>
<td>Intron</td>
<td>0.1186</td>
<td>210636_at</td>
<td>PPARD</td>
<td>C/C</td>
<td>0.2193</td>
<td>0.0840</td>
<td>0.0396</td>
<td></td>
</tr>
<tr>
<td>SLCO1B1_37041T &gt; C (V174A)</td>
<td>12</td>
<td>rs4149056</td>
<td>Coding-synon</td>
<td>0.1186</td>
<td>210366_at</td>
<td>SLCO1B1</td>
<td>C/C</td>
<td>0.1357</td>
<td>0.0869</td>
<td>0.0104</td>
<td></td>
</tr>
</tbody>
</table>

Table includes SNP annotation, probe details, regression coefficient estimates (Estimate), and standard error (SE). It also shows the P-value from the analysis of variance (ANOVA). MAF, minor allele frequency; SNP, single-nucleotide polymorphism.
transcription factor *SPI*. *SPI* and *SFPQ* are known to dynamically regulate growth factor-regulated gene expression response in mammalian cells [48,49]. The CYP450 network identified potential regulatory mechanisms via the xenosensors (Fig. 4c–e). The CYP2D6 network showed interaction with other proteins that regulate one of the significantly associated expression genes, Flavin Containing Monoxygenase 5 (*FMO5*), which is known to mediate the oxidative metabolism of several xenobiotics [50]. Cyp enzymes are regulated by xenosensors such as *NR1I2* (*PXR*) and *NR1H4* (*FXR*) [51], which also induce Fetuin-B, a tumor suppressor, and may contribute to interindividual variability in drug response [52,53] (Fig. 4d). Heterogeneous nuclear ribonucleoproteins (hnRNPU, hnRNPA, hnRNPC) positively regulate the expression of cytokines such as *TNF-α* [54], which in turn can reduce the expression of xenosensors and their CYP enzyme targets [55] and drug transporters such as *SLCOIB1* [56] (Fig. 4e).

Pathway analysis of top genes from SNP comparative analysis. (a) Network created using genes common to the NCI-60 and GWAS: gemcitabine + placebo data sets; (b) Network using the significant genes in the NCI-60 data set; (c) Network using the significant genes in the GWAS: gemcitabine + placebo data set. The significant genes have been indicated using the blue highlight. GWAS, genome-wide association studies; SNP, single-nucleotide polymorphism.
**Discussion**

Abnormal expression and activity of drug transporters and metabolizing enzymes may arise from inherent or acquired polymorphisms and lead to undesirable drug response [2]. Identification of such variants can aid in the selection of patients who may benefit from gemcitabine treatment.

For the GWAS data sets we compared patients treated with gemcitabine + bevacizumab and gemcitabine + placebo as the addition of bevacizumab was shown to have no clinical benefit over gemcitabine treatment alone in these patients. Bevacizumab is an anti-angiogenesis drug that inhibits VEGF-A and its action is different from gemcitabine’s, which inhibits RRM1, a gene that encodes the regulatory subunit of ribonucleotide reductase critical to the synthesis of deoxyribonucleotides. The GWAS: gemcitabine + placebo data set is a smaller, less noisy data set compared with the larger and more diverse GWAS: all patients data set. These differences could be reflected at the molecular level and potentially explain the small difference in results between these two GWAS data sets in the SNP comparative analyses of gemcitabine response and pathway analysis.

ABC transporters have the ability to transport diverse types of anticancer drugs including gemcitabine. Members of ABC transporter family are also involved in diverse processes that impact cell proliferation and apoptotic pathways. Variants in two of these genes, *ABCC1* [57] and *ABCC4* [58], were significant in both studies, the overexpression of which has been observed to contribute to gemcitabine resistance in pancreatic cancer cells. As the functional significance of the SNPs in regulating...
these genes is unknown, we looked at molecular interactions that might explain their effect on gemcitabine. We identified shared pathways for these genes, and suspect that one or more of these SNPs could have a combined effect on gemcitabine response. CYP4F12, whose role in pancreatic cancer or gemcitabine resistance is yet unknown, is one of the enzymes induced by p53 in response to xenobiotic signals such as chemotherapy [59].

Our analyses also revealed significant variants in CHST genes that have not been reported to be involved in gemcitabine transport or metabolism. These enzymes utilize 3′-phospho-5′-adenyl sulfate (PAPS) as sulfonate donor and catalyze the transfer of sulfate to selective glycoprotein ligands such as chondroitin, and the resulting chondroitin sulfates play a critical role in oncogenic HRAS signaling [60] and diverse biological functions including cell proliferation, adhesion, migration, and differentiation [61,62]. Members of the CHST family also mediate inflammation, immunity, angiogenesis, and extracellular matrix reorganization [63,64]. So far there is one study [65] that reports the involvement of chondroitin sulfates in enhancing antitumor activity of gemcitabine in human bladder cancer cells.

Looking at SNPs within the NCI-60 data that affect their own gene expression we found 10 unique SNPs in PPARD and one in SLCO1B1. SLCO1B1 is known to regulate the uptake of drugs and compounds in the liver and SNP rs4149056 is predicted by SIFT and polyphen to be damaging. The C allele of this SNP has been linked to reduced transport activity and hence impaired drug effect [66,67]. Another SLCO1B1 variant, rs4149015, known to affect its own expression [68], may impact its transporting activity and, consequently, drug response. Variants in PPARG2 have been known to affect mRNA expression of PPARG1 and its target genes [69,70], which play a diverse role in cellular energy metabolism and angiogenesis. Activation of PPARβ/δ, in particular, has been reported to favor angiogenesis, and is suggested to play a critical role in inducing angiogenesis in pancreatic cancer [71]. Such alterations in tumor cell characteristics can consequently modify drug response. Variants in ABCC4 have been observed to affect intracellular concentrations of some drugs even though the SNPs were synonymous or within 3′UTR [72,73].

SNPs can influence the expression levels of multiple genes either through cis-regulatory or trans-regulatory effects; and variations in gene expression levels can directly or indirectly impact drug response [74,75]. To understand these effects, we integrated the gene expression data with DMET SNP and drug response data for the NCI-60 cell lines. Associations involving detoxifying proteins such as CYP enzymes and transporters such as SLCO1B1 revealed pathways closely related to drug response regulation. Associations involving other variant genes such as CHST3 and PPARD suggest a major role for cell adhesion, cell growth, and tumor-invasive processes in the alteration's drug response. Although trans-regulatory effects, if present, will have to be experimentally validated, the pathway analysis provides a potential mechanism to explain how SNPs could have a genome-wide impact on drug response, directly or indirectly, by altering tumor cell characteristics.

Conclusion and future perspective

There is strong agreement among clinicians about the need to implement pharmacogenomic discoveries in clinical practice due to the changing needs in medicine. However, most physicians have limited familiarity with genomics and inadequate knowledge of the pharmacogenomic tests available to help their patients [76,77]. Many associations have been identified between variant genotypes and drug response phenotypes, some of which are now identified in FDA-approved medical product labels [78]. Despite the growing evidence of pharmacogenetic involvement in drug response, physician uptake of pharmacogenetic testing has been poor. We believe this is in part due to lack of strong evidence of clinical utility, evolving clinical guidelines around these markers, and lack of integrated decision support tools. Another major barrier to address is the lack of adequate physician education in the field [76].

To overcome some of these challenges, adequate computational and informatics data analysis and support will be necessary to analyze and interpret results from clinical studies and to help evaluate pharmacogenomics biomarker panels. Multidisciplinary teams of clinicians, geneticists, bioinformaticians, and systems biologists need to partner to apply growing pharmacogenomic knowledge in the clinic. Decision support tools must be tightly coupled to electronic health records and clinical workflows to alert healthcare providers to actionable pharmacogenetic information.

Although we present an integrative systems biology and genetic analysis of ‘omics’ data for understanding variants that are strongly associated with gemcitabine response (e.g. CHST family of proteins), this report is intended as a framework to systematically integrate varied data types and strengthen the evidence behind pharmacogenomics markers. Such frameworks can ultimately provide motivation to develop clinical decision support tools and interfaces to help clinicians use pharmacogenomics data to improve patient care.

Acknowledgements

This work was supported, in part, by the FDA Centers of Excellence in Regulatory Science and Innovation (CERSI) (FDA U01 FD004319), and by the National Cancer Institute (P30CA051008). The content is solely the responsibility of the authors and does not necessarily represent the official views of the FDA, the National...
Cancer Institute, or the National Institutes of Health. Support for research computing resources was provided by the Amazon Corporation.

Conflicts of interest
There are no conflicts of interest.

References
Pharmacogenomics of gemcitabine response Harris et al. 93


74 FDA Table of Pharmacogenomic Biomarkers in Drug Labels. Available at: http://www.fda.gov/drugs/scienceresearch/researchareas/unresolvedissues/


81 FDA Table of Pharmacogenomic Biomarkers in Drug Labels. Available at: http://www.fda.gov/drugs/scienceresearch/researchareas/