Knowledge-Guided Prioritization of Genes Determinant of Drug Response using ProGENI

Amin Emad, Ph.D.

Carl R. Woese Institute for Genomic Biology, University of Illinois at Urbana-Champaign

Junmei Cairns, Ph.D.

Department of Molecular Pharmacology and Experimental Therapeutics, Mayo Clinic.

Krishna R. Kalari, Ph.D.

Department of Health Sciences Research, Mayo Clinic.

Liewei Wang, M.D., Ph.D.

Department of Molecular Pharmacology and Experimental Therapeutics, Mayo Clinic.

Saurabh Sinha, Ph.D.

Department of Computer Science and Institute of Genomic Biology, University of Illinois at Urbana-Champaign

Corresponding Authors:

Saurabh Sinha
2122 Siebel Center
201 N. Goodwin Ave,
Urbana, IL 61801. USA.
Phone: 217-333-3233
Email: sinhas@illinois.edu

Liewei Wang
Gonda 19, Mayo Clinic Rochester
200, 1st St. SW,
Rochester, MN 55905. USA.
Phone: 507-284-5264
Email: Wang.Liewei@mayo.edu

ABSTRACT

Identification of genes whose basal mRNA expression can predict the sensitivity of tumor cells to treatments can play an important role in individualized cancer medicine. Screening the expression of these genes in the tumor tissue may suggest the best course of chemotherapy or suggest a combination of drugs to overcome chemoresistance. In this study, we developed a computational method called Prioritization of Genes Enhanced with Network Information (ProGENI), to identify such genes by leveraging their basal expressions and prior knowledge in the form of protein-protein and genetic interactions. ProGENI is based on identifying a small set of genes where a combination of their expression and the activity level of the network module surrounding them shows a high correlation with drug response, followed by the ranking of the genes based on their relevance to this set using random walk techniques.

Our analysis on two relatively new and large datasets of cell lines and their response to a large compendium of drugs revealed a significant improvement in predicting drug sensitivity using ProGENI compared to methods that do not consider network information. In addition, we used literature evidence and siRNA knockdown experiments to confirm the effect of highly ranked genes on the sensitivity of three chemotherapy drugs: cisplatin, docetaxel and doxorubicin. Our results confirmed the role of 73% of the genes (33 out of 45) identified using ProGENI in the sensitivity of cell lines to these drugs. These results suggest ProGENI to be a powerful computational technique in identifying genes determining the drug response.

Keywords: Cytotoxicity, drug sensitivity, gene interaction network, gene prioritization, network-based algorithm

INTRODUCTION

The goal of gene prioritization is to rank genes with respect to their relationship to a phenotype (e.g., occurrence of a disease, response to a drug, etc.), providing an experimentalist a way to prioritize genetic perturbation tests and leading to discovery of genes affecting the phenotype. In the context of drug design and chemosensitivity, various gene prioritization techniques have been used to identify drug targets, reveal mechanisms of actions (MoAs) of drugs, and identify genes associated with drug response, as well as for drug repositioning (Emig et al. 2013; Guo et al. 2015; Isik et al. 2015; Rees et al. 2016).

It has been previously shown that gene expression is the most informative currently available 'omic' feature with respect to drug sensitivity prediction (Costello et al. 2014). Basal gene expression of cancer cell lines (CCLs) has been used to rank genes by their role in cytotoxic drug resistance, utilizing correlation analysis (Scherf et al. 2000; Mariadason et al. 2003; Bussey et al. 2006; Rees et al. 2016) or feature selection and regression techniques (Barretina et al. 2012; Garnett et al. 2012; Basu et al. 2013; Iorio et al. 2016) to statistically associate drug response with gene expression profiles of cell lines. At the same time, many genes with key roles escape identification based on expression profiling alone, due to the complexity of drug MoA and noisy data (Rees et al. 2016), and due to the fact that current methods overlook known functional and biochemical relationships among genes involved in the drug MoA. Indeed, several studies have shown that utilizing such prior knowledge can improve gene prioritization based on identification of differentially expressed genes in drug-treated CCLs (Chen et al. 2012; Kotlyar et al. 2012; Guo et al. 2015; Isik et al. 2015). We posited therefore that knowledge-guided techniques should also improve analysis of basal gene expression data for identifying genes involved in drug MoA and chemosensitivity.

Although many aspects of drug MoA can be uncovered through analysis of drug-perturbed gene expression in CCLs, analysis of basal gene expression is valuable because it sheds light on the relationship between the cell's resting physiological state and its chemosensitivity. In addition to the direct targets of a compound, genes and proteins involved in the processes that precede and follow the binding of the compound to its targets also play a crucial role in the compound's MoA (Palmer 2016), and variations in their expression levels may underlie individual variations in drug response, even if they are not found to be differentially expressed in response to drug treatment. Thus, our primary goal here was not to identify biochemical targets of a drug or genes whose expression are affected by the drug, but rather to identify genes whose basal expression affects the drug response. Over- or under-expression of specific genes can be experimentally shown to influence drug sensitivity (Chen et al. 2005; Le and Bast 2011), but performing these experiments for all genes is infeasible and computational methods that can suggest candidates for such tests are necessary. Shortlisting such genes can provide complementary insight into the MoAs of a drug, offer a better understanding of drug resistance mechanisms, suggest novel targets to overcome drug resistance, and identify biomarkers of drug resistance.

We describe here a novel knowledge-guided gene prioritization algorithm called Prioritization of Genes Enhanced with Network Information (ProGENI), that discovers the relationship between basal gene expression and drug response while incorporating prior knowledge in the form of an experimentally verified network of protein-protein interactions (PPI) and genetic interactions. We used the ProGENI gene prioritization technique to analyze two large and relatively new datasets, one that includes nearly 300 human lymphoblastoid cell lines (LCLs) and another that spans over 600 CCLs of different tissues-of-origin. We employed a systematic way to evaluate different methods for gene prioritization and demonstrated the advantage of the ProGENI method. In addition, we used siRNA knockdown experiments to confirm the role of the highly

ranked genes in drug sensitivity for three cytotoxic treatments widely used in chemotherapy. The results of our analysis demonstrate ProGENI to be a powerful computational technique for identifying genes that play key roles in determining drug response.

RESULTS

A network-based method of gene prioritization from basal expression and phenotype data

In a recent study, Rees et al. (Rees et al. 2016) identified the genes most associated with drug response variation in a collection of cell lines based on Pearson's correlation coefficient (PCC) between basal gene expression and response, one gene at a time. We call this the 'Pearson correlation' scheme for gene prioritization. As an alternative to this 'single gene' analysis, we used the Elastic Net algorithm (Barretina et al. 2012; Iorio et al. 2016) to perform linear regression on the drug response against the expression levels of all genes, employing regularization to enforce sparsity of features and thus learn the most relevant genes. Henceforth, we call this the 'Elastic Net' scheme. See Supplemental Methods for details.

We then developed a new method called Prioritization of Genes Enhanced with Network Information (ProGENI) that incorporates a network of known biological relationships among genes in the gene prioritization task. The method is illustrated in Fig. 1A. It is given a gene expression matrix with genes as columns and samples as rows, and a network with genes as nodes and inter-gene relationships as edges. It first performs a 'network-based smoothing' (Hofree et al. 2013; Cho et al. 2015) of the expression matrix so that the transformed expression value of a gene also reflects the activity level of the gene's network-neighborhood (see Methods). Next, it identifies a pre-set (say *m*) number of genes with the highest correlation (both positive and negative) between the transformed expression values and the given phenotype

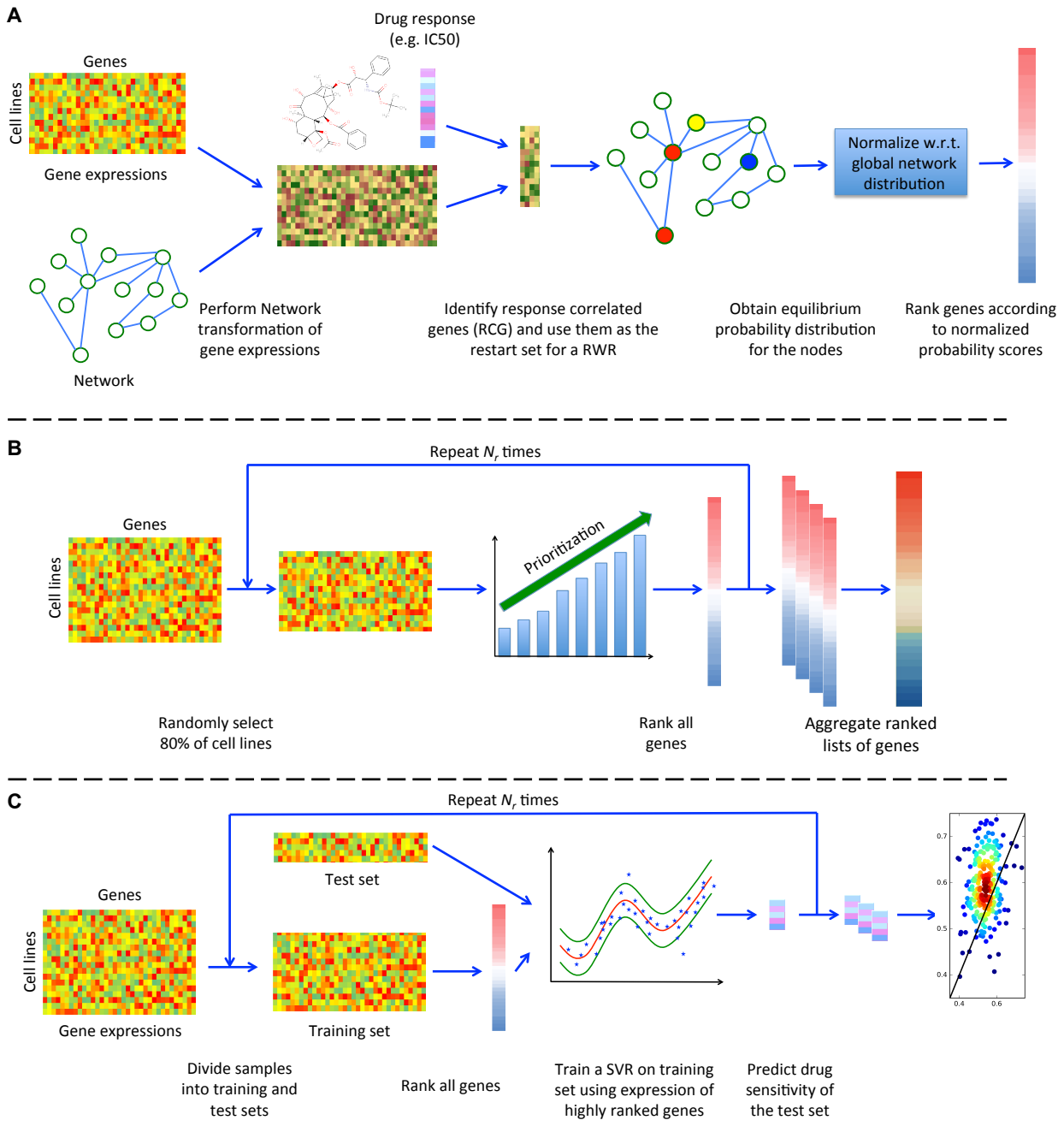

Figure 1: Overview of computational pipelines. (A) ProGENI: An RWR is used to obtain a vector representation of each gene and is used to perform a network transformation on the gene expressions. The response-correlated genes (RCGs) are identified as 100 genes whose transformed expressions have the highest absolute PCC with the drug response. An RWR is used to score each gene based on similarity to the RCG. These scores are then normalized to remove the network bias. (B) Robust ranking: 80% of the cell lines are selected randomly and used with a prioritization method to obtain a ranked list of genes. This procedure is repeated r times and the acquired ranked lists are aggregated to obtain a final ranked list. (C) Cross-validation scheme: a nonlinear support vector machine is trained on the training set using the top 500 genes to predict drug sensitivity of the test set and evaluate the accuracy of prediction.

measurements on samples; these are called the 'response-correlated genes' (RCGs). Then, it performs a random walk with restarts (RWR) on the network, using the genes from the previous step as the restart set, to obtain an equilibrium probability distribution on all the nodes of the network. These probabilities are then normalized with respect to a global equilibrium distribution over all gene nodes that does not depend on the RCG set. Finally, the normalized score for each node is used as the ranking criterion. This approach places the strongest RCGs at or near the top of the list, but the algorithm also makes use of prior knowledge encoded in the network.

To make the reported gene rankings more robust to the effect of noise in the data, we used a bootstrap sampling technique (illustrated in Fig. 1B, also see Methods), whereby prioritization is performed repeatedly on randomly selected subsets of samples and the resulting ranked lists are aggregated to produce the final ranking of genes. Henceforth, we use the name 'Robust-ProGENI' whenever we refer to this bootstrapping scheme and the name 'ProGENI' for the basic method without bootstrapping. We used this robust ranking method with ProGENI as well as the baseline methods in our comparative evaluations described below.

Genes prioritized by ProGENI are more predictive of cytotoxic response than alternatives that do not use network information

We sought to identify the genes associated with individual variation in sensitivity to cytotoxic treatments. Towards this goal, we obtained gene expression and cytotoxic response data (EC50 values for 24 treatments) on approximately 300 LCLs from (Niu et al. 2010; Hanson et al. 2015) (see Methods). We analyzed this LCL dataset with ProGENI, using a network obtained from the STRING database (Szklarczyk et al. 2015) based on protein-protein and genetic interaction data, and focusing on one treatment at a time.

In order to evaluate the gene ranking provided by this method and other prioritization methods (Pearson correlation or Elastic Net), we used a support vector regression (SVR) algorithm to predict cytotoxic response from expression levels of the top 500 ranked genes, and assessed its accuracy with 5-fold cross-validation (see Methods and Fig. 1C). This cross validation scheme was repeated 50 times, resulting in 250 assessments. In each assessment, the performance of the SVR was summarized using the 'scaled probabilistic concordance index' (SPCI) (Costello et al. 2014). This measure ranges between 0 (bad) and 1 (good) and was specifically developed to compare drug sensitivity prediction algorithms in the DREAM 7 challenge (see Supplemental Methods). SPCI values on the 250 test sets were compared between ProGENI and a baseline method, separately for each cytotoxic treatment. According to the above evaluation scheme, gene ranking by ProGENI was significantly better (one-sided paired t-test, $\alpha = 0.05$) than the Pearson correlation scheme for 15 (of 24) cytotoxic treatments (Table 1 and Supplemental Fig. S1), while six treatments showed the opposite trend. When comparing results on all 24 treatments together, ProGENI significantly outperformed Pearson correlation (p-value = 9.93E-36, one-sided Wilcoxon signed-rank test). Fig. 2A shows SPCI measures for these two prioritization schemes over all 250 test sets, for five treatments. Comparisons with the Elastic net scheme showed similar trends (Table 1, Fig. 2B, and Supplemental Fig. S2) with ProGENI significantly outperforming the Elastic Net scheme (p-value = 6.49E-46, one-sided Wilcoxon signed-rank test) overall. (Also see Supplemental Methods for comparison of ProGENI with Elastic Net with a different parameters.)

To gain further confidence in the above observations, we proceeded to repeat the evaluation on a completely different dataset. We obtained drug response data in the form of IC50 values for 139 cytotoxic treatments and gene expression data for more than 600 CCLs from the Genomics of Drug Sensitivity in Cancer (GDSC) database from 13 tissues of origin (Yang et al. 2013). In

Table 1: Performance of drug sensitivity prediction using 500 features selected by ProGENI compared to 500 features selected using baseline schemes, for the LCL dataset. The p-values are calculated using a one-sided paired t-test. The treatments are sorted based on the smallest p-value of the improvement obtained using ProGENI compared to any of the baseline schemes.

Treatment	ProGENI > PCC? (P-value)	ProGENI > Elastic Net? (P-value)	SPCI Average (ProGENI)	SPCI Average (PCC)	SPCI Average (Elastic Net)
MPA	Yes (1.81E-30)	Not significant (0.117)	0.732	0.714	0.728
everolimus	Yes (4.00E-09)	Yes (1.72E-25)	0.595	0.570	0.538
oxaliplatin	No (5.40E-09)	Yes (2.33E-25)	0.676	0.694	0.620
TCN	Yes (5.76E-23)	Yes (5.33E-04)	0.581	0.522	0.555
paclitaxel	Yes (5.44E-08)	Yes (1.29E-17)	0.559	0.542	0.521
docetaxel	Yes (5.20E-13)	Yes (5.23E-14)	0.608	0.585	0.573
6MP	No (6.72E-05)	Yes (2.18E-12)	0.627	0.638	0.588
doxorubicin	Yes (4.52E-12)	Yes (5.74E-04)	0.660	0.641	0.642
cladribine	Yes (4.89E-08)	Yes (7.91E-12)	0.622	0.602	0.592
radiation	Yes (1.82E-09)	Yes (2.19E-07)	0.583	0.560	0.557
epirubicin	Yes (2.55E-09)	Not significant (0.445)	0.646	0.628	0.646
rapamycin	Yes (3.04E-09)	No (1.58E-06)	0.620	0.606	0.641
carboplatin	Yes (5.90E-09)	Not significant (0.265)	0.615	0.596	0.612
NAPQI	Yes (2.44E-08)	Yes (3.45E-08)	0.597	0.575	0.566
arsenic	Yes (1.64E-07)	Yes (3.15E-08)	0.530	0.513	0.509
arac	Yes (3.59E-03)	Yes (1.49E-07)	0.658	0.650	0.629
TMZ	Yes (1.23E-03)	Not significant (0.106)	0.482	0.471	0.489
6TG	No (2.16E-21)	Yes (1.90E-02)	0.717	0.746	0.707
gemcitabine	No (3.12E-04)	Not significant (0.070)	0.739	0.745	0.733
MTX	No (1.34E-03)	Not significant (0.264)	0.583	0.588	0.580
metformin	Not significant (0.347)	Not significant (0.299)	0.583	0.584	0.580
fludarabine	Not significant (0.386)	No (2.76E-03)	0.539	0.540	0.556
hypoxia	Not significant (0.318)	No (3.58E-03)	0.588	0.590	0.601
cddp	No (2.84E-03)	No (4.07E-04)	0.602	0.607	0.619

evaluations similar to those above, ProGENI outperformed the Pearson correlation scheme (one-sided paired t-test, $\alpha = 0.05$) for 68 of 139 treatments (Fig. 2C, Supplemental Table S1), with 45 treatments showing the opposite trend. Similarly, ProGENI outperformed the Elastic Net scheme (one-sided paired t-test with $\alpha=0.05$) for 104 treatments (Fig. 2D, Supplemental Table S1), with the opposite being true in only 9 cases. Considering all drugs simultaneously, ProGENI outperformed both Pearson correlation and Elastic Net schemes (p-value = 4.9E-71 and p-value < 1.0E-307, respectively).

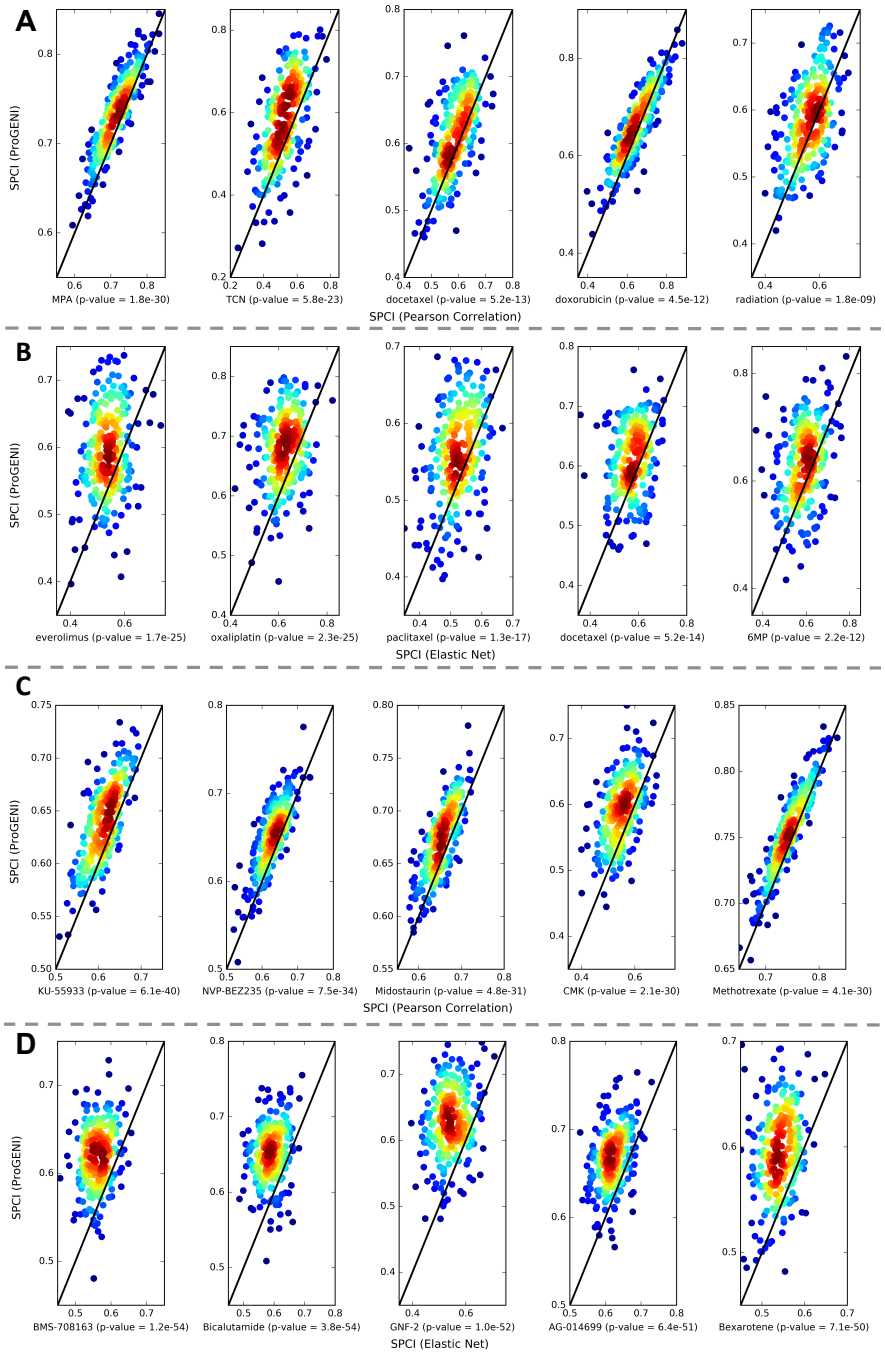

Figure 2: The performance of drug sensitivity prediction based on ProGENI compared to the baseline methods using the LCL (A, B) and GDSC (C, D) datasets for a selection of drugs. The y-axis shows the SPCI corresponding to ProGENI while the x-axis shows the SPCI corresponding to the baseline. Each point in the scatter plot corresponds to one random choice of training/test set. The color of each point represents the density of points in that region: a dark red color on a point means that the point is surrounded by many other points, while a blue color on a point means that the point is isolated. The p-values are calculated using one-sided paired t-test. A) Performance of ProGENI vs. Pearson correlation for the LCL dataset. B) Performance of ProGENI vs. Elastic Net for the LCL dataset. C) Performance of ProGENI vs. Pearson correlation for the GDSC dataset. D) Performance of ProGENI vs. Elastic Net for the GDSC dataset.

Functional validations confirm the role of ProGENI-identified genes in drug response

We sought to verify whether genes associated with drug response variation identified by ProGENI could be linked in vivo to significant changes in drug sensitivity. To this end, we selected the top 15 genes identified using Robust-ProGENI for three drugs – cisplatin, docetaxel, and doxorubicin – from the GDSC dataset. (These drugs belong to three different classes of cytotoxic drugs.) The selections included genes with high Pearson correlation (positive and negative) with drug response (henceforth called 'HPC' genes), as well as genes that were prioritized because their network neighbors' activity was correlated with drug response. As shown in Fig. 3D, four genes for cisplatin, five genes for docetaxel, and eight genes for doxorubicin that were ranked among the top 15 by Robust-ProGENI, are not among the top 15 HPC genes. For example, the expression of *CSNK2A1*, a gene known for its role in doxorubicin response, is not highly correlated with the response to doxorubicin; however it is directly connected in the network to two HPC genes (*NOL3* and *ATF1*) and also has 23 neighbors that are directly connected to HPC genes.

For each identified drug-gene pair, we mined the literature for direct evidence of the gene's role in response to that drug. Out of the 45 pairs examined, we found 'direct' literature evidence for 23 drug-gene pairs in that the gene's knockdown was previously shown to affect chemosensitivity (Table 2 and Supplemental Table S2). For predicted drug-gene pairs that were not validated by literature evidence, we performed siRNA knockdown experiments in two different cell lines of clinical significance: the human triple negative breast cancer MDA-MB-231 and BT549 cells. The siRNA knockdowns were performed for 21 candidate genes predicted by ProGENI to be associated with doxorubicin (8) docetaxel (7), or cisplatin (6), with negative siRNA as a control. The results of these assays for the 21 drug-gene pairs are shown in Fig. 3 and Supplemental Figs. S3 and S4, revealing that 10 of the 21 pairs were validated. Therefore,

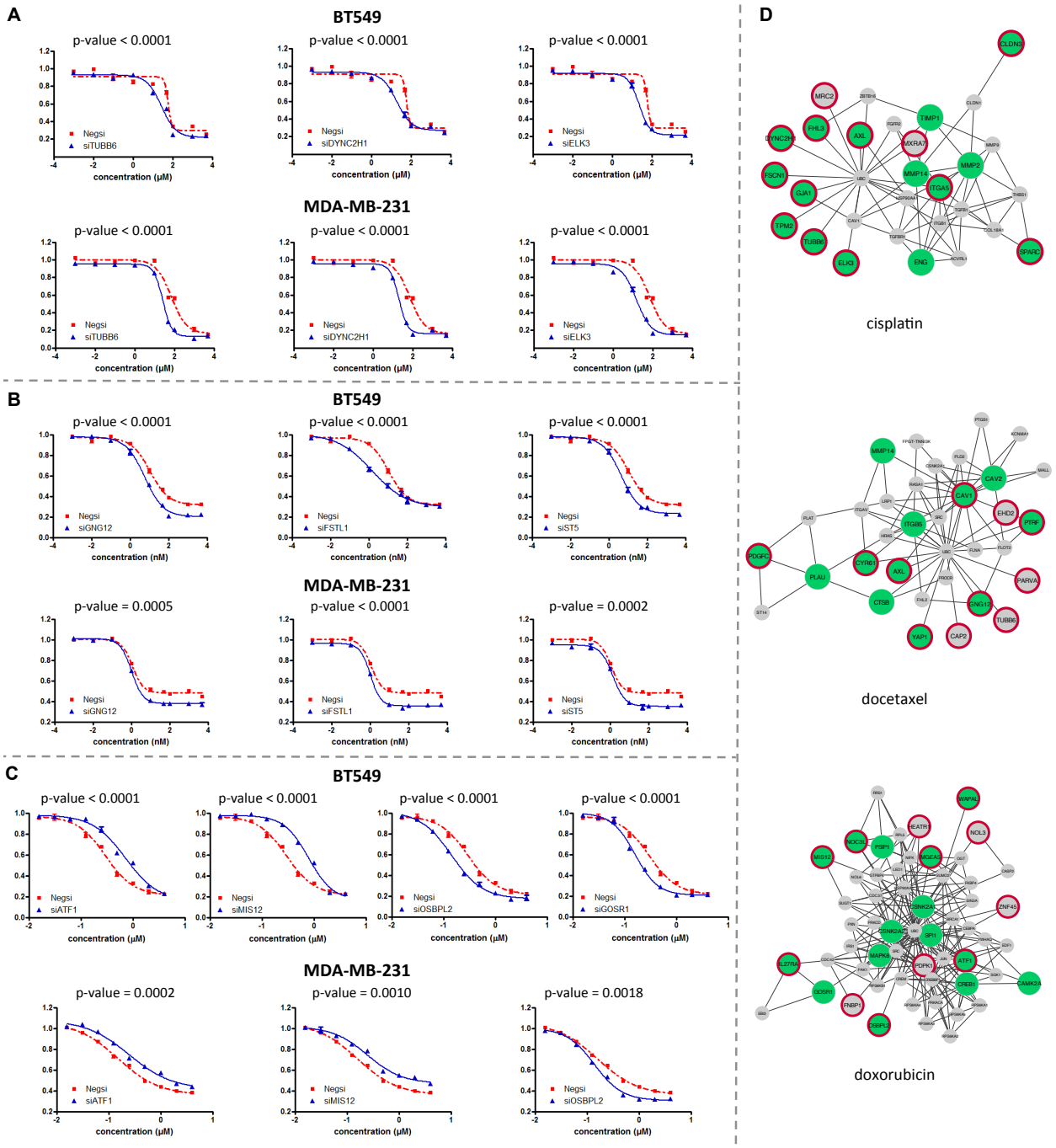

Figure 3: Dosage-response curves for the genes identified using Robust-ProGENI which showed significant change compared to control for (A) cisplatin, (B) docetaxel, and (C) doxorobicin in BT549 and MDA-MB-231 cell lines. P-values are calculated using a two-tailed unpaired t-test. D) The interaction network of genes highly ranked using Robust-ProGENI (green circles), genes highly ranked using Pearson correlation analysis (HPC) (circles with red border), and the shared neighbors of these two groups (small grey circles with no borders). Edges correspond to experimentally obtained PPI and genetic interactions extracted from STRING database; only edges with high affinity scores (>500) are depicted. The degree of a gene that is highly ranked using ProGENI but not among the HPC genes (green circle with no border) shows the number of its HPC neighbors and the number of its shared neighbors with HPC genes. These figures are drawn using Cytoscape (Shannon et al. 2003).

Table 2: Experimental evidence for top 15 genes identified using ProGENI from the GDSC dataset for (A) cisplatin, (B) docetaxel, and (C) doxorubicin. The first column shows the gene symbols, the second column shows the rank of each gene using Robust-ProGENI, the third column shows the rank of each gene using the Pearson correlation scheme, the forth column shows the absolute value of the PCC, and the fifth column shows the nature of the evidence. The green or yellow colors show that siRNA knockdown or similar experiments have shown that the mRNA expression of the gene affects sensitivity to the treatment; such a relationship was found either in literature (green), or was confirmed by siRNA knockdown experiments performed by us (yellow).

A

Gene Symbol	Rank (ProGENI)	Rank (Pearson)	Absolute value of Pearson correlation coefficient	Evidence
TUBB6	2	2	0.2759	Direct (this study)
DYNC2H1	3	4	0.2680	Direct (this study)
CLDN3	4	7	0.2602	Direct (literature)
SPARC	5	8	0.2574	Direct (literature)
GJA1	6	6	0.2623	Direct (literature)
ITGA5	7	11	0.2466	Direct (literature)
TPM2	8	9	0.2567	Direct (literature)
MMP2	9	37	0.2160	Direct (literature)
AXL	12	15	0.2373	Direct (literature)
ENG	13	47	0.2089	Direct (literature)
ELK3	14	13	0.2394	Direct (this study)
TIMP1	15	29	0.2207	Direct (literature)
FSCN1	1	1	0.2879	Not found
FHL3	10	10	0.2477	Not found
MMP14	11	39	0.2143	Not found

B

Gene Symbol	Rank (ProGENI)	Rank (Pearson)	Absolute value of Pearson correlation coefficient	Evidence
CAV1	1	8	0.3713	Direct (literature)
YAP1	2	1	0.4148	Direct (literature)
WWTR1	3	4	0.4075	Direct (literature)
AXL	6	2	0.4098	Direct (literature)
MMP14	7	22	0.3525	Direct (literature)
CYR61	9	6	0.3791	Direct (literature)
CAV2	10	16	0.3566	Direct (literature)
GNG12	11	5	0.3792	Direct (this study)
CTSB	12	27	0.3462	Direct (literature)
FSTL1	14	17	0.3557	Direct (this study)
ST5	15	7	0.3782	Direct (this study)
PDGFC	4	13	0.3659	Not found
PTRF	5	3	0.4094	Not found
ITGB5	8	21	0.3534	Not found
PLAU	13	110	0.3033	Not found

C

Gene Symbol	Rank (ProGENI)	Rank (Pearson)	Absolute value of Pearson correlation coefficient	Evidence
ATF1	1	1	0.2000	Direct (this study)
MIS12	2	4	0.1887	Direct (this study)
OSBPL2	5	6	0.1865	Direct (this study)
CSNK2A1	7	1587	0.0752	Direct (literature)
PSIP1 (LEDGF)	8	46	0.1537	Direct (literature)
CAMK2A	9	6991	0.0157	Direct (literature)
CSNK2A2	10	4870	0.0347	Direct (literature)
GOSR1	11	6867	0.0167	Direct (this study)
MAPK8	13	7574	0.0112	Direct (literature)
CREB1	15	665	0.1000	Direct (literature)
NOC3L	3	3	0.1893	Not found
IL27RA	4	2	0.1911	Not found
MGEA5	6	7	0.1814	Not found
WAPAL	12	8	0.1805	Not found
SPI1	14	6287	0.0217	Not found

overall 33 (73%) of our 45 top predictions for these three drugs have knockdown-based evidence in their favor.

Out of the top 15 genes identified using ProGENI for their role in cisplatin sensitivity, we found direct literature evidence for 9 genes (Table 2A and Supplemental Table S2). For example, *CLDN3* (Claudin-3) (ProGENI rank 4), a gene that is involved in tight junction-specific obliteration of the intercellular space, has been shown to regulate sensitivity to cisplatin by controlling expression of cisplatin influx transporter *CTR1*; in addition, knockdown of *CLDN3* has been shown to increase resistance to cisplatin in human ovarian carcinoma cells in both in vitro culture and in vivo xenograft model (Shang et al. 2013). As another example *MMP2*, a member of the matrix metalloproteinase family involved in the breakdown of the extracellular matrix, was ranked 9 using ProGENI, while Pearson correlation analysis did not place it among the top 15. An inhibitor of *MMP2* has been shown to significantly increase cytotoxicity in cisplatin resistant ovarian carcinoma cell line, A2780cis (Laios et al. 2013). In addition to the 9 (of 15) genes with direct literature evidence, our own experiments revealed that knockdown of three of the remaining 6 predicted genes, *TUBB6*, *DYNC2H1*, and *ELK3*, significantly sensitized both cell lines to cisplatin treatment (Fig. 3A). β-tubulin, of which *TUBB6* is a sub-type, plays a prominent role in cell survival allowing cancer cells to survive, and these cell survival pathways can also be responsible for resistance to chemotherapy (Derry et al. 1997). Suppression of *ELK3* induces sensitivity of MDA-MB-231 cells to doxorubicin treatment by inhibiting autophagy (Park et al. 2016). However, no previous study had linked these three genes to cisplatin sensitivity, making our experimental validation a novel finding.

Among the top 15 genes identified using ProGENI for docetaxel, we found direct literature evidence for 8 genes (Table 2B and Supplemental Table S2). For example, *YAP1* (yes-associated protein 1) (ProGENI rank 2) regulates genes involved in cell proliferation and

apoptosis; induction of this gene has been shown to induce resistance to docetaxel, and its knockdown has been shown to sensitize esophageal cancer cells to this drug (Song et al. 2015). Knockdowns of three of the seven remaining genes, *GNG12*, *FSTL1*, and *ST5*, significantly increased docetaxel sensitivity in both MDA-MB-231 and BT549 cells (Fig. 3B). These three genes are differentially expressed in some cancers. For example, *GNG12* is found to be down-regulated in endometrial cancer (Orchel et al. 2012). *FSTL1* was found to be downregulated in v-myc and v-ras oncogene-transformed cells, with a possible role in carcinogenesis (Johnston et al. 2000), poor prognosis of glioblastoma (Reddy et al. 2008), and progression of prostate cancer (Trojan et al. 2005). *ST5* (*DENND2B*) activates guanosine triphosphatase Rab13 at the leading edge of migrating cells and promotes metastatic behavior (Ioannou et al. 2015). However, none of these three genes were previously known to affect docetaxel sensitivity.

We also found direct literature evidence for 7 genes among the top 15 genes for doxorubicin (Table 2C and Supplemental Table S2). As an example, *CSNK2A1* (Casein Kinase 2 Alpha 1) and its paralog *CSNK2A2* are serine/threonine protein kinases that have regulatory roles in cell proliferation, differentiation and apoptosis. Both of these genes were ranked among the top 15 for doxorubicin using ProGENI, while Pearson correlation analysis places them at ranks 1587 and 4870, respectively. Several studies have shown the role of these genes in resistance to doxorubicin and the synergistic effect between their inhibition and cytotoxicity of doxorubicin (Di Maira et al. 2008; Sandholt et al. 2009; Stolarczyk et al. 2012; Zanin et al. 2012; Tubi et al. 2013). As another example, Daugaard et al. have shown that the ectopic expression of *PSIP1* (*LEDGF*) (ProGENI rank 8) protects MFC-7 cells against several cytotoxic drugs including doxorubicin (Daugaard et al. 2007). Through siRNA knockdown experiments, we found that three genes out of the eight remaining ‘top 15’ predictions for doxorubicin - *ATF1*, *MIS12*, and *OSBPL2* - changed doxorubicin sensitivity in both MDA-MB-231 and BT549 cells (Fig. 6C and

7c). Knockdown of *ATF1* and *MIS12* significantly desensitized both cell lines to doxorubicin treatment, while knockdown of *OSBPL2* significantly sensitized both cell lines to doxorubicin treatment. Additionally, knockdown of *GOSR1* also increased doxorubicin sensitivity in BT549 cells, but had less effect on doxorubicin response in MDA-MB-231 cells. *ATF1*, a negative regulator of apoptosis, is upregulated in metastatic melanoma cells, and inactivation of *ATF1* in melanoma cells resulted in inhibition of tumor growth and metastasis in vivo (Jean et al. 2000). The *MIS12* complex makes an important contribution to kinetochore assembly during cell division (Cheeseman et al. 2006). Defects in kinetochore proteins often lead to aneuploidy and cancer. However, no previous study had linked these genes to doxorubicin sensitivity.

Finally, we note that though several of associations were not corroborated experimentally in selected CCLs, this is expected to an extent as the selection of CCLs was based on clinical indications for each drug and gene transcription regulation is often cell type specific.

Genes highly ranked for many drugs point to common pathways of cytotoxic response

Close examination revealed that some genes are highly ranked for many treatments. Supplemental Table S3 contains a list of 137 genes that were among the top 500 Robust-ProGENI-identified genes for at least 40 (over a quarter of 139 studied) treatments in the GDSC dataset. Functional enrichment analysis using DAVID (Huang da et al. 2009)) (Supplemental Table S3) revealed that these genes are involved in regulation of cell proliferation (43 genes, FDR = 9.19 E-18) and regulation of cell death (28 genes, FDR = 1.67 E-5), which can be explained by the cytotoxic nature of the considered drugs. On the other hand, some of these genes encode proteins that are involved in different processes at the cell surface, such as plasma membrane (76 genes, FDR = 4.05E-08) and cell surface receptor linked signal transduction (54 genes, FDR = 2.33 E-11). Several studies have shown the involvement of

plasma membrane components in multidrug-resistance (MDR) (Milosavljevic et al. 2011; Ferreira et al. 2015; Yu et al. 2015), and transport through the cell membrane, particularly vesicular transport (exosomes), has been linked to resistance to cytotoxic drugs (Yu et al. 2015). Other enrichments include cell adhesion (45 genes, FDR = 1.11 E-21) and focal adhesion (40 genes, FDR = 5.11 E-28) and particularly the integrin family (19 genes, FDR = 3.17 E-18), which has been shown to play an important role in chemo-resistance (Eke and Cordes 2015; Seguin et al. 2015).

Seeking additional global insights about common drug-associated genes, we next formed a drugs x genes matrix indicating the top 500 genes identified for each drug, and used agglomerative clustering to identify four dense biclusters of drugs and genes that are associated with each other (Fig. 4A and Supplemental Table S4). We performed pathway enrichment analysis on the genes in each bicluster, using DAVID (Figs. 4B-E). We noted that one of the biclusters (Fig. 4C) includes genes enriched in the MAPK signaling pathway (FDR = 6.72 E-12). This bicluster includes drugs such as ABT-263, AICAR, ATRA, bicalutamide, IPA3, lenalidomide, methotrexate, nilotinib, PAC1, Vorinostat, and VX-702 for which either the inhibition of the MAPK signaling pathway affects drug-resistance, or this pathway is involved in their MoA (Farooq et al. 2003; Du et al. 2005; Wishart et al. 2006; Yu and Xing 2006; Konig et al. 2008; Kim et al. 2009; Mendy et al. 2009; Keuling et al. 2010; Terakawa et al. 2010; Menges et al. 2012; Uehara et al. 2012). Another bicluster (Fig. 4D) includes genes enriched in the Wnt signaling pathway, and drugs such as QS11, doxorubicin, etoposide, OSU03012, thapsigargin, and tipifarnib, whose association with the Wnt signaling pathway has been confirmed in previous studies (Zhang et al. 2007; Baryawno et al. 2010; Thiago et al. 2010; Lu et al. 2011; Vangipuram et al. 2012; Termglinchan et al. 2013). We also observed a bicluster (Fig. 4E) with a group of MEK inhibitors (AZD6244, CI-1040, PD-0325901, RDEA119) and genes enriched in the inflammation response

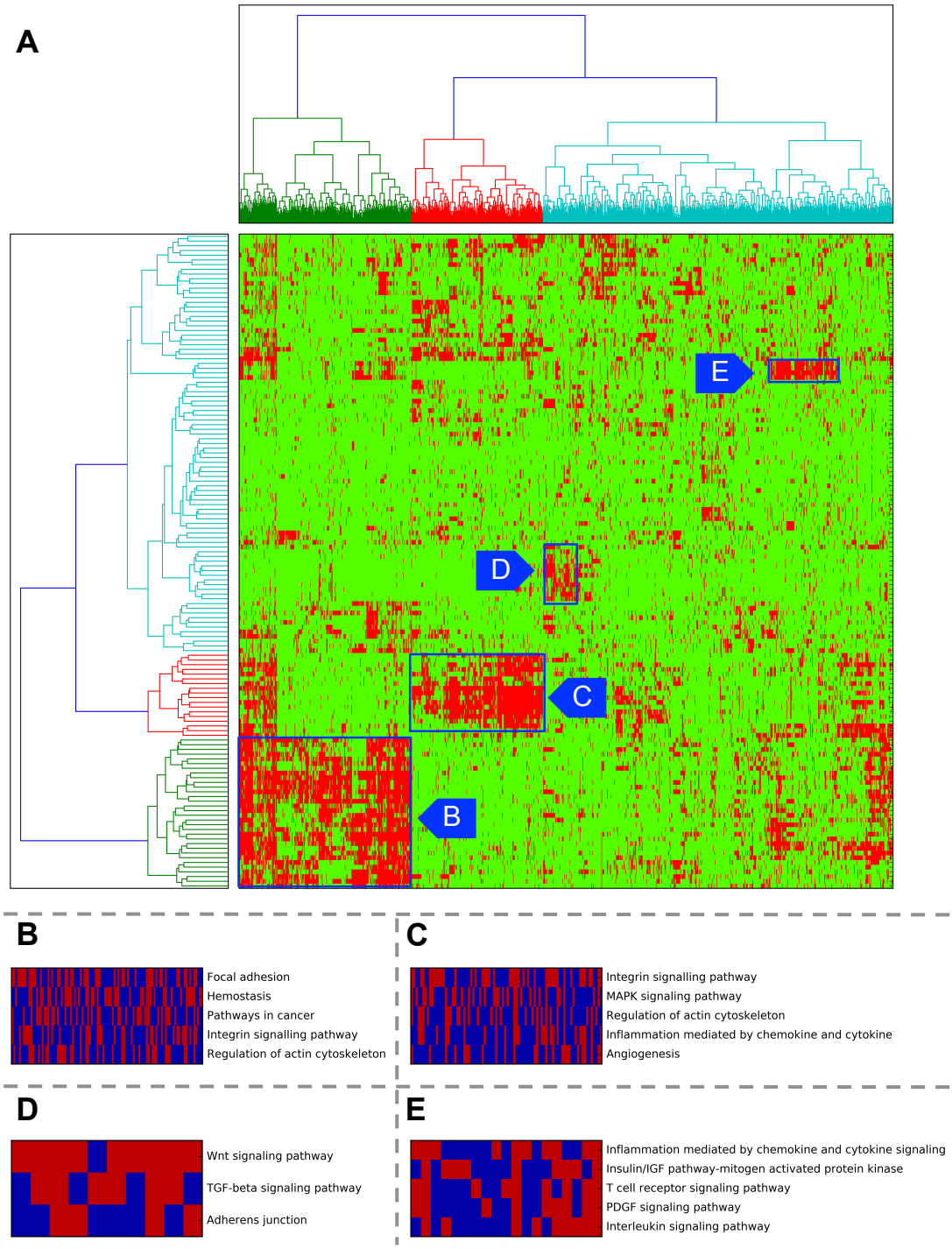

Figure 4: Agglomerative clustering results applied to all drugs and their Robust-ProGENI identified genes from the GDSC dataset. A) Clusters formed for drugs and genes. Rows of the matrix correspond to different drugs and columns correspond to different genes. Only columns (genes) with variance larger than 0.1 were used in the analysis (1177 genes). B, C, D, E) Enriched pathways identified for 4 cluster of genes using DAVID. Pathways with Benjamini-Hochberg corrected p-value <0.1 were sorted based on the number of shared genes and top entries were kept. Columns correspond to genes and rows correspond to pathways, with red indicating that the gene is annotated with that pathway name.

pathways, consistent with prior reports of MEK inhibition resulting in anti-inflammatory response (Jaffee et al. 2000). To summarize, examination of a global map of drug-gene associations predicted by ProGENI reveals sub-groups of similarly acting compounds and pathways involved in their MoA.

Systematic performance analysis of ProGENI

The ProGENI pipeline consists of several steps, each of which plays an important role in prioritizing genes determinant of drug sensitivity. We used the cross validation evaluation (depicted in Fig. 1C) on all 24 treatments of the LCL dataset to study the contribution of these steps (Fig. 5A). First, we noted that the significant gap in performance between ProGENI and the Pearson correlation scheme (p-value = 9.9E-36) becomes much more modest (p-value = 1.5E-3) when the latter scheme is applied to network transformed gene expression values. This demonstrates the significance of network-based transformation of the gene expression matrix by ProGENI. Use of a small set of RCGs (m = 100) as the restart set also contributes to the performance of ProGENI, as was demonstrated by testing a variant algorithm (ProGENI-AGC) where all genes formed the restart set. ProGENI outperforms ProGENI-AGC with p-value = 3.2E-3.

In light of the above evidence in favor of network-guided gene prioritization, we next asked if similarly high performance can be obtained by ignoring RCGs altogether, using only the network information. This may be possible for instance if network hubs are good predictors of drug response in general. We tested this variant method ('NHDS', Fig. 5A), which runs an RWR on the network with all nodes as restart set and thus prioritizes genes with high degree or genes in dense sub-networks, ignoring drug response data altogether. ProGENI significantly outperforms NHDS with p-value = 9.4E-40, showing that a combination of network information and information about gene-phenotype correlations is necessary to achieve the improved

performance of ProGENI. Finally, we tested a variant (ProGENI-NH) that omits the final step of adjusting for the global equilibrium distribution over gene nodes (Fig. 1A, Methods), thereby potentially advancing the ranks of network hubs. Cross validation evaluation showed ProGENI and ProGENI-NH to have very similar performance (p-value = 0.17). However, we noted that omitting this step heavily biases the final ranked list towards network hubs (Table 3A), regardless of the phenotype being studied.

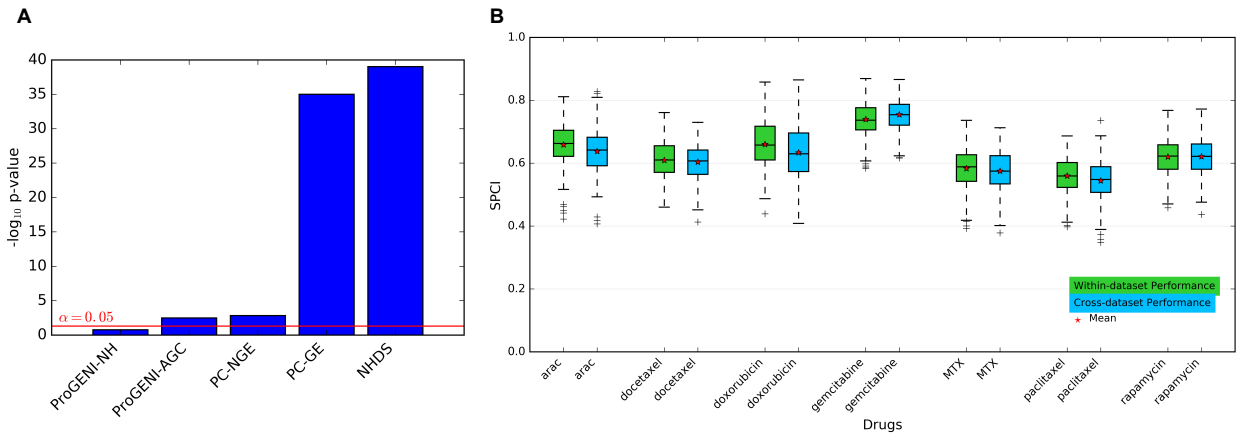

Figure 5: A) The performance of ProGENI compared to other variations of the algorithm using all 24 treatments in the LCL dataset. The y-axis shows -$\log_{10}$ p-value of the improvement provided by ProGENI compared to other algorithms (x-axis), calculated using one-sided Wilcoxon sign-rank test. ProGENI-NH corresponds to ProGENI without normalization of the equilibrium probabilities with respect to the global equilibrium distribution of the nodes. PrOGENI-AGC corresponds to ProGENI with RCG selected as all the nodes in the network. PC-NGE corresponds to prioritization of genes based on Pearson correlation applied to network transformed gene expressions. PC-GE corresponds to prioritization of genes based on Pearson correlation applied to original gene expressions. NHDS corresponds to selection of 500 genes consisting of network hubs and genes in dense sub-networks obtained by running an RWR independent of drug response on the network. B) The Performance of predicting drug sensitivity of the test sets in the LCL dataset using 500 features selected by PrOGENI using the LCL dataset (within-dataset) and the GDSC dataset (cross-dataset). The box plot shows the distribution of the SPCI values for each drug.

ProGENI prioritizes drug-specific genes

We noted above that a network-based prioritization method (ProGENI-NH) may show high accuracy in our cross-validation evaluation despite being heavily biased towards network hubs. If so, it is also possible that the prioritized genes are not specific to the drug being analyzed. To investigate this, we tested whether ProGENI provides a drug-specific ranking of genes. We randomly partitioned the LCL cell lines into two groups of approximately equal sizes and ran

Table 3: A) Presence of network hubs among the highly ranked genes provided by ProGENI and ProGENI-NH. For each treatment, this table shows the size of intersection between the top 500 genes obtained using Robust-ProGENI (or Robust-ProGENI-NH) and the set of 200 genes in the network with the highest degree. P-value for the intersection is calculated using a hypergeometric test. B) Table of drug-specificity of the top 500 genes identified using ProGENI and Pearson correlation sheme using the LCL dataset. A high rank (small entry) shows that the average size of intersection between genes identified using the prioritization method on the two sets of cell lines for the same drug is larger than the intersection when the drug is compared with other drugs. The geometric means of all ranks for prioritization using ProGENI and Pearson correlation are 4.5 and 3.1, respectively.

A

Treatment	Intersection with hubs for ProGENI (p-value)	Intersection with hubs for ProGENI-NH (p-value)
6MP	5 (0.82)	186 (5.75E-269)
6TG	0 (1)	173 (2.10E-235)
arac	0 (1)	187 (9.61E-272)
arsenic	4 (0.92)	195 (1.51E-295)
carboplatin	0 (1)	173 (2.10E-235)
cddp	4 (0.92)	141 (1.20E-165)
cladribine	1 (0.99)	195 (1.51E-295)
docetaxel	3 (0.97)	141 (1.20E-165)
doxorubicin	12 (0.04)	176 (8.56E-243)
epirubicin	4 (0.92)	167 (3.82E-221)
everolimus	8 (0.38)	199 (<1.0E-307)
fludarabine	0 (1)	197 (3.66E-302)
gemcitabine	4 (0.92)	176 (8.56E-243)
hypoxia	38 (3.98E-18)	194 (2.26E-292)
metformin	5 (0.82)	185 (3.19E-266)
MPA	15 (0.0038)	100 (6.28E-94)
MTX	1 (0.99)	182 (3.54E-258)
NAPQI	2 (0.99)	167 (3.82E-221)
oxaliplatin	1 (0.99)	163 (5.56E-212)
paclitaxel	21 (5.30E-06)	175 (2.61E-240)
radiation	1 (0.99)	199 (<1.0E-307)
rapamycin	4 (0.916)	192 (3.16E-286)
TCN	2 (0.99)	194 (2.26E-292)
TMZ	4 (0.92)	189 (2.08E-277)

B

Treatment	Rank (ProGENI)	Rank (Pearson correlation)
6MP	3	1
6TG	23	5
arac	9	4
arsenic	12	12
carboplatin	5	2
cddp	8	3
cladribine	2	3
docetaxel	1	1
doxorubicin	2	2
epirubicin	2	2
everolimus	12	2
fludarabine	11	8
gemcitabine	2	1
hypoxia	2	1
metformin	3	7
MPA	1	1
MTX	13	18
NAPQI	11	1
oxaliplatin	1	1
paclitaxel	1	1
radiation	24	24
rapamycin	2	2
TCN	11	23
TMZ	15	16

ProGENI for all drugs on these two sets. Supplemental Table S5 reports the intersection of the top 500 genes identified for any pair of drugs using ProGENI (or Pearson correlation scheme), averaged over 100 repeats of this procedure. An expected sign of drug-specificity is that gene lists for the same drug (but based on different subsets of cell lines) have a greater intersection

than gene lists for different drugs. Indeed, we noted that for 10 of the 24 treatments the intersection between gene lists based on different subsets of cell lines was ranked 1 or 2 compared to their intersection with gene lists for different drugs (Table 3B). This analysis shows that ProGENI provides a drug-specific set of genes. However, the results are not as specific as provided by the Pearson correlation scheme (Table 3B), which is expected since the latter relies exclusively on response data for each drug.

Cross dataset evaluation of ProGENI

We sought to determine whether drug-associated genes identified using a heterogeneous set of cell lines (GDSC dataset) can help predict drug response in a more homogeneous cohort of cell lines (LCL dataset). We identified seven drugs shared between these two datasets, applied Robust-ProGENI on the GDSC dataset to identify top 500 genes for each drug and evaluated these genes on the LCL dataset using the SVR-based cross-validation scheme (Fig. 1C). We also compared this cross-dataset evaluation to the 'within-dataset' evaluation (Fig. 5B), where top genes are selected using the LCL cell lines in the training set and used to predict the drug sensitivity of the LCL cell lines in the testing set. As expected, the within-dataset evaluations yield stronger performance than the cross-dataset evaluations when aggregating tests over all seven drugs (p-value = 4.6E-15). However, for any given drug the average SPCI values for the two schemes are similar (Fig. 5B), suggesting that it is practical to utilize gene prioritization results from a diverse cohort such as GDSC in a more specific context such as LCLs.

DISCUSSION

Profiling of cell lines is a promising means to better understand the mechanisms that relate the genomic and transcriptomic features to many different phenotypic outcomes (Masters 2000). In

this study, we used both cancer cell lines and LCLs obtained from healthy individuals to identify genes whose over/under expression influences drug response of an individual. To achieve this goal, we proposed ProGENI, a novel method that integrates information on gene interactions and relationships with data on basal mRNA expression and drug cytotoxicity in a panel of cell lines to prioritize genes that determine chemosensitivity. We showed that genes prioritized by ProGENI can together predict drug response more accurately than top genes identified by a single gene method (Pearson correlation (Rees et al. 2016)) as well as a multiple regression method (Elastic Net). Several major steps in ProGENI differentiate it from other methods, including existing methods that utilize network information. We showed that the network transformation it performs on the gene expression matrix enables it to consider the expression of each gene in the context of its network neighbors, and greatly improves performance. The systematic removal of network bias from the output of RWR in the last step of ProGENI allows it to prioritize drug-specific genes; without this step the top ranked genes are significantly enriched in high-degree nodes (network hubs), limiting our ability to obtain a complete picture of drug resistance mechanisms, and divert our attention to generic, phenotype-independent mechanisms. Similar effects of high-degree nodes on network-based analysis have been noted before in other contexts (Gillis and Pavlidis 2012). In addition to cross-validation performance, we also used knockdown evidence from the literature and our own experiments to confirm the role of many of the ProGENI-identified genes in drug resistance. These included genes whose expression had a low correlation with drug response, but the activity of their surrounding neighbors had a high correlation.

Of the 12 genes for which siRNA knockdown did not affect drug sensitivity, eight have expression highly correlated with drug response. Since these genes have high phenotype correlation both individually and in the context of the interaction network, we speculate that the experimental validation failed because these genes are in a family of genes with similar

function, and knockdown of one member is compensated by other genes in that family, or because the role of these genes in drug resistance can only be captured through their corresponding pathways, which do not become disrupted by single-gene knockdown. For example, both ProGENI and correlation analysis placed *PTRF* (Cavin-1) as an influential gene for docetaxel-resistance. Cavin-1 is a member of the cavin family proteins, which along with the caveolin family are responsible for assembly of caveolae (Hansen and Nichols 2010). Although our analysis showed that siRNA knockdown of *PTRF* is not sufficient to affect sensitivity to docetaxel, several studies have shown the role of caveolae and the cavin and caveolin protein families on multi drug resistance and sensitivity to various drugs including docetaxel (Hehlgans and Cordes 2011; Park et al. 2012; Yi et al. 2013; Kang et al. 2016). We speculate therefore that the role of cavin-1 in docetaxel resistance can only be captured in the context of cavin and caveolin families and the pathways with which it is associated.

The modeling techniques we employed in this study can be extended and improved in several directions. First, in our analysis we did not consider the similarities between different treatments, and the identification of genes for each drug was performed independent of other drugs. However, we expect that many of the genes that affect drugs from the same family would be the same. As a result, incorporating drug similarity information based on their chemical structure, their known targets or known MoAs can improve prediction accuracy (Zhang et al. 2015). Another area that can potentially improve these results is incorporating genomic and epigenomic data in the analysis, as has been shown for example by Hanson et al. (Hanson et al. 2015). However, one should note that incorporating such data requires great caution, as the drastic increase in number of features necessitates a much larger number of samples to recover the signal and avoid over-fitting. While, an increase in the number of samples can be obtained by conducting comprehensive experiments and measurements on many cell lines (Iorio et al. 2016), an alternative approach is to combine various datasets obtained in different studies.

However, the success of this approach highly depends on consistency of the combined datasets; unfortunately, several studies have shown a lack of consistency between the drug response of large public datasets (Haibe-Kains et al. 2013). As a result, new standards and protocols may be necessary to ensure reproducibility in large-scale drug screening studies (Hatzis et al. 2014). In addition, as more accurate and comprehensive datasets become available on PPI and genetic interactions among the genes, we expect that new aspects of drug resistance mechanism can be uncovered using network-based methods.

METHODS

Data collection

LCL dataset: We obtained basal gene expression and drug response (half maximal effective concentration or 'EC50') data on 284 LCLs and 24 cytotoxic treatments from (Niu et al. 2010; Hanson et al. 2015). GDSC dataset: We obtained gene expression and drug response (half maximal inhibitory concentration or 'IC50') data on 624 CCLs from 13 different tissue origins and 139 cytotoxic drugs from the Genomics of Drug Sensitivity in Cancer (GDSC) database (release-5.0) (Yang et al. 2013).

The gene interaction network used here was obtained form the STRING database (Szklarczyk et al. 2015), and consists of genetic interactions, protein associations and protein colocalizations obtained experimentally. It includes 2,961,786 undirected weighted edges (relationships) among 15,589 nodes (genes). Additional details are in Supplemental Methods.

Incorporating network information using random walk with restart

Several steps in the proposed algorithm ProGENI use the random walk with restart (RWR) method (Tong et al. 2006) to incorporate network information in the prioritization task. RWR is a

method for quantifying the similarity between any given node of a weighted network and a given set of the nodes, called the restart set. When at a node, the walker can either move to a neighboring node, or it can jump to one of the nodes in the restart set. The probability of each of these decisions is determined by the weights of the adjacent edges and the restart probability p. The equilibrium probability of visiting each node in the network determines the similarity between that node and the restart set.

More formally, let $\boldsymbol{A}$ be an $N_n \times N_n$ symmetric adjacency matrix of the network (with N_n nodes) such that $\boldsymbol{A}(i,j)$ determines the weight of the edge between nodes i and j. Also, let $\boldsymbol{B}$ be the corresponding probability transition matrix obtained by normalizing each column of $\boldsymbol{A}$ to sum up to 1. Let $\boldsymbol{v}$ denote the equilibrium probability of all the nodes. This vector can be obtained iteratively using $\boldsymbol{v}^{(t+1)} = (1-p)\boldsymbol{B}\boldsymbol{v}^{(t)} + p\boldsymbol{w}$, where $\boldsymbol{w}$ is a probability vector of length N_n determining initial probability of restart for each node. An entry in vector $\boldsymbol{w}$ is equal to zero if the corresponding node is not in the restart set, and is nonzero otherwise. See Supplemental Methods for the details of the convergence criterion.

Prioritization of genes enhanced with network information (ProGENI)

ProGENI is a method for gene prioritization that incorporates prior information on gene-gene interactions with basal gene expression and drug response data obtained from a large panel of samples (Fig. 1A). As input, this algorithm accepts a weighted undirected network of gene-gene relationships, a matrix $\boldsymbol{X}$ of gene expression data (samples x genes), and a vector $\boldsymbol{d}$ of drug response values for the samples. First, a log_2 transformation followed by a Z-transform ensures that the expression of each gene across all cell lines follows a distribution with mean of zero and variance of one.

Next, a network transformation is performed on the gene expression matrix $\boldsymbol{X}$ to generate a 'network-smoothed' matrix $\boldsymbol{X}'$ as described next. Let N_n denote the number of nodes in the network and N_s denote the number of genes shared between the gene expression dataset and the network. For each such gene, an N_n dimensional vector representation with respect to other genes in the network is obtained using a random walk with restart (RWR). This representation is equal to the vector of equilibrium probabilities, $\boldsymbol{v}_i$, when the restart set only consists of node i: $\boldsymbol{w}(i) = 1$, and $\boldsymbol{w}(j) = 0$ for $j \neq i$. Using these vector representations, an $N_s \times N_s$ matrix $\boldsymbol{V}$ is formed, where its ith column is obtained from $\boldsymbol{v}_i$ by removing entries corresponding to network nodes not in the expression dataset and normalizing it to sum to 1. Finally, the network-smoothed expression matrix is obtained according to $\boldsymbol{X}' = \boldsymbol{X}\boldsymbol{V}$, followed by a Z-transformation on each column.

Next, we compute for each gene i the absolute Pearson correlation coefficient between their network-smoothed expression (a column of $\boldsymbol{X}'$) and drug response ($\boldsymbol{d}$), across all samples; this is denoted by r_i. Then, 'response-correlated genes' (RCGs) are identified as the set of m genes with the highest values of r_i. The RCG set is used as the restart set in a RWR, in which $\boldsymbol{w}(i) \propto r_i$ if gene i is an RCG. The vector $\boldsymbol{w}$ is scaled so that it sums to 1 and is used in a RWR to generate the equilibrium probability vector $\boldsymbol{v}_{RCG}$. In addition, a global equilibrium probability vector $\boldsymbol{v}_{global}$ is obtained by performing a RWR on the network, with the same probability of restart that was used to obtain $\boldsymbol{v}_{RCG}$, and with all the nodes as the restart set ($\boldsymbol{w}(i) = 1/N_n$ for all i). Finally, $\boldsymbol{v}_{RCG} - \boldsymbol{v}_{global}$ is used as the ranking criterion for gene prioritization. In this study, we used a probability of restart $p = 0.5$ for all RWRs, since this value provides a good balance between the local and global topology of the network.

Robust prioritization using bootstrap sampling and Borda rank aggregation

To obtain rankings robust to noise in the data, we used the bootstrap sampling technique (Fig. 1B). A pre-specified number of samples (80% of the cell lines) are randomly sampled, and used in the prioritization method to obtain a ranked list of genes. This procedure is repeated N_r times (a user specified number) and the geometric mean of the N_r Borda scores obtained for each gene is computed and is used as the final ranking criterion (Sculley 2007). See Supplemental Methods for more details.

Cross validation scheme for prediction of drug response

We used a cross validation scheme (Fig. 1C) to evaluate the ability of different prioritization methods in identifying genes that determine and predict drug sensitivity. We used a 5-fold cross validation procedure, repeated 50 times. In each repeat, the cell lines were randomly grouped into 5 folds; 4 folds (80% of the cell lines) were used as the training set and the remaining cell lines were used as the testing set. Prioritization methods were used to analyze gene expression of cell lines within the training set and identify 500 genes. These genes were then used to train a nonlinear support vector regression (SVR) model with Gaussian kernel, using their expression values (smoothed expression for ProGENI and original expression for baseline methods) as features. (Thus, each cell line was described by a 500 dimensional feature vector.) Hyperparameters of the SVR were learnt using a 4-fold cross validation applied inside the training set. The trained model was then used with the feature vectors corresponding to the cell lines in the test set to predict their drug sensitivity. See Supplemental Methods for the details on the set of parameters used to train the SVR. Comparisons among methods were based on the same cross-validation partitions of cell lines.

Cell culture and treatments for knockdown experiments

Human triple negative breast cancer MDA-MB-231 and BT549 cell lines were obtained from the American Type Culture Collection (Manassas, VA). MDA-MB-231 cells were cultured in L-15 medium containing 10% FBS at 37°C without CO2. BT549 cells were cultured in RPMI 1640 containing 10% FBS at 37°C with 5% CO2.

Doxorubicin, docetaxel, and cisplatin were purchased from Sigma-Aldrich (St. Louis, MO). Drugs were dissolved in DMSO and aliquots of stock solutions were frozen at −80°C. RNA interference (siRNAs) for the candidate genes and negative control siRNA, as well as the Real-time quantitative reverse transcription-PCR (qRT-PCR), were purchased from Dharmacon. Reverse transfection was performed for MDA-MB231and BT549 cells in 96-well plates. Specifically, 3000–4000 cells were mixed with 0.3 μL of lipofectamine RNAi-MAX reagent (Invitrogen) and 10 nM siRNA for each experiment.

Total RNA was isolated from cultured cells transfected with control or specific siRNAs with the Qiagen RNeasy kit (QIAGEN, Inc.), followed by qRT-PCR performed with the one-step, Brilliant SYBR Green qRT-PCR master mix kit (Stratagene). Specifically, primers purchased from QIAGEN were used to perform qRT-PCR using the Stratagene Mx3005P Real-Time PCR detection system (Stratagene). All experiments were performed in triplicate with beta-actin as an internal control. Reverse transcribed Universal Human reference RNA (Stratagene) was used to generate a standard curve. Control reactions lacked RNA template.

MTS cytotoxicity assay

Cell proliferation assays were performed in triplicate at each drug concentration. Cytotoxicity assays with the lymphoblastoid were performed in triplicate at each dose. Specifically, 90 μL of cells (5×10^4 cells) were plated into 96-well plates (Corning, NY) and were treated with

increasing dose of specific drug or Radiation. After incubation for 72 hours, 20 μL of CellTiter 96® AQueous Non-Radioactive Cell Proliferation Assay solution (Promega Corporation, Madison, WI) was added to each well. Plates were read in a Safire2 plate reader (Tecan AG, Switzerland).

Cytotoxicity assays with the tumor cell lines were performed with the CellTiter 96® AQueous Non-Radioactive Cell Proliferation Assay (Promega Corporation, Madison, WI). Specifically, 90 μL of cells (5×10^3 cells) were plated into 96-well plates and were treated with increasing dose of specific drug. The escalation of concentrations is provided in Supplemental Methods. After incubation for 72 hours, 20 μL of CellTiter 96® AQueous Non-Radioactive Cell Proliferation Assay solution (Promega Corporation, Madison, WI) was added to each well. Plates were read in a Safire2 plate reader (Tecan AG, Switzerland). Cytotoxicity was assessed by plotting cell survival versus drug concentration (on a log scale). Significance of the IC50 values between negative control siRNA and gene-specific siRNA was determined by a two-tailed unpaired t-test.

Code availability

An implementation of ProGENI in python, with appropriate documentation, is freely available at: https://github.com/KnowEnG/ProGENI. ProGENI will also be available, along with the underlying network, through the cloud-based analysis framework KnowEnG (knowing.org) upon publication.

ACKNOWLEDGMENTS

This research was supported by grant 1U54GM114838 awarded by NIGMS through funds provided by the trans-NIH Big Data to Knowledge (BD2K) initiative (www.bd2k.nih.gov). This work was also supported by the Mayo Clinic Center for Individualized Medicine (CIM).

CONFLICT OF INTEREST

The authors declare no conflict of interest.

REFERENCES

Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, Wilson CJ, Lehar J, Kryukov GV, Sonkin D et al. 2012. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**: 603-607.

Baryawno N, Sveinbjornsson B, Eksborg S, Chen CS, Kogner P, Johnsen JI. 2010. Small-molecule inhibitors of phosphatidylinositol 3-kinase/Akt signaling inhibit Wnt/beta-catenin pathway cross-talk and suppress medulloblastoma growth. *Cancer Res* **70**: 266-276.

Basu A, Bodycombe NE, Cheah JH, Price EV, Liu K, Schaefer GI, Ebright RY, Stewart ML, Ito D, Wang S et al. 2013. An interactive resource to identify cancer genetic and lineage dependencies targeted by small molecules. *Cell* **154**: 1151-1161.

Bussey KJ, Chin K, Lababidi S, Reimers M, Reinhold WC, Kuo WL, Gwadry F, Ajay, Kouros-Mehr H, Fridlyand J et al. 2006. Integrating data on DNA copy number with gene expression levels and drug sensitivities in the NCI-60 cell line panel. *Mol Cancer Ther* **5**: 853-867.

Cheeseman IM, Chappie JS, Wilson-Kubalek EM, Desai A. 2006. The conserved KMN network constitutes the core microtubule-binding site of the kinetochore. *Cell* **127**: 983-997.

Chen T, Pengetnze Y, Taylor CC. 2005. Src inhibition enhances paclitaxel cytotoxicity in ovarian cancer cells by caspase-9-independent activation of caspase-3. *Mol Cancer Ther* **4**: 217-224.

Chen X, Jiang W, Wang Q, Huang T, Wang P, Li Y, Chen X, Lv Y, Li X. 2012. Systematically characterizing and prioritizing chemosensitivity related gene based on Gene Ontology and protein interaction network. *BMC Med Genomics* **5**: 43.

Cho H, Berger B, Peng J. 2015. Diffusion component analysis: unraveling functional topology in biological networks. In *International Conference on Research in Computational Molecular Biology*, pp. 62-64. Springer.

Costello JC, Heiser LM, Georgii E, Gönen M, Menden MP, Wang NJ, Bansal M, Hintsanen P, Khan SA, Mpindi J-P. 2014. A community effort to assess and improve drug sensitivity prediction algorithms. *Nature biotechnology* **32**: 1202-1212.

Daugaard M, Kirkegaard-Sørensen T, Ostenfeld MS, Aaboe M, Høyer-Hansen M, Ørntoft TF, Rohde M, Jäättelä M. 2007. Lens epithelium-derived growth factor is an Hsp70-2

regulated guardian of lysosomal stability in human cancer. *Cancer research* **67**: 2559-2567.

Derry WB, Wilson L, Khan IA, Luduena RF, Jordan MA. 1997. Taxol differentially modulates the dynamics of microtubules assembled from unfractionated and purified beta-tubulin isotypes. *Biochemistry* **36**: 3554-3562.

Di Maira G, Brustolon F, Tosoni K, Belli S, Krämer SD, Pinna LA, Ruzzene M. 2008. Comparative analysis of CK2 expression and function in tumor cell lines displaying sensitivity vs. resistance to chemical induced apoptosis. *Molecular and cellular biochemistry* **316**: 155-161.

Du JH, Xu N, Song Y, Xu M, Lu ZZ, Han C, Zhang YY. 2005. AICAR stimulates IL-6 production via p38 MAPK in cardiac fibroblasts in adult mice: a possible role for AMPK. *Biochem Biophys Res Commun* **337**: 1139-1144.

Eke I, Cordes N. 2015. Focal adhesion signaling and therapy resistance in cancer. *Semin Cancer Biol* **31**: 65-75.

Emig D, Ivliev A, Pustovalova O, Lancashire L, Bureeva S, Nikolsky Y, Bessarabova M. 2013. Drug target prediction and repositioning using an integrated network-based approach. *PLoS One* **8**: e60618.

Farooq A, Plotnikova O, Chaturvedi G, Yan S, Zeng L, Zhang Q, Zhou MM. 2003. Solution structure of the MAPK phosphatase PAC-1 catalytic domain. Insights into substrate-induced enzymatic activation of MKP. *Structure* **11**: 155-164.

Ferreira RJ, dos Santos DJ, Ferreira MJ. 2015. P-glycoprotein and membrane roles in multidrug resistance. *Future Med Chem* **7**: 929-946.

Garnett MJ, Edelman EJ, Heidorn SJ, Greenman CD, Dastur A, Lau KW, Greninger P, Thompson IR, Luo X, Soares J et al. 2012. Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature* **483**: 570-575.

Gillis J, Pavlidis P. 2012. “Guilt by association” is the exception rather than the rule in gene networks. *PLoS Comput Biol* **8**: e1002444.

Guo H, Dong J, Hu S, Cai X, Tang G, Dou J, Tian M, He F, Nie Y, Fan D. 2015. Biased random walk model for the prioritization of drug resistance associated proteins. *Sci Rep* **5**: 10857.

Haibe-Kains B, El-Hachem N, Birkbak NJ, Jin AC, Beck AH, Aerts HJ, Quackenbush J. 2013. Inconsistency in large pharmacogenomic studies. *Nature* **504**: 389-393.

Hansen CG, Nichols BJ. 2010. Exploring the caves: cavins, caveolins and caveolae. *Trends Cell Biol* **20**: 177-186.

Hanson C, Cairns J, Wang L, Sinha S. 2015. Computational discovery of transcription factors associated with drug response. *Pharmacogenomics J* doi:10.1038/tpj.2015.74.

Hatzis C, Bedard PL, Birkbak NJ, Beck AH, Aerts HJ, Stem DF, Shi L, Clarke R, Quackenbush J, Haibe-Kains B. 2014. Enhancing reproducibility in cancer drug screening: how do we move forward? *Cancer Res* **74**: 4016-4023.

Hehlgans S, Cordes N. 2011. Caveolin-1: an essential modulator of cancer cell radio-and chemoresistance. *Am J Cancer Res* **1**: 521-530.

Hofree M, Shen JP, Carter H, Gross A, Ideker T. 2013. Network-based stratification of tumor mutations. *Nat Methods* **10**: 1108-1115.

Huang da W, Sherman BT, Lempicki RA. 2009. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* **37**: 1-13.

Ioannou MS, Bell ES, Girard M, Chaineau M, Hamlin JN, Daubaras M, Monast A, Park M, Hodgson L, McPherson PS. 2015. DENND2B activates Rab13 at the leading edge of migrating cells and promotes metastatic behavior. *The Journal of cell biology* **208**: 629-648.

Iorio F, Knijnenburg TA, Vis DJ, Bignell GR, Menden MP, Schubert M, Aben N, Goncalves E, Barthorpe S, Lightfoot H et al. 2016. A Landscape of Pharmacogenomic Interactions in Cancer. *Cell* **166**: 740-754.

Isik Z, Baldow C, Cannistraci CV, Schroeder M. 2015. Drug target prioritization by perturbed gene expression and network information. *Sci Rep* **5**: 17417.

Jaffee BD, Manos EJ, Collins RJ, Czerniak PM, Favata MF, Magolda RL, Scherle PA, Trzaskos JM. 2000. Inhibition of MAP kinase kinase (MEK) results in an anti-inflammatory response in vivo. *Biochem Biophys Res Commun* **268**: 647-651.

Jean D, Tellez C, Huang S, Davis DW, Bruns CJ, McConkey DJ, Hinrichs SH, Bar-Eli M. 2000. Inhibition of tumor growth and metastasis of human melanoma by intracellular anti-ATF-1 single chain Fv fragment. *Oncogene* **19**: 2721-2730.

Johnston IM, Spence HJ, Winnie JN, McGarry L, Vass JK, Meagher L, Stapleton G, Ozanne BW. 2000. Regulation of a multigenic invasion programme by the transcription factor, AP-1: re-expression of a down-regulated gene, TSC-36, inhibits invasion. *Oncogene* **19**: 5348-5358.

Kang J, Park JH, Lee HJ, Jo U, Park JK, Seo JH, Kim YH, Kim I, Park KH. 2016. Caveolin-1 Modulates Docetaxel-Induced Cell Death in Breast Cancer Cell Subtypes through Different Mechanisms. *Cancer research and treatment: official journal of Korean Cancer Association* **48**: 715-726.

Keuling AM, Andrew SE, Tron VA. 2010. Inhibition of p38 MAPK enhances ABT-737-induced cell death in melanoma cell lines: novel regulation of PUMA. *Pigment Cell Melanoma Res* **23**: 430-440.

Kim YJ, Song M, Ryu JC. 2009. Inflammation in methotrexate-induced pulmonary toxicity occurs via the p38 MAPK pathway. *Toxicology* **256**: 183-190.

Konig H, Holtz M, Modi H, Manley P, Holyoake T, Forman S, Bhatia R. 2008. Enhanced BCR-ABL kinase inhibition does not result in increased inhibition of downstream signaling pathways or increased growth suppression in CML progenitors. *Leukemia* **22**: 748-755.

Kotlyar M, Fortney K, Jurisica I. 2012. Network-based characterization of drug-regulated genes, drug targets, and toxicity. *Methods* **57**: 499-507.

Laios A, Mohamed BM, Kelly L, Flavin R, Finn S, McEvoy L, Gallagher M, Martin C, Sheils O, Ring M et al. 2013. Pre-Treatment of platinum resistant ovarian cancer cells with an MMP-9/MMP-2 inhibitor prior to cisplatin enhances cytotoxicity as determined by high content screening. *Int J Mol Sci* **14**: 2085-2103.

Le XF, Bast RC, Jr. 2011. Src family kinases and paclitaxel sensitivity. *Cancer Biol Ther* **12**: 260-269.

Lu D, Choi MY, Yu J, Castro JE, Kipps TJ, Carson DA. 2011. Salinomycin inhibits Wnt signaling and selectively induces apoptosis in chronic lymphocytic leukemia cells. *Proc Natl Acad Sci U S A* **108**: 13253-13257.

Mariadason JM, Arango D, Shi Q, Wilson AJ, Corner GA, Nicholas C, Aranes MJ, Lesser M, Schwartz EL, Augenlicht LH. 2003. Gene expression profiling-based prediction of response of colon carcinoma cells to 5-fluorouracil and camptothecin. *Cancer Res* **63**: 8791-8812.

Masters JR. 2000. Human cancer cell lines: fact and fantasy. *Nat Rev Mol Cell Biol* **1**: 233-236.

Mendy D, Gaidarova S, Brady H, Lopez-Girona A. 2009. Abstract# 1266: Lenalidomide treatment interferes with Ras/MAPK activation and induces apoptosis in multiple myeloma. *Cancer Research* **69**: 1266-1266.

Menges CW, Sementino E, Talarchek J, Xu J, Chernoff J, Peterson JR, Testa JR. 2012. Group I p21-Activated Kinases (PAKs) Promote Tumor Cell Proliferation and Survival through the AKT1 and Raf–MAPK Pathways. *Molecular Cancer Research* **10**: 1178-1188.

Milosavljevic N, Blanchard A, Wahl ML, Harguindey S, Poet M, Counillon L, Rauch C. 2011. Teaching new dogs old tricks: membrane biophysical properties in drug delivery and resistance. *Recent Pat Anticancer Drug Discov* **6**: 334-346.

Niu N, Qin Y, Fridley BL, Hou J, Kalari KR, Zhu M, Wu TY, Jenkins GD, Batzler A, Wang L. 2010. Radiation pharmacogenomics: a genome-wide association approach to identify radiation response biomarkers using human lymphoblastoid cell lines. *Genome Res* **20**: 1482-1492.

Orchel J, Witek L, Kimsa M, Strzalka-Mrozik B, Kimsa M, Olejek A, Mazurek U. 2012. Expression patterns of kinin-dependent genes in endometrial cancer. *Int J Gynecol Cancer* **22**: 937-944.

Palmer AC. 2016. Chemical probes: The many genes of drug mechanism. *Nat Chem Biol* **12**: 57-58.

Park JH, Kang JH, Jo UH, Gil EY, Park JK, Lee ES, Kim YH, Kim IS, Park KH. 2012. Caveolin-1 modulates docetaxel activity by inducing p53 expression in breast cancer. *Cancer Research* **72**: 4895-4895.

Park JH, Kim KP, Ko JJ, Park KS. 2016. PI3K/Akt/mTOR activation by suppression of ELK3 mediates chemosensitivity of MDA-MB-231 cells to doxorubicin by inhibiting autophagy. *Biochem Biophys Res Commun* **477**: 277-282.

Reddy SP, Britto R, Vinnakota K, Aparna H, Sreepathi HK, Thota B, Kumari A, Shilpa B, Vrinda M, Umesh S. 2008. Novel glioblastoma markers with diagnostic and prognostic value identified through transcriptome analysis. *Clinical Cancer Research* **14**: 2978-2987.

Rees MG, Seashore-Ludlow B, Cheah JH, Adams DJ, Price EV, Gill S, Javaid S, Coletti ME, Jones VL, Bodycombe NE. 2016. Correlating chemical sensitivity and basal gene expression reveals mechanism of action. *Nature chemical biology* **12**: 109-116.

Sandholt IS, Olsen BB, Guerra B, Issinger O-G. 2009. Resorufin: a lead for a new protein kinase CK2 inhibitor. *Anti-cancer drugs* **20**: 238-248.

Scherf U, Ross DT, Waltham M, Smith LH, Lee JK, Tanabe L, Kohn KW, Reinhold WC, Myers TG, Andrews DT et al. 2000. A gene expression database for the molecular pharmacology of cancer. *Nat Genet* **24**: 236-244.

Sculley D. 2007. Rank Aggregation for Similar Items. In *SDM*, pp. 587-592. SIAM.

Seguin L, Desgrosellier JS, Weis SM, Cheresh DA. 2015. Integrins and cancer: regulators of cancer stemness, metastasis, and drug resistance. *Trends Cell Biol* **25**: 234-240.

Shang X, Lin X, Manorek G, Howell SB. 2013. Claudin-3 and claudin-4 regulate sensitivity to cisplatin by controlling expression of the copper and cisplatin influx transporter CTR1. *Mol Pharmacol* **83**: 85-94.

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**: 2498-2504.

Song S, Honjo S, Jin J, Chang SS, Scott AW, Chen Q, Kalhor N, Correa AM, Hofstetter WL, Albarracin CT et al. 2015. The Hippo Coactivator YAP1 Mediates EGFR Overexpression and Confers Chemoresistance in Esophageal Cancer. *Clin Cancer Res* **21**: 2580-2590.

Stolarczyk EI, Reiling CJ, Pickin KA, Coppage R, Knecht MR, Paumi CM. 2012. Casein kinase 2α regulates multidrug resistance-associated protein 1 function via phosphorylation of Thr249. *Molecular pharmacology* **82**: 488-499.

Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, Simonovic M, Roth A, Santos A, Tsafou KP et al. 2015. STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res* **43**: D447-452.

Terakawa T, Miyake H, Kumano M, Sakai I, Fujisawa M. 2010. The antiandrogen bicalutamide activates the androgen receptor (AR) with a mutation in codon 741 through the mitogen

activated protein kinase (MARK) pathway in human prostate cancer PC3 cells. *Oncol Rep* **24**: 1395-1399.

Termglinchan V, Wanichnopparat W, Suwanwongse K, Teeyapant C, Chatpermporn K, Leerunyakul K, Chuadpia K, Sirimaneethum O, Wijitworawong P, Mutirangura W et al. 2013. Candidate cancer-targeting agents identified by expression-profiling arrays. *Onco Targets Ther* **6**: 447-458.

Thiago LS, Costa ES, Lopes DV, Otazu IB, Nowill AE, Mendes FA, Portilho DM, Abreu JG, Mermelstein CS, Orfao A et al. 2010. The Wnt signaling pathway regulates Nalm-16 b-cell precursor acute lymphoblastic leukemic cell line survival and etoposide resistance. *Biomed Pharmacother* **64**: 63-72.

Tong HH, Faloutsos C, Pan JY. 2006. Fast random walk with restart and its applications. *Ieee Data Mining*: 613-622.

Trojan L, Schaaf A, Steidler A, Haak M, Thalmann G, Knoll T, Gretz N, Alken P, Michel MS. 2005. Identification of metastasis-associated genes in prostate cancer by genetic profiling of human prostate cancer cell lines. *Anticancer Res* **25**: 183-191.

Tubi LQ, Gurrieri C, Brancalion A, Bonaldi L, Bertorelle R, Manni S, Pavan L, Lessi F, Zambello R, Trentin L. 2013. Inhibition of protein kinase CK2 with the clinical-grade small ATP-competitive compound CX-4945 or by RNA interference unveils its role in acute myeloid leukemia cell survival, p53-dependent apoptosis and daunorubicin-induced cytotoxicity. *Journal of hematology & oncology* **6**: 1.

Uehara N, Kanematsu S, Miki H, Yoshizawa K, Tsubura A. 2012. Requirement of p38 MAPK for a cell-death pathway triggered by vorinostat in MDA-MB-231 human breast cancer cells. *Cancer Lett* **315**: 112-121.

Vangipuram SD, Buck SA, Lyman WD. 2012. Wnt pathway activity confers chemoresistance to cancer stem-like cells in a neuroblastoma cell line. *Tumour Biol* **33**: 2173-2183.

Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, Stothard P, Chang Z, Woolsey J. 2006. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res* **34**: D668-672.

Yang W, Soares J, Greninger P, Edelman EJ, Lightfoot H, Forbes S, Bindal N, Beare D, Smith JA, Thompson IR et al. 2013. Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res* **41**: D955-961.

Yi JS, Mun DG, Lee H, Park JS, Lee JW, Lee JS, Kim SJ, Cho BR, Lee SW, Ko YG. 2013. PTRF/cavin-1 is essential for multidrug resistance in cancer cells. *J Proteome Res* **12**: 605-614.

Yu DD, Wu Y, Shen HY, Lv MM, Chen WX, Zhang XH, Zhong SL, Tang JH, Zhao JH. 2015. Exosomes in development, metastasis and drug resistance of breast cancer. *Cancer Sci* **106**: 959-964.

Yu Z, Xing Y. 2006. atRA-induced apoptosis of mouse embryonic palate mesenchymal cells involves activation of MAPK pathway. *Toxicol Appl Pharmacol* **215**: 57-63.

Zanin S, Borgo C, Girardi C, O'Brien SE, Miyata Y, Pinna LA, Donella-Deana A, Ruzzene M. 2012. Effects of the CK2 inhibitors CX-4945 and CX-5011 on drug-resistant cells. *PLoS One* **7**: e49193.

Zhang N, Wang H, Fang Y, Wang J, Zheng X, Liu XS. 2015. Predicting Anticancer Drug Responses Using a Dual-Layer Integrated Cell Line-Drug Network Model. *PLoS Comput Biol* **11**: e1004498.

Zhang Q, Major MB, Takanashi S, Camp ND, Nishiya N, Peters EC, Ginsberg MH, Jian X, Randazzo PA, Schultz PG et al. 2007. Small-molecule synergist of the Wnt/beta-catenin signaling pathway. *Proc Natl Acad Sci U S A* **104**: 7444-7448.