Citation: CPT Pharmacometrics Syst. Pharmacol. (2015) 4, 415–425; doi:10.1002/psp4.51

ORIGINAL ARTICLE

IDEA: Integrated Drug Expression Analysis—Integration of Gene Expression and Clinical Data for the Identification of Therapeutic Candidates

MH Ung[1], FS Varn[1] and C Cheng[1,2,3]*

Cancer drug discovery is an involved process spanning efforts from several fields of study and typically requires years of research and development. However, the advent of high-throughput genomic technologies has allowed for the use of *in silico*, genomics-based methods to screen drug libraries and accelerate drug discovery. Here we present a novel approach to computationally identify drug candidates for the treatment of breast cancer. In particular, we developed a Drug Regulatory Score similarity metric to evaluate gene expression profile similarity, in the context of drug treatment, and incorporated time-to-event patient survival information to develop an integrated analysis pipeline: Integrated Drug Expression Analysis (IDEA). We were able to predict drug candidates that have been known and those that have not been known in the literature to exhibit anticancer effects. Overall, our method enables quick preclinical screening of drug candidates for breast cancer and other diseases by using the most important indicator of drug efficacy: survival.

CPT Pharmacometrics Syst. Pharmacol. (2015) **4**, 415–425; doi:10.1002/psp4.51; published online on 18 June 2015.

Study Highlights

WHAT IS THE CURRENT KNOWLEDGE ON THE TOPIC? ☑ Large-scale genomic projects have allowed for the utilization of drug treatment profiles and disease profiles to aid in drug discovery. However, the direct clinical usefulness of these data has not been demonstrated in a rigorous manner. • WHAT QUESTION DID THIS STUDY ADDRESS? ☑ This investigation aimed to integrate patient survival information from large-scale breast cancer genomic datasets to evaluate drug efficacy. It also introduces a novel approach to assessing the similarity between drug treatment profiles and disease gene expression profiles. • WHAT THIS STUDY ADDS TO OUR KNOWLEDGE ☑ Molecular information captured in drug treatment profiles derived from cancer cell lines can be used in conjunction with clinical and gene expression data from primary tumors to identify novel breast cancer drugs. It introduces an *in silico* screening method that can be introduced into current preclinical drug discovery pipelines to make them faster and more efficient. • HOW THIS MIGHT CHANGE CLINICAL PHARMACOLOGY AND THERAPEUTICS ☑ The results from this study introduce an effective computational drug screening schema that can modify how new therapeutic compounds are identified. Instead of using time-consuming *in vitro* screening experiments to identify lead compounds, it may be possible to narrow down potential leads computationally before committing to more resource-intensive methods. Additionally, it suggests that tumor gene expression profiles can be used to guide drug treatment regimens in the clinic.

Early-stage attrition of candidate therapeutic compounds for cancer is commonplace in today's drug discovery pipelines due to the complexity associated with drug action. As such, the combination of large-scale data integration and computing power shows significant promise in transforming drug discovery into a more analytical, systematic, and comprehensive enterprise.[1] In recent years, the Connectivity Map (CMap) has gained popularity by providing an invaluable resource containing drug effect information obtained through large-scale molecular profiling of drug-treated cell lines.[2] Many methods have been proposed to utilize these vast resources of genomic information to accelerate drug development, including key studies by Sirota *et al*., Dudley *et al*., and Hassane *et al.*[1,3–12]

Despite the encouraging results of these studies, using treatment profile similarity based on *ex vivo* or *in vivo* cell line experiments as proxies for overall drug effect/action similarity has certain limitations, especially if it is to be extrapolated to clinical settings.[13] Thus, in this study we hypothesized that the inclusion of cancer patient time-to-event survival information into drug treatment screening procedures can be used to more accurately identify and reposition drug candidates. In addition, we claim that it is also possible to stratify patients into different prognostic groups using this approach. Specifically, we integrated breast cancer gene expression and clinical survival information with CMap drug treatment profiles (DTPs) in a systematic analysis to identify lead candidate therapeutics for breast cancer.[2] In our approach, we circumvent obstacles encountered by Amelio *et al.* including incomplete drug-protein interactomes and determining *P*-value cutoffs by developing the Drug Regulatory Score (DRS) metric system that is more sensitive to gene expression changes than

[1]Department of Genetics, Geisel School of Medicine at Dartmouth, Hanover, New Hampshire, USA; [2]Institute for Quantitative Biomedical Sciences, Geisel School of Medicine at Dartmouth, Lebanon, New Hampshire, USA; [3]Norris Cotton Cancer Center, Geisel School of Medicine at Dartmouth, Lebanon, New Hampshire, USA. *Correspondence: C Cheng (chao.cheng@dartmouth.edu)

Received 8 January 2015; accepted 30 April 2015; published online on 18 June 2015. doi:10.1002/psp4.51

methods that involve defining gene sets and calculating enrichment as performed by Amelio *et al.*, or using simple correlation analysis of DTPs.[14,15] The DRS captures information about the magnitude of change of all genes in a tumor gene expression profile (GEP) and utilizes the entire DTP, which is more representative of the overall effect of a drug on the final phenotype.

Broadly, we interrogated all MCF7 cell line DTPs from CMap against primary breast tumor GEPs to derive a DRS; a modified similarity metric adapted from our previous study.[2,16,17] We then implemented Cox proportional hazards survival analysis using DRS as input to identify drugs associated with patient survival.[18] Furthermore, we show that DRS can predict tumor metastatic potential in patients from an independent dataset published by Vijver *et al.* using unsupervised clustering and a random forest machine learning classifier.[19] Finally, we verify that DRS can be used as a predictive marker for paclitaxel-based neoadjuvant chemotherapy using another independent dataset published by Hatzis *et al.*[20]

METHODS

Datasets

Normalized METABRIC ($n = 2{,}136$) breast cancer gene expression datasets were downloaded from the European Genome-Phenome Archive (http://www.ebi.ac.uk/ega/) under the accession number EGAS00000000083.[16] Normal samples were removed leaving a total of 1,996 samples for our analysis. The normalized Vijver ($n = 260$) dataset was downloaded from the Netherlands Cancer Institute's data portal (http://ccb.nki.nl/data/). The normalized Hatzis ($n = 508$) and Ur-Rehman ($n = 1{,}570$) breast cancer gene expression datasets were downloaded from the gene expression omnibus (GEO) under the accession numbers GSE25066 and GSE47561, respectively.[20,21] The Hatzis dataset consisted of HER2-negative breast cancer tumors GEPs measured prior to neoadjuvant paclitaxel treatment. Time-to-event clinical information was collected over the course of neoadjuvant paclitaxel treatment.[20] All datasets included time-to-event survival information and other clinical information including subtype, tumor grade, metastatic potential, etc. Raw DTPs (.CEL files) derived from MCF7, HL60, and PC3 cell lines and supplementary drug information were downloaded from the Connectivity Map data portal (https://www.broadinstitute.org/cmap/).[2]

Data preprocessing of DTPs

Robust Microarray Analysis (RMA) was used for background correction of .CEL DTPs ($n = 1{,}215$), followed by quantile normalization, and fitting of a multichip linear model to each probe set; all techniques were implemented using the "affy" library from Bioconductor.[22] Probe sets were collapsed based on average intensity values. Gene fold-change was calculated by taking the ratio of treatment and control intensities followed by a $\log_2$-transformation. Genes with a $\log_2$ fold-change >0 were labeled as the "up" group, and genes with a $\log_2$ fold-change <0 were labeled as the "dn" group. Log-transformed values were z-transformed so that intensity values followed a standard normal distribution $[log_2(treatment/control) \sim N(0,1)]$. From this distribution, a two-sided *P*-value was derived for each gene and $-\log_{10}$-transformed to yield exponentially distributed values ($-\log_{10}$(p-value) $\sim$ Exp(λ)), which assigned heavier weights to more statistically significant genes when calculating DRS scores. Values >20 were set to 20; all values were then divided by 20 so that they took on a value with a range of 0 to 1. The final set of relative values corresponding to a drug was defined as the drug treatment profile or DTP.

Calculation of DRS

For each DTP, a DRS was calculated for each patient GEP using a modified version of an algorithm published by Zhu *et al.*[17] Namely, each DTP was treated as a vector $\mathbf{d} = [d_1,d_2,d_3...d_j...d_n]$; where $d_j = -\log(P\text{-value})$ and $n = \#$ of genes. We then defined a patient's molecular profile as the vector $\mathbf{g} = \langle g_1,g_2,g_3...g_j...g_n\rangle$;, containing the sorted (decreasing) gene expression intensities for each gene g_j. **d** was sorted according to **g**. We then calculated a pre-DRS (pDRS) for the "up" and "down" gene categories for each DTP by applying the following formulas: First, we calculate the foreground *f(i)* and background *b(i)* functions.

$$f(i) = \frac{\sum_{j=1}^{i} |g_j d_j|}{\sum_{j=1}^{n} |g_j d_j|}, 1 \leq i \leq n \tag{1}$$

$$b(i) = \frac{\sum_{j=1}^{i} |g_j (1-d_j)|}{\sum_{j=1}^{n} |g_j (1-d_j)|}, 1 \leq i \leq n \tag{2}$$

Second, we calculate the pDRS for the "up" group ($pDRS_{up}$) and the "dn" group ($pDRS_{dn}$), separately.

$$pDRS^{+}_{up/dn} = \max[f(i_{\max}) - b(i_{\max}), 0], \text{ where } i_{\max} = \text{argmax}_{i=1,2,3...n}[f(i) - b(i)] \tag{3a}$$

$$pDRS^{-}_{up/dn} = \min[f(i_{\min}) - b(i_{\min}), 0], \text{ where } i_{\min} = \text{argmin}_{i=1,2,3...n}[f(i) - b(i)] \tag{3b}$$

$$pDRS_{up/dn} = \begin{cases} pDRS^{+}, \; pDRS^{+} > |pDRS^{-}| \\ pDRS^{-}, \; otherwise \end{cases} \tag{3c}$$

The $pDRS_{up/dn}$ is then normalized by 1) permuting the gene labels in the patient profile m times (e.g., m = 1,000) and repeating steps 1–3 to yield a null $pDRS_{up/dn}$ distribution and 2) dividing by the average of the null to yield $DRS_{up/dn}$. We then computed DRS_{up} – DRS_{dn} to yield the final DRS. For each breast cancer dataset, we constructed a matrix where each element contained the DRS for each drug-sample pair (**Supplementary File 1**).

Survival analysis

Survival analysis was carried out for each drug, using the drug's DRS profile and patient clinical information included in breast cancer datasets. A univariate Cox proportional hazards model was fitted for each drug using the DRS as the independent variable.[18] The model is formulized below:

$$h(t|DRS) = h_0(t)e^{\beta_{DRS} * DRS}$$

where *h(t|DRS)* is the hazard function given the DRS and $h_0(t)$ is the baseline hazard, both at time *t*. We also fitted multivariate models to correct for potential confounding clinical factors by including patient age at diagnosis, tumor grade, tumor size, tumor stage, estrogen receptor (ER) status, HER2 status, and PR status as additional model covariates. The multivariate model is formulized below:

$$h(t|\mathbf{x}) = h_0(t)e^{\sum_i^n \beta_i x_i}$$

where **x** is the design matrix containing the drug's DRS profile and clinicopathological covariates, and their associated value for each patient. Schoenfeld residuals were analyzed to evaluate the proportional hazards assumption for all fitted models. The Wald test was used to assess significance of model parameters and outputted *P*-values were adjusted for multiple hypothesis testing using the Benjamini-Hochberg procedure (all *P*-values presented in the **Results** were adjusted unless noted otherwise).[23] Drugs yielding a $P < 1\text{E-}5$ were presumed to exhibit a significant pharmacological effect in breast cancer patients. Kaplan-Meier estimators and log-rank tests were used to compare survival rates between patients stratified on DRS. Patients were generally split at $DRS = 0$; if this yielded disproportionate sample groups, patients were stratified at DRS_{median} instead. Survival analysis was implemented using the "survival" R package (**Supplementary File 1**).

GO enrichment analysis

For ciclosporin, genes with a fold-change <0.5 between treatment and control samples were used as the downregulated gene list that was inputted into the DAVID functional annotation tool (http://david.abcc.ncifcrf.gov/) to calculate enrichment of GO Biological Process terms in the downregulated gene set.[24]

Machine learning analysis

Unsupervised hierarchical clustering of the patients in the Vijver dataset was performed using DRS profiles that exhibited a statistically significant difference between metastatic and nonmetastatic tumors.[19] In total, 84 DRS profiles yielded $P < 0.005$ (Wilcoxon test), and were included as features in the clustering analysis. A random forest classifier was trained using the same features to predict the metastatic identity of tumors based on the 84 DRS profiles. Ten-fold cross-validation and calculation of the area under the curve (AUC) of the receiver operating characteristic (ROC) curve were used to evaluate model performance. Clustering analysis and the random forest model were implemented using the R packages "gplots" and "randomForest," respectively.

Drug information

Information about US Food and Drug Administration (FDA)-approved drugs was derived from the National Cancer Institute's cancer drug information webpage (http://www.cancer.gov/cancertopics/druginfo/breastcancer), which also provides links to more detailed drug descriptions. Literature information about drugs mentioned in this study is provided in the **Supplementary Information.**

RESULTS

Overview of analysis

Our computational pipeline begins by first calculating a DRS between a DTP and a tumor GEP for each drug-tumor pair; a high DRS indicates that the baseline GEP of a tumor sample closely reflects the DTP that is induced by a drug, and vice versa for a low DRS **(Figure 1)**. A Cox proportional hazards model was then fitted to the DRS profile (DRS across all patient tumors) of each drug to evaluate its association with disease-specific patient survival **(Figure 1)**.[18] If the DRS profile of a drug was significantly correlated with patient survival, we considered the drug to be a potential therapeutic candidate. Moreover, to underscore the application of our methodology to precision medicine, we modified our models to identify candidates in several molecular subtypes of breast cancer.

To further evaluate DRS as an effective metric, we demonstrated that DRS captures the molecular differences between tumors by training a random forest classifier that could differentiate tumors with high and low metastatic proclivities using DRS across significant drugs as features. Additionally, we confirmed that there was a statistically significant difference between the DRS of patients with different residual cancer burden (RCB) scores. Overall, Integrated Drug Expression Analysis (IDEA) identifies drug candidates that can modulate the baseline expression of primary breast tumors in a way that affects patient survival time.

Systematic identification of survival associated drugs in breast cancer

Global summary of results. By hypothesizing that drugs with DRS profiles that significantly correlate with patient survival are pharmacologically active, we were able to identify several candidates using the METABRIC dataset from Curtis *et al*. **(Supplementary Table S1)**.[16] These candidates belonged to a variety of pharmacological classes including known antineoplastic agents, antioxidants, hormone therapeutics, and immunosuppressants **(Supplementary Chart S1)**. To note, the CMap dataset contains data corresponding to several replicate treatment experiments with varying concentrations for each drug.[2] Thus, we chose the replicate that yielded the most significant *P*-value from the survival analysis to represent the drug. Replicates where higher drug concentrations were used typically yielded more significant results. **Figure 2** shows the *P*-value and hazard ratio (HR) distribution (Cox proportional hazards model) of all drugs after selecting the most significant replicate. In total, there were 169 drugs (out of 1,215 drugs) that yielded $P < 1\text{E-}05$ with 110 of these having HR <1 and 59 having HR >1, when all samples were included in the analysis **(Supplementary Table S1)**. We reasoned that drugs whose DRS profiles yield $P < 1\text{E-}5$ from the model will exhibit a pharmacologic effect (therapeutic or toxic). For example, alpha-estradiol was predicted to have an effect with $P = 1.1\text{E-}20$, Wald test, HR = 0.91, along with scoulerine with $P = 1.3\text{E-}15$, Wald test, HR = 1.1 **(Supplementary Table S1)**. **Figure 2b,c** shows Kaplan-Meier plots for these two drugs, respectively, where patients are stratified by DRS. Patients with high alpha-estradiol DRS

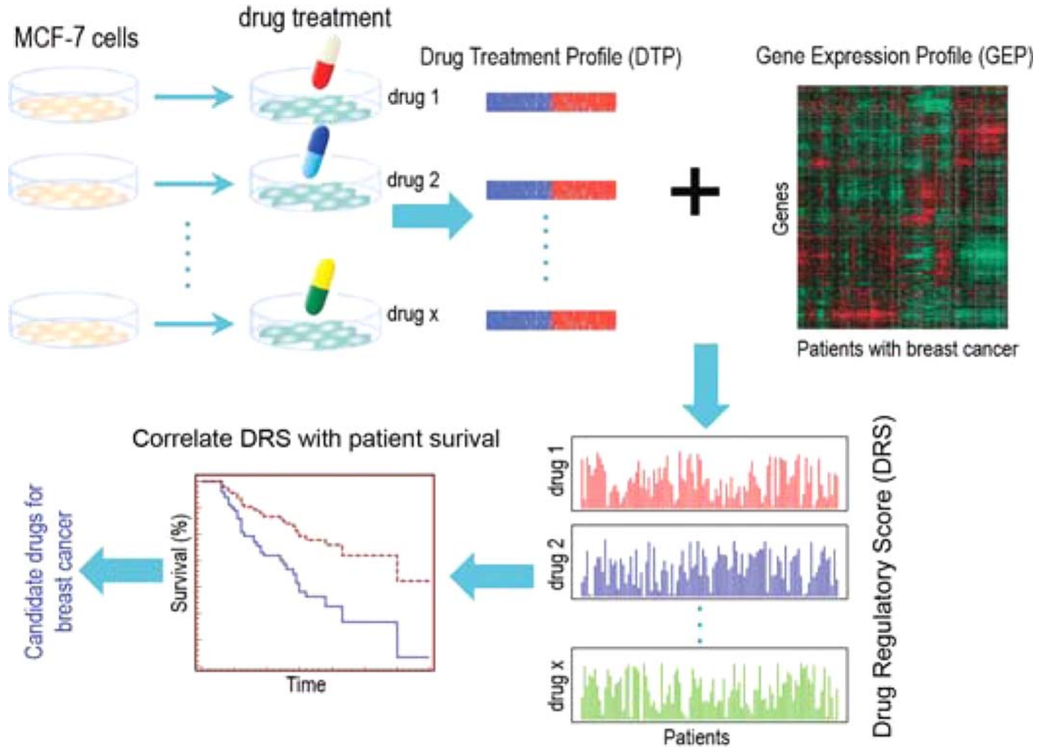

MCF-7 cells
drug treatment
drug 1
drug 2
drug x
Drug Treatment Profile (DTP)
Gene Expression Profile (GEP)
Genes
Patients with breast cancer
Correlate DRS with patient surival
Candidate drugs for breast cancer
Survival (%)
Time
drug 1
drug 2
drug x
Patients
Drug Regulatory Score (DRS)

Figure 1 Overview of IDEA. Treatment and control DTPs derived from MCF-7 cell lines were downloaded from CMap and combined to construct a DTP. DTPs were then compared against GEPs (GEP) of each patient from the breast cancer dataset of interest to generate a DRS. This resulted in a DRS profile for each drug. Each DRS profile was then used as the covariate(s) in Cox proportional hazards models. All drugs with DRS profiles that were significantly associated with patient survival were considered potential therapeutic candidates.

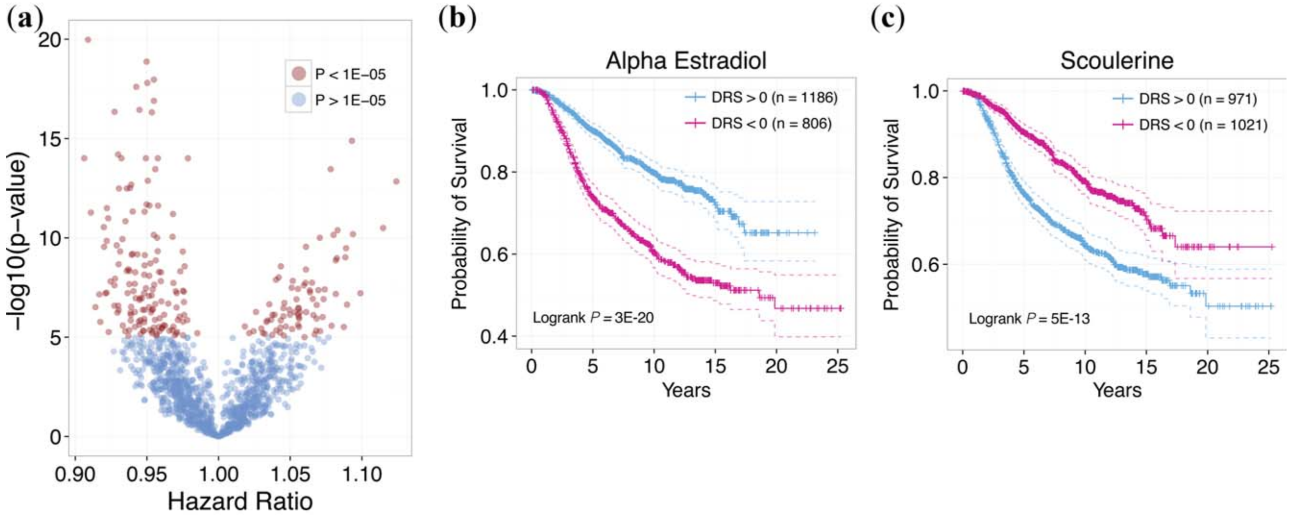


Figure 2 Output of IDEA. (**a**) Distribution of hazard ratios and *P*-values for all drugs. Each point corresponds to a drug, with red points corresponding to drugs with $P <$ 1E-5 and blue points corresponding to drugs with $P >$ 1E-5. Drugs with HR <1 indicates that survival is correlated with increased similarity of the DTP with breast cancer GEPs. Drugs with HR >1 indicates that survival is anticorrelated with increased similarity of the DTP with breast cancer GEPs. (**b**) Kaplan-Meier plot of patients with DRS >0 and DRS <0 for alpha estradiol. Patients with high DRS exhibit significantly more favorable prognosis than patients with low DRS ($P=$ 3E-20, Logrank test). (**c**) Kaplan-Meier plot of patients with DRS >0 and DRS <0 for scoulerine. Patients with low DRS exhibit significantly more favorable prognosis than patients with high DRS ($P=$ 5E-13, Logrank test).

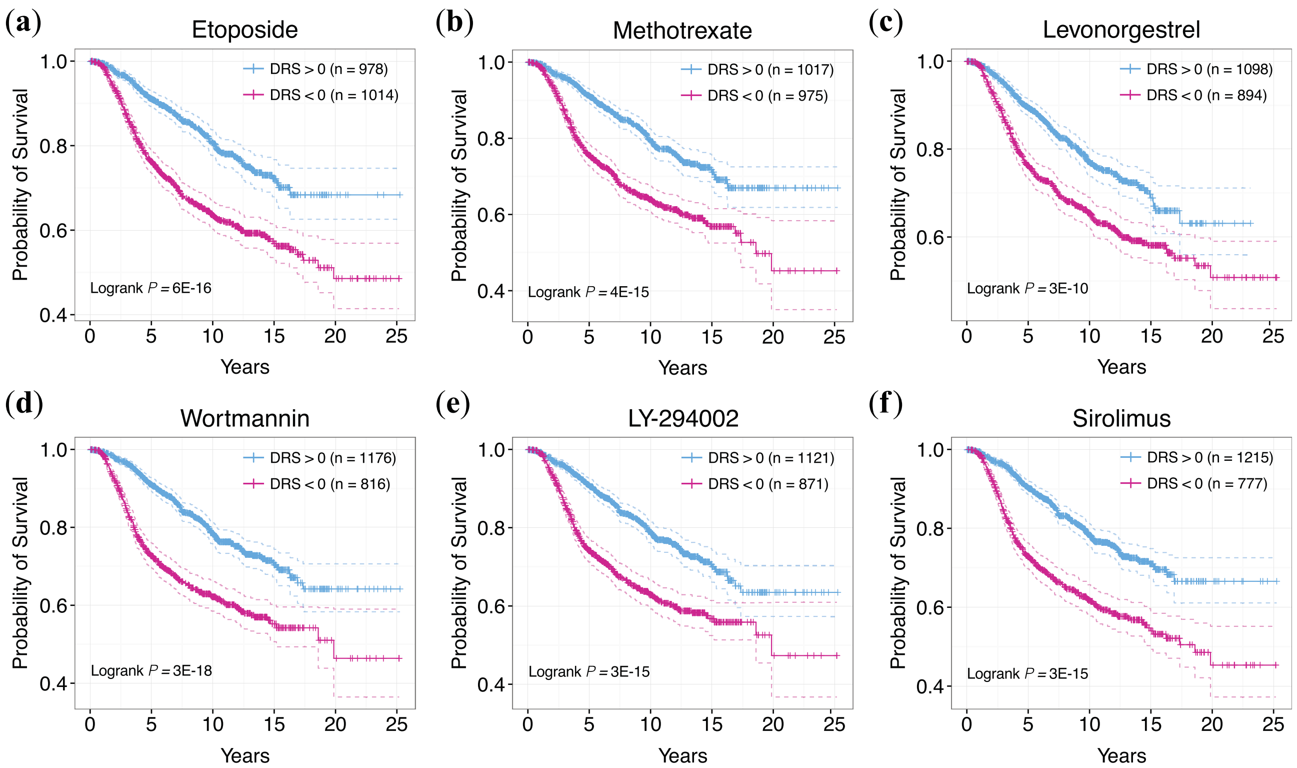


Figure 3 Examples of drug candidates belonging to different pharmacological classes. (**a**) Kaplan-Meier curves of patients with DRS >0 and DRS <0 for etoposide. (**b**) Kaplan-Meier curves of patients with DRS >0 and DRS <0 for methotrexate. (**c**) Kaplan-Meier curves of patients with DRS >0 and DRS <0 for levonorgestrel. (**d**) Kaplan-Meier curves of patients with DRS >0 and DRS <0 for wortmannin. (**e**) Kaplan-Meier curves of patients with DRS >0 and DRS <0 for LY-294002. (**f**) Kaplan-Meier curves of patients with DRS >0 and DRS <0 for sirolimus.

exhibited improved survival, whereas patients with low scoulerine DRS exhibited better prognosis.

Association of model coefficients with drug effect. In light of these results, we note that the interpretability of the HR is limited in the context of this study. Specifically, DTPs yielding an HR <1 indicates that patients with GEPs similar to that induced by the drugs in MCF7 cell lines (high DRS) tend to have better prognosis. This may suggest that patients with lower DRS would be most responsive to the drug because their GEPs are most dissimilar to the drug's DTP, possibly indicating that the drug may "reverse" the patients' GEPs to mimic that of high DRS patients. Alternatively, it may be that patients with high DRS would be more responsive because it may be easier for the drug to enhance an already similar GEP rather than act against a dissimilar one. Indeed, these two possibilities are not mutually exclusive, given the complexity of drug action, and may vary depending on the drug (see **Discussion**). Despite the fact that predicted drugs may exhibit a general therapeutic or toxic effect, it is reasonable to conclude that not every patient will respond similarly to each drug. Hence, our analysis also allows us to identify patients who will not respond in a way similar to that of the general sample population. More specifically, patient samples can be stratified based on their DRS. For example, if we stratify patients into groups based on their DRS for alpha estradiol, patients with DRS >0 exhibited a more favorable prognosis than patients with DRS <0 (P = 3E-20, Wald test, **Figure 2b**). This indicates that, in general, alpha-estradiol may have a pharmacological effect in breast cancer patients but only a select group may be responsive. This demonstrates that we can both evaluate the overall effect of drugs and subsequently predict individual patient response to them based on their DRS.

Identification of chemotherapeutics, PI3K inhibitors, and novel drug indications. Moreover, we identified chemotherapy drugs in our analysis including etoposide and methotrexate **(Figure 3a,b)**. These results suggest that our analysis was able to identify known anticancer agents, thus validating our results.[25,26] Second, we identified lovastatin and levonorgestrel, both of which are currently undergoing clinical trials for the treatment of breast and ovarian cancer, respectively (ClinicalTrials.gov Identifiers: NCT00285857 (lovastatin), NCT00445887 (levonorgestrel)) **(Supplementary Chart S1, Figure 3c)**.[27]

In support of our hypothesis that survival information can reveal bioactive drug leads, we were able to identify a number of drugs that have been experimentally shown to exhibit anticancer activity. In particular, we identified several phosphatidylinositol-4,5-bisphosphate 3-kinase (PI3K)

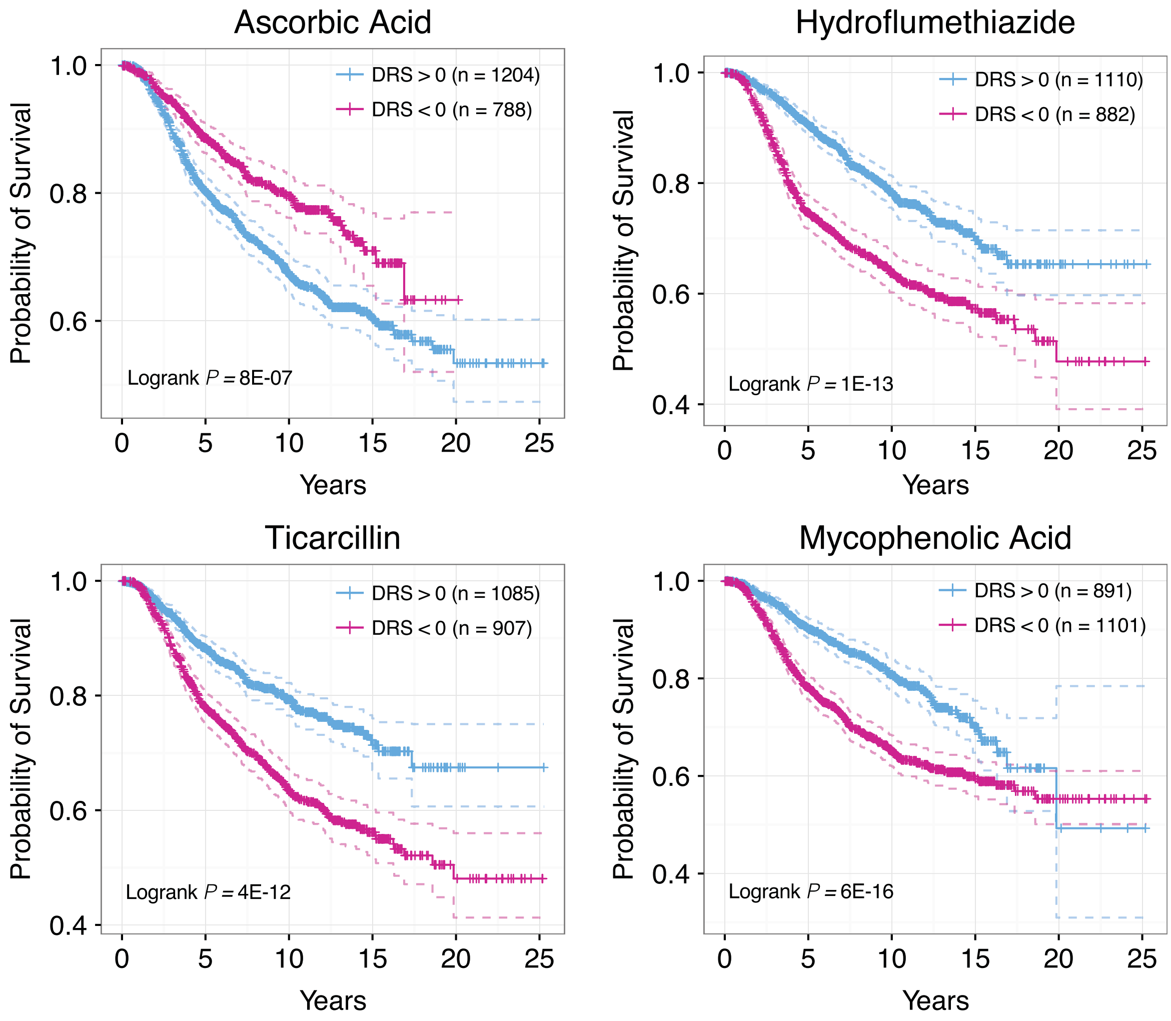
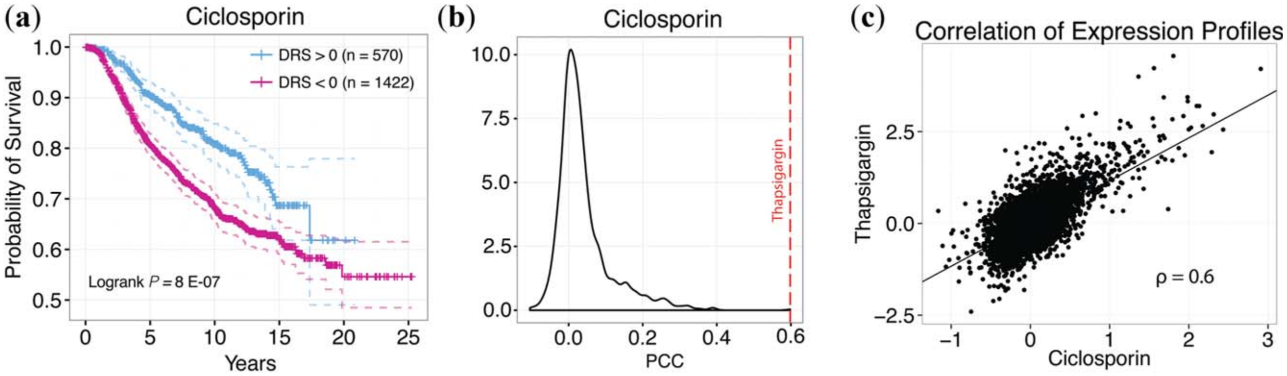


Figure 4 Examples of drug candidates not previously considered for chemotherapy. (**a**) Kaplan-Meier curves of patients with DRS >0 and DRS <0 for ascorbic acid. (**b**) Kaplan-Meier curves of patients with DRS >0 and DRS <0 for hydroflumethiazide. (**c**) Kaplan-Meier curves of patients with DRS >0 and DRS <0 for ticarcillin. (**d**) Kaplan-Meier curves of patients with DRS >0 and DRS <0 for mycophenolic acid.

pathway inhibitors that are known to inhibit proliferation, reduce angiogenesis, and induce apoptosis in cancer cell lines.[28–30] Identified PI3K/mTOR inhibitors include wortmannin, LY-294002, and sirolimus **(Figure 3d–f)**.[28,31–47] In addition to known drugs, we identified drug candidates that have not yet been experimentally tested or lack rigorous testing. Among these compounds are ciclosporin, ascorbic acid, hydroflumethiazide, ticarcillin, and mycophenolic acid **(Figures 4**, **5)**. To show that these candidates have therapeutic potential, we exemplify the case of ciclosporin ($P = 4.69$E-12, HR = 0.94, **Figure 5a)**, which has a DTP similar to that of thapsigargin, a drug shown to delay tumor growth in multiple cancer mouse xenograft models.[48,49] Ciclosporin's DTP was significantly correlated with the DTPs of only a few drugs, as shown by the distribution of correlation coefficients **(Figure 5b)**.[2] This suggests that the mechanism of action of ciclosporin, if it does possess anticancer activity, is inherently different from that of other top drugs identified in our analysis. Interestingly, ciclosporin was significantly correlated with thapsigargin in their GEPs with a Pearson's correlation coefficient of 0.6 **(Figure 5c)**. The fact that simple correlation analysis was able to detect a high similarity between the two DTPs suggests that ciclosporin may be related to thapsigargin in terms of drug effect. Additionally, we calculated the enrichment of Gene Ontology Biological Process terms in genes that were significantly downregulated between ciclosporin treatment and control groups **(Figure 5d, Supplementary Table S2)**. Interestingly, we found highly relevant cancer-associated pathways such as "steroid hormone receptor signaling pathway," "regulation of cell cycle," and "regulation of apoptosis" enriched in the downregulated gene set, suggesting that ciclosporin may have a bioactive effect on cancer growth. Overall, these results indicate that the drugs we identified that are not currently recognized as cancer therapeutics may in fact be strong candidates for further experimental testing.



(d)

Term	P-Value	Fold Enrichment	Bonferroni	Benjamini	FDR
GO:0030518~steroid hormone receptor signaling pathway	5.61E-06	6.53	1.27E-02	1.27E-02	0.01
GO:0030522~intracellular receptor-mediated signaling pathway	9.74E-06	5.51	2.20E-02	1.11E-02	0.02
GO:0006974~response to DNA damage stimulus	1.19E-05	2.58	2.68E-02	9.01E-03	0.02
GO:0051726~regulation of cell cycle	3.63E-05	2.6	7.96E-02	2.05E-02	0.06
GO:0010942~positive regulation of cell death	6.87E-05	2.29	1.45E-01	3.09E-02	0.12
GO:0043068~positive regulation of programmed cell death	1.55E-04	2.23	2.98E-01	5.72E-02	0.27
GO:0033554~cellular response to stress	2.41E-04	2.01	4.23E-01	7.54E-02	0.42
GO:0051789~response to protein stimulus	2.69E-04	3.86	4.59E-01	7.40E-02	0.47
GO:0009615~response to virus	3.17E-04	3.79	5.15E-01	7.72E-02	0.55
GO:0043065~positive regulation of apoptosis	3.35E-04	2.16	5.35E-01	7.37E-02	0.58
GO:0010941~regulation of cell death	3.92E-04	1.77	5.92E-01	7.82E-02	0.68
GO:0030521~androgen receptor signaling pathway	5.31E-04	6.69	7.03E-01	9.61E-02	0.92
GO:0006281~DNA repair	6.28E-04	2.42	7.62E-01	1.05E-01	1.09
GO:0043067~regulation of programmed cell death	7.03E-04	1.74	7.99E-01	1.08E-01	1.22
GO:0010605~negative regulation of macromolecule metabolic process	7.22E-04	1.78	8.08E-01	1.04E-01	1.25

Figure 5 Ciclosporin DTP. (**a**) Kaplan-Meier curves of patients with DRS >0 and DRS <0 for ciclosporin. (**b**) Empirical distribution of Pearson correlation coefficients (PCC) from comparing the ciclosporin DTP with all other DTPs. Red line indicates PCC from comparing ciclosporin DTP with thapsigargin DTP. (**c**) Scatterplot comparing ciclosporin and thapsigargin DTPs. Each point corresponds to a single gene in the DTP. (**d**) GO enrichment analysis of downregulated genes between treatment and control DTPs for ciclosporin.

Drugs association analysis in stratified breast cancer samples

In the clinic, immunohistochemical categorization of breast cancer is an essential component of determining patient prognosis and designing treatment regimens. Thus, the effectiveness of drug treatment may vary according to ER, p53, and/or HER2 expression. As such, we predicted drug candidates for patients belonging to each of the six breast cancer histological subtypes to increase the resolution of our analysis and to approach drug discovery from a precision medicine standpoint. We found several significant drug candidates for ER+ (P < 1E-05, Wald Test), ER– (P < 0.05, Wald Test), p53+ (P < 1E-04, Wald Test), and HER2– (P < 1E-04, Wald Test) breast cancer subtypes and no significant candidates for p53– and HER2+ subtypes **(Supplementary Tables S3–S8)**. Furthermore, to demonstrate that drug efficacy varies across breast cancer subtypes, we compared drugs between ER+ (n = 1,518) and ER– (n = 474) breast cancer groups. **Figure 6a,b** shows the global P-value and hazard ratio distribution of each drug for ER+ and ER– tumors, respectively. Comparing the top 100 drug candidates from each subtype, we found that only 10 drugs were common between the two groups. This suggests that ER status is an important predictive biomarker for drug efficacy. For instance, wortmannin yielded P = 8.42E-12 (Wald test) and HR = 0.94 in ER+ samples but was not significant in ER– samples **(Figure 6c,d)**. Conversely, repaglinide was significant in ER– samples (P = 0.02, Wald test) but not significant in ER+ samples **(Figure 6e,f)**. These results suggest that our analysis is sensitive to differences in prognostic features exhibited by different tumor groups. Indeed, being able to identify drugs for a specific subset of patients is powerful in that it allows for the development of tailored treatments based on optimal drug efficacy.

Integrative model for metastasis prediction using drug regulatory scores

Since metastasis is an important driver of cancer progression, we postulated that there would be a relationship between metastatic tumors and their DRS profiles. If DRS does indeed capture information regarding metastatic tendency, then the distribution of DRS should be different between metastatic and nonmetastatic tumors. Thus, we also aimed to evaluate the predictive power of DRS in stratifying patients based on whether their tumors became metastatic by applying IDEA to an independent breast cancer dataset from van de Vijver *et al.*[19] We first identified 84 DRS profiles that exhibited the greatest difference between metastatic and nonmetastatic tumor samples and then clustered samples based on these DRS profiles to derive three apparent clusters corresponding to the metastatic identity of the samples (see **Methods**). The first cluster had 48 metastatic samples out of 103 (47%), the second cluster

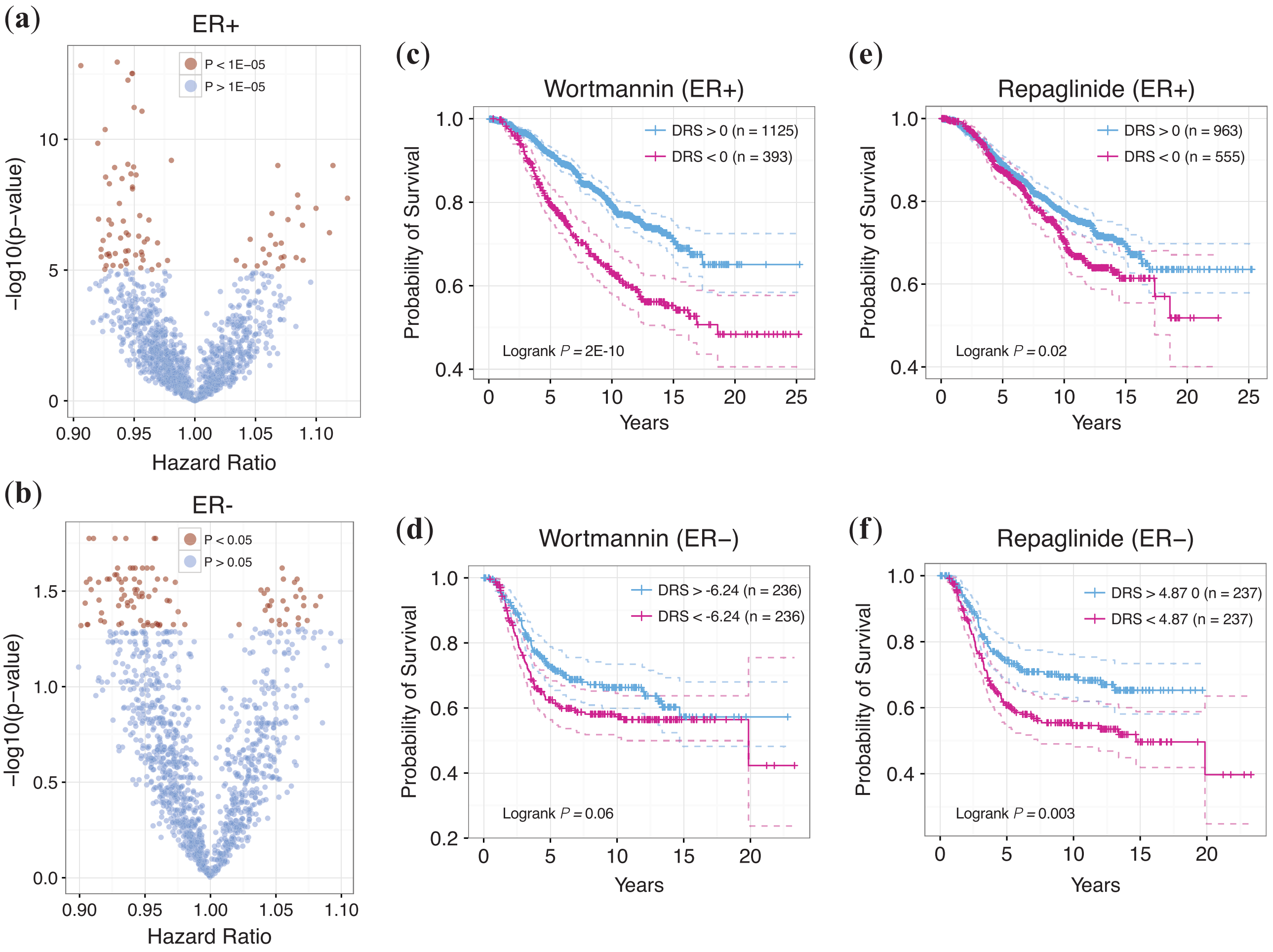


Figure 6 Drug candidates for ER+ and ER– patients. (**a**) *P*-value and hazard ratio distribution for all drugs for ER+ patients. (**b**) *P*-value and hazard ratio distribution for all drugs for ER– patients. Red points indicate drugs with $P < 1E\text{-}5$ and blue points indicate drugs with $P > 1E\text{-}5$. (**c**) Kaplan-Meier curves of patients with and for wortmannin for ER+ patients. (**d**) Kaplan-Meier curves of patients with and for wortmannin for ER– patients. (**e**) Kaplan-Meier curves of patients with and for repaglinide for ER+ patients. (**f**) Kaplan-Meier curves of patients with and for wortmannin for ER– patients. For ER– patients, stratifying patients at yields disproportionate sample sizes; therefore, the median DRS was used instead.

had 27 metastatic samples out of 99 (27%), and the third cluster had eight metastatic clusters out of 58 (14%) **(Figure 7a)**. This indicates that DRS profiles can distinguish differences in metastatic potential between breast cancer tumors, and implies that treatment with a subset (cluster) of these drugs may predispose a tumor to take on a prometastatic molecular identity. Conversely, another subset may reverse a tumor's tendency to metastasize. To further evaluate this idea, we trained a random forest classifier with the same DRS profile features to determine if it could accurately stratify metastatic and nonmetastatic tumors, and evaluated its performance using 10-fold cross-validation. The model achieved an AUC of 0.71 calculated from the ROC curve, suggesting that DRS is an informative predictor when classifying tumors based on a known survival-associated phenotype **(Figure 7b)**. Furthermore, we implemented IDEA in the van de Vijver *et al.* dataset and were able to validate 28 drugs ($P < 0.05$, Wald test) including wortmannin, LY-294002, and etoposide **(Supplementary Table S9)**. Additionally, we were able to validate 20 of the 25 top candidates in an independent meta-dataset compiled and normalized by Ur-Rehman *et al.*,[21] indicating that our methodology is robust across different datasets. **(**$P < 0.05$, **Supplementary Table S10)**.

Predicting patient response to neoadjuvant chemotherapy using drug regulatory scores

Because our previous results were derived using survival data that were collected over the course of entire treatment periods, we aimed to determine if DRS could stratify patients that would respond well to neoadjuvant chemotherapy. Thus, we focused on the paclitaxel DRS profile since paclitaxel is a common chemotherapeutic administered in the clinic. Following our claim that the efficacy of a drug can be determined by its DRS profile, we constructed a Kaplan-Meier estimator for patient survival in the

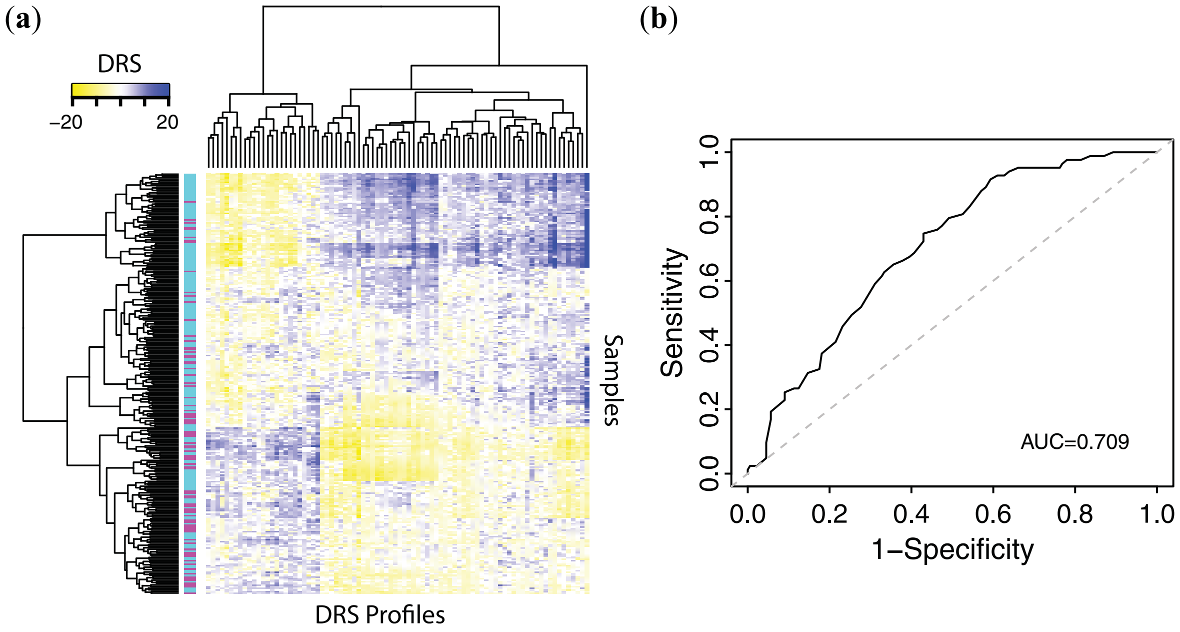


Figure 7 Use of DRS profiles to predict metastasis. (**a**) Hierarchical clustering of DRS profiles. Magenta sample labels indicate metastatic tumors and aqua sample labels indicate nonmetastatic tumors. (**b**) Receiver operating characteristic curve for random forest machine learning classifier, using 84 most significant DTPs.

METABRIC dataset using the paclitaxel DRS profile and observed that patients with DRS >0 show poor survival **(Supplementary Figure S1A)**. Although patients in the METABRIC dataset with DRS >0 exhibited poor prognosis, the observed difference in survival rates is still indicative of pharmacological activity. We then applied the paclitaxel DTP to an independent breast cancer dataset, published by Hatzis *et al.*,[20] that includes patients who were administered neoadjuvant paclitaxel therapy. We generated a DRS profile for paclitaxel in the Hatzis dataset and found that the DRS of RCB-I (low residual cancer burden) tumors were significantly larger than the DRS of RCB-II/III (high residual cancer burden) tumors **(**$P=0.01$, Wilcoxon test, **Supplementary Figure S1B)**. This indicates that the paclitaxel DRS profile can be used as a predictive marker. Additionally, this result suggests that the patients with DRS >0 (poor survival group) in the METABRIC dataset would have been more responsive to the effects of paclitaxel. Overall, we were able to validate predicted paclitaxel pharmacological activity in a separate dataset where patients were actually treated with paclitaxel.

DISCUSSION

In this study we introduce the DRS and show that by using it in conjunction with time-to-event clinical survival information we can identify novel cancer therapeutics for different breast cancer subtypes. Despite the success of our method, we remark that there are several limitations to our approach, many of which partially stem from a lack of appropriate data. First, interpreting the hazard ratio is difficult in that we, as of yet, do not understand the biological implications regarding high-matching profiles (high DRS) and low-matching profiles (low DRS). In other words, it is unknown whether patients with high DRS will benefit from the drug due to enhancement of its current GEP or if patients with low DRS will benefit more because the drug will induce an opposite effect. Indeed, both mechanisms are simultaneously possible, making interpretation even more difficult. Moreover, this implies that our drug candidates may either be therapeutic or toxic in nature, and the effects may and probably do vary across individuals. Second, our method is based on the assumption that DTPs provide a functional readout of a drug's overall biological activity, which may not always be the case. Likewise, a patient's GEP may not always reflect the molecular events responsible for cancer development. Regardless, we hold that we are able to identify bioactive agents that are indeed associated with patient survival and that this drug identification schema is a reasonable first *in silico* screening step. Finally, we note that DTPs are derived from MCF7 cell lines, which are ER+, and thus our results may be less accurate in ER– tumor samples. However, we reimplemented IDEA using wortmannin, LY-294002, and sirolimus DTPs from PC3 and HL60 cell lines to show that drug treatment effects remain stable across different cell types **(Supplementary Table S11)**. More generally, cell lines themselves may not be completely representative of actual patient tumors, which are heterogeneous and complex in nature. Despite issues with tissue specificity, cell lines are still informative about the core genes that drive cancer development. We also note that our datasets contain more ER+ samples than ER– samples, which may be why no significant drug candidates were identified for ER– tumor samples after correcting for clinical confounders.

Despite these limitations, our method improves on the current state of *in silico* drug screening, by using the terminal effect of a drug and molecular profiling of patient tumors to guide preclinical drug development. Furthermore, our approach can be extended to other cancers and diseases for which molecular profiles and clinical information are available. As more drug and clinical data become available, this analytical pipeline can further be modified to incorporate additional information. Moreover, IDEA can be applied to any disease in which GEPs and terminal phenotype data (i.e., survival, flare-ups, outbreaks) are available. Overall, we have presented a novel, integrated, and flexible approach to *in silico* drug profiling that can achieve results consistent with the known drug literature, identify novel therapeutic candidates, and be potentially applied to other diseases.

Acknowledgments. This work was supported by the American Cancer Society Research grant IRG-82-003-270, the Centers of Biomedical Research Excellence (COBRE) grant GM103534, and the start-up funding package provided to CC by the Geisel School of Medicine at Dartmouth College.

Conflict of Interest. The authors declare no conflicts of interest. The funders had no role in the design of the method, data selection and analysis, decision to publish, or preparation of the article.

Author Contributions. M.H.U. and F.S.V. wrote the article; C.C. designed the research; C.C., M.H.U., and F.S.V. performed the research; C.C. and M.H.U. analyzed the data; C.C. contributed new reagents/analytical tools.

1. Hurle, M.R., Yang, L., Xie, Q., Rajpal, D.K., Sanseau, P. & Agarwal, P. Computational drug repositioning: from data to therapeutics. *Clin. Pharmacol. Ther.* **93**, 335–341 (2013).
2. Lamb, J. *et al.* The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* **313**, 1929–1935 (2006).
3. Sirota, M. *et al.* Discovery and preclinical validation of drug indications using compendia of public gene expression data. *Sci. Transl. Med.* **3**, 96ra77 (2011).
4. Dudley, J.T. *et al.* Computational repositioning of the anticonvulsant topiramate for inflammatory bowel disease. *Sci. Transl. Med.* **3**, 96ra76 (2011).
5. Hassane, D.C. *et al.* Discovery of agents that eradicate leukemia stem cells using an in silico screen of public gene expression data. *Blood* **111**, 5654–5662 (2008).
6. Jahchan, N.S. *et al.* A drug repositioning approach identifies tricyclic antidepressants as inhibitors of small cell lung cancer and other neuroendocrine tumors. *Cancer Disc.* **3**, 1364–1377 (2013).
7. Kinnings, S.L., Liu, N., Buchmeier, N., Tonge, P.J., Xie, L. & Bourne, P.E. Drug discovery using chemical systems biology: repositioning the safe medicine Comtan to treat multi-drug and extensively drug resistant tuberculosis. *PLoS Comput. Biol.* **5**, e1000423 (2009).
8. van Noort, V. *et al.* Novel drug candidates for the treatment of metastatic colorectal cancer through global inverse gene-expression profiling. *Cancer Res.* **74**, 5690–5699 (2014).
9. Iorio, F. *et al.* Discovery of drug mode of action and drug repositioning from transcriptional responses. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 14621–14626 (2010).
10. Cheng, F. *et al.* Prediction of drug-target interactions and drug repositioning via network-based inference. *PLoS Comput Biol.* **8**, e1002503 (2012).
11. Zhao, H. *et al.* Novel modeling of cancer cell signaling pathways enables systematic drug repositioning for distinct breast cancer metastases. *Cancer Res.* **73**, 6149–6163 (2013).
12. Lussier, Y.A. & Chen, J.L. The emergence of genome-based drug repositioning. *Sci. Transl. Med.* **3**, 96ps35 (2011).
13. Crockett, S.D., Schectman, R., Sturmer, T. & Kappelman, M.D. Topiramate use does not reduce flares of inflammatory bowel disease. *Dig. Dis. Sci.* **59**, 1535–1543 (2014).
14. Amelio, I., Gostev, M., Knight, R.A., Willis, A.E., Melino, G. & Antonov, A.V. DRUGSURV: a resource for repositioning of approved and experimental drugs in oncology based on patient survival information. *Cell Death Dis.* **5**, e1051 (2014).
15. Pacini, C. *et al.* DvD: An R/Cytoscape pipeline for drug repurposing using public repositories of gene expression data. *Bioinformatics* **29**, 132–134 (2013).
16. Curtis, C. *et al.* The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**, 346–352 (2012).
17. Zhu, M., Liu, C.C. & Cheng, C. REACTIN: regulatory activity inference of transcription factors underlying human diseases with application to breast cancer. *BMC Genomics* **14**, 504 (2013).
18. Cox, D.R. Regression models and life-tables. *J. R. Stat. Soc. B* **34**, 187 (1972).
19. van de Vijver, M.J. *et al.* A gene-expression signature as a predictor of survival in breast cancer. *N. Engl. J. Med.* **347**, 1999–2009 (2002).
20. Hatzis, C. *et al.* A genomic predictor of response and survival following taxane-anthracycline chemotherapy for invasive breast cancer. *JAMA* **305**, 1873–1881 (2011).
21. Ur-Rehman, S., Gao, Q., Mitsopoulos, C. & Zvelebil, M. ROCK: a resource for integrative breast cancer data analysis. *Breast Cancer Res. Treat.* **139**, 907–921 (2013).
22. Irizarry, R.A. *et al.* Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**, 249–264 (2003).
23. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate—a practical and powerful approach to multiple testing. *J. Roy. Stat. Soc. B Met.* **57**, 289–300 (1995).
24. Huang, D.W. *et al.* DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. *Nucleic Acids Res.* **35**, W169–175 (2007).
25. Wright, J.C., Prigot, A., Wright, B., Weintraub, S. & Wright, L.T. An evaluation of folic acid antagonists in adults with neoplastic diseases: a study of 93 patients with incurable neoplasms. *J. Natl. Med. Assoc.* **43**, 211–240 (1951).
26. Busse, D. *et al.* Pharmacokinetics of intravenous etoposide in patients with breast cancer: influence of dose escalation and cyclophosphamide and doxorubicin coadministration. *Naunyn Schmiedeberg's Arch. Pharmacol.* **366**, 218–225 (2002).
27. Corbelli, J. & Bimla Schwarz, E. Emergency contraception: a review. *Minerva Ginecol.* **66**, 551–564 (2014).
28. Li, J., Li, F., Wang, H., Wang, X., Jiang, Y. & Li, D. Wortmannin reduces metastasis and angiogenesis of human breast cancer cells via nuclear factor-kappaB-dependent matrix metalloproteinase-9 and interleukin-8 pathways. *J. Int. Med. Res.* **40**, 867–876 (2012).
29. Will, M. *et al.* Rapid induction of apoptosis by PI3K inhibitors is dependent upon their transient inhibition of RAS-ERK signaling. *Cancer Discov.* **4**, 334–347 (2014).
30. Fruman, D.A. & Rommel, C. PI3K and cancer: lessons, challenges and opportunities. *Nat. Rev. Drug Discov.* **13**, 140–156 (2014).
31. Schultz, R.M. *et al.* In vitro and in vivo antitumor activity of the phosphatidylinositol-3-kinase inhibitor, wortmannin. *Anticancer Res.* **15**, 1135–1139 (1995).
32. Yun, J., Lv, Y.G., Yao, Q., Wang, L., Li, Y.P. & Yi, J. Wortmannin inhibits proliferation and induces apoptosis of MCF-7 breast cancer cells. *Eur. J. Gynaecol. Oncol.* **33**, 367–369 (2012).
33. Howes, A.L. *et al.* The phosphatidylinositol 3-kinase inhibitor, PX-866, is a potent inhibitor of cancer cell motility and growth in three-dimensional cultures. *Mol. Cancer Ther.* **6**, 2505–2514 (2007).
34. Ihle, N.T. *et al.* Molecular pharmacology and antitumor activity of PX-866, a novel inhibitor of phosphoinositide-3-kinase signaling. *Mol. Cancer Ther.* **3**, 763–772 (2004).
35. Akter, R., Hossain, M.Z., Kleve, M.G. & Gealt, M.A. Wortmannin induces MCF-7 breast cancer cell death via the apoptotic pathway, involving chromatin condensation, generation of reactive oxygen species, and membrane blebbing. *Breast Cancer* **4**, 103–113 (2012).
36. Liu, Y., Shreder, K.R., Gai, W., Corral, S., Ferris, D.K. & Rosenblum, J.S. Wortmannin, a widely used phosphoinositide 3-kinase inhibitor, also potently inhibits mammalian polo-like kinase. *Chem. Biol.* **12**, 99–107 (2005).
37. Zhou, Y. *et al.* LY294002 inhibits the malignant phenotype of osteosarcoma cells by modulating the phosphatidylinositol 3kinase/Akt/fatty acid synthase signaling pathway in vitro. *Mol. Med. Rep.* **11**, 1352–1357 (2015).
38. Long, X.H. *et al.* LY294002 suppresses the malignant phenotype and sensitizes osteosarcoma cells to pirarubicin chemotherapy. *Mol. Med. Rep.* **10**, 2967–2972 (2014).
39. Zhao, K. *et al.* SN50 enhances the effects of LY294002 on cell death induction in gastric cancer cell line SGC7901. *Arch. Med. Sci. AMS* **9**, 990–998 (2013).
40. Qian, Y.J. *et al.* Effect of LY294002 on adriamycin-induced epithelial-mesenchymal transition in human breast carcinoma cells. *Zhongguo Yi Xue Ke Xue Yuan Xue Bao Acta Acad. Med. Sin.* **34**, 319–323 (2012).
41. Ma, J., Xie, S.L., Geng, Y.J., Jin, S., Wang, G.Y. & Lv, G.Y. In vitro regulation of hepatocellular carcinoma cell viability, apoptosis, invasion, and AEG-1 expression by LY294002. *Clin. Res. Hepatol. Gastroenterol.* **38**, 73–80 (2014).
42. Carminati, P.O., Donaires, F.S., Marques, M.M., Donadi, E.A., Passos, G.A. & Sakamoto-Hojo, E.T. Cisplatin associated with LY294002 increases cytotoxicity and induces changes in transcript profiles of glioblastoma cells. *Mol. Biol. Rep.* **41**, 165–177 (2014).

43. Wu, D. *et al.* Phosphatidylinositol 3-kinase inhibitor LY294002 suppresses proliferation and sensitizes doxorubicin chemotherapy in bladder cancer cells. *Urol. Int.* **87**, 105–113 (2011).
44. Yamada, T., Horinaka, M., Shinnoh, M., Yoshioka, T., Miki, T. & Sakai, T. A novel HDAC inhibitor OBP-801 and a PI3K inhibitor LY294002 synergistically induce apoptosis via the suppression of survivin and XIAP in renal cell carcinoma. *Int. J. Oncol.* **43**, 1080–1086 (2013).
45. Badinloo, M. & Esmaeili-Mahani, S. Phosphatidylinositol 3-kinases inhibitor LY294002 potentiates the cytotoxic effects of doxorubicin, vincristine, and etoposide in a panel of cancer cell lines. *Fundam. Clin. Pharmacol.* **28**, 414–422 (2014).
46. Yoshioka, T., Yogosawa, S., Yamada, T., Kitawaki, J. & Sakai, T. Combination of a novel HDAC inhibitor OBP-801/YM753 and a PI3K inhibitor LY294002 synergistically induces apoptosis in human endometrial carcinoma cells due to increase of Bim with accumulation of ROS. *Gynecol. Oncol.* **129**, 425–432 (2013).
47. Alayev, A., Berger, S.M., Kramer, M.Y., Schwartz, N.S. & Holz, M.K. The combination of rapamycin and resveratrol blocks autophagy and induces apoptosis in breast cancer cells. *J. Cell. Biochem.* **116**, 450–457 (2015).
48. Desmedt, C. *et al.* Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter independent validation series. *Clin. Cancer Res.* **13**, 3207–3214 (2007).
49. Schmidt, M. *et al.* The humoral immune system has a key prognostic impact in node-negative breast cancer. *Cancer Res.* **68**, 5405–5413 (2008).

Supplementary information accompanies this paper on the *CPT: Pharmacometrics & Systems Pharmacology* website (http://www.wileyonlinelibrary.com/psp4)