SCIENTIFIC REPORTS

OPEN

Discordancy Partitioning for Validating Potentially Inconsistent Pharmacogenomic Studies

Received: 28 June 2017
Accepted: 25 October 2017
Published online: 09 November 2017

J. Sunil Rao[1,2] & Hongmei Liu[1]

The Genomics of Drug Sensitivity in Cancer (GDSC) and Cancer Cell Line Encyclopedia (CCLE) are two major studies that can be used to mine for therapeutic biomarkers for cancers of a large variety. Model validation using the two datasets however has proved challenging. Both predictions and signatures do not consistently validate well for models built on one dataset and tested on the other. While the genomic profiling seems consistent, the drug response data is not. Some efforts at harmonizing experimental designs has helped but not entirely removed model validation difficulties. In this paper, we present a partitioning strategy based on a data sharing concept which directly acknowledges a potential lack of concordance between datasets and in doing so, also allows for extraction of reproducible novel gene-drug interaction signatures as well as accurate test set predictions. We demonstrate these properties in a re-analysis of the GDSC and CCLE datasets.

It's now widely believed that cancers are far more heterogeneous than once thought - that in fact, they represent a myriad of different diseases with varying biological determinants rather than a single entity whose effective treatment will rely on some over-arching theoretical construct of our understanding of the disease. To this end, high throughput pharmacogenomic screening of small molecules and other compounds has the potential to implicate new drug leads (or drug combinations) that can be used for more personalized treatments.

A typical pharmacogenomic workflow involves characterizing interesting compounds for dose-response effects on cancer cell lines, and then doing functional genomic characterization in additional screens. Of interest is the elucidation of therapeutic biomarkers whose patterns might be predictive of a compound's activity against a particular cancer cell line.

Given the complexity of such assays, the necessity of proper validation is paramount. Given a fresh set of data generated from the same workflow, one should be able to demonstrate both accurate predictions of drug activity as well as reproducibility of therapeutic genomic signatures.

The GDSC and CCLE datasets

The opportunity to study validation of models arose with the generation of two major pharmacogenomic datasets - the Genome Drug Sensitivity in Cancer (GDSC) project[1,2] and the Cancer Cell Line Encyclopedia (CCLE) project[3]. Both represent large-scale studies in which experimental and approved drug compounds were screened against panels of molecularly characterized cancer cell lines.

A number of recent studies looked at the 15 drugs and 706 cell lines which were in common between the two studies[4–6]. One study acted as a training dataset and the other the test dataset. Focus was put initially on examining concordance of drug response profiles and genomic profiles across datasets, then later moving to reproducibility of detected drug-gene interactions. Some concerning results were published[5] that reported poor reproducibility of drug response profiles but high concordance of genomic profiles. A more stringent analysis was then done[4] which used more biologically grounded analytical considerations yielding improved consistency of pharmacological data but still only reaching 67% of evaluable compounds. Given that the ultimate goal of such studies is to detect novel drug-gene interactions, successful validation also implies that training data models of drug-gene interactions should yield low test set prediction errors of drug response *and* reproducibility of therapeutic genomic signatures on the test dataset. *It should be noted however, that in general developing accurate signatures* (*those with true drug-gene interactions*) *does not always imply low prediction errors*. Overfit models are also

[1]Division of Biostatistics, Department of Public Health Sciences, University of Miami, Florida, USA. [2]Sylvester Comprehensive Cancer Center, University of Miami, Florida, USA. Correspondence and requests for materials should be addressed to J.S.R. (email: jrao@miami.edu)

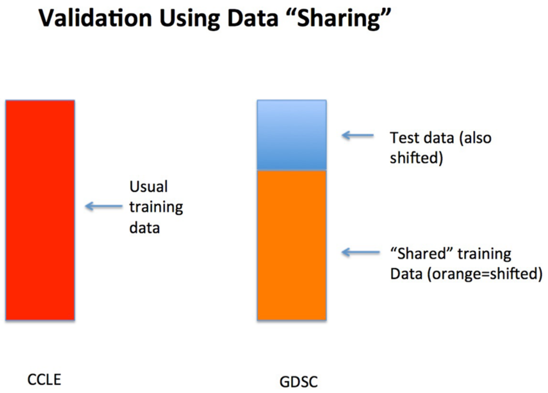
Validation Using Data "Sharing"
Usual training data
Test data (also shifted)
"Shared" training Data (orange=shifted)
CCLE
GDSC

Figure 1. Schematic of data sharing for the combined GDSC-CCLE analysis.

known to predict well (and these might also yield reproducible signatures) - however the test set error differences between overfit and accurate more sparse models is usually not large[7].

Some additional studies have been carried out by adding a third new drug screening dataset to the mix to see if improved concordance could be achieved[8–10]. In all cases, the GDSC-CCLE discordancies could not be completely resolved. For instance, a third independent dataset called the Genentech Cell Line Screening Initiative (gSCI)[10] was generated. Improved agreement appears to be present between gSCI and CCLE, and concordance with GDSC was weaker. However, careful reading of this paper shows that there are still issues with the improved agreement that was found. The proportion of shared genomic features found is low, association values are not compared (meaning the actual magnitudes and signs of the effects are not displayed), and the elastic net model they used was only fit once and all non-zero estimates were regarded as candidate biomarkers. Typically these models need proper setting of tuning parameters which are found by methods like cross-validation. Cross-validation in itself can be unstable and thus it's more robust to repeat the process many times and look at a more rigorous definition of a candidate biomarker in terms of repeated detection over many iterations of cross-validation. In another study, a third independent dataset from the Institute for Molecular Medicine Finland (FIMM)[9] was generated which shared a much more similar experimental protocol to CCLE than GDSC did. This dataset had a significantly higher level of consistency with CCLE than with GDSC although correlations across cell lines were markedly lower than between cell lines. Note that only 26 cell lines were shared while CCLE and GDSC share 268 cell lines, so even though the conclusions reached by the authors about the need to carefully standardize experimental protocols may be correct, a selection bias cannot be ruled out. Still the improved results does not completely resolve the discordancy issue. Recently, even a detailed re-analyses of an updated (and larger version) of the GDSC and CCLE data using many additional suggestions from the research community did not result in significantly improved validation findings[11]. The same trends of low concordance of drug profiles, high concordance of genomic profiles and poor detection of novel drug-gene interactions remain - and these patterns are not constant across drugs. So while new datasets and improved harmonization of protocols is certainly important, because validation is a multi-faceted concept, discordancy issues of one type or another may still remain - although our hope is over time, these will continue to be mitigated.

In this paper we take a more agnostic view: we assume that discordancies *may* exist for a given drug, and we build this assumption into our modeling framework. This is followed up with using shrinkage estimation of model parameters which can *zero out* discordancies *if in fact they are small*. That way, we can handle both scenarios - drugs that are concordant and those that might not be (whatever the underlying reasons). Our approach also can achieve both low test set prediction errors and detect reproducible gene-drug interactions - many of which we found to be novel. We develop a **discordancy partitioning** approach using the data sharing strategy[12,13] across datasets which can easily be extended to more than two datasets.

Methods

Figure 1 presents a schematic of our analysis strategy. We use all of the CCLE dataset and a *portion* of the GDSC dataset as training. The remainder of the GDSC dataset will be withheld from modeling and used purely for evaluating test set prediction errors. In order to have proper representation of all tissue types, we ensure that the portion of GDSC used for modeling contains cell lines sampled at random from all tissue types.

The goal of the analysis is to identify therapeutic genomic signatures for each drug (and potentially further by cancer type). The response variable y comes from the drug response assay and has typically been taken to be the $log(IC50)$ value or the drug-specific area under the curve (AUC) value from the dose-response curve. The predictors $\mathbf{x}$, come from the molecular assays and include, mutation status of known cancer genes, gene expression profiles, copy number variations and cancer tissue types. We consider the following two model formulations:

Formulation 1:

$$\begin{aligned} y_i &= \boldsymbol{\beta}^T \mathbf{x}_i + \varepsilon_i,\ i \in G \\ y_i &= (\boldsymbol{\beta} + \boldsymbol{\delta})^T \mathbf{x}_i + \varepsilon_i,\ i \in C, \end{aligned} \tag{1}$$

where $\in G$ or $\in C$ most generally represent two independent datasets, but here can represent cell line membership in GDSC or CCLE datasets respectively. The model dimensions are $y_i \in \mathbb{R}$, $\mathbf{x}_i \in \mathbb{R}^p$, $\boldsymbol{\beta} \in \mathbb{R}^p$, $\boldsymbol{\delta} \in \mathbb{R}^p$. The errors $\varepsilon_i \in \mathbb{R}$ are identically and independent distributed with zero mean and finite variance. Model (1) can be generalized to more than two datasets and can also allow for multiple drug responses where $\mathbf{y}_i \in \mathbb{R}^q$. We will report on a new model fitting strategy for the multivariate model elsewhere.

Model (1) formally partitions the common genomic effects across datasets ($\boldsymbol{\beta}$) from the potential dataset discordancies ($\boldsymbol{\delta}$). A version of this was called data shared lasso[12] which used lasso estimation[14] to induce sparsity in $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$. An earlier version of this approach traces to the data enriched linear regression[13]. We use a different constraint on the model parameters by using a generalized version of the elastic net[15]. Specifically, we add an additional L_2 constraint on the model parameters which allows finding sparse signal with potential grouped effects which represent correlated genomic features. The following penalized optimization function is used,

$$\begin{aligned}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\delta}}) \; = \; \arg\min \frac{1}{2}\sum_{i\in G}(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2 + \sum_{i\in C}(y_i - \mathbf{x}_i^T(\boldsymbol{\beta} + \boldsymbol{\delta}))^2 \\ +\lambda_1\left(\sum_{j=1}^{p}\omega_j\left|\beta_j\right| + r\sum_{j=1}^{p}\psi_j\left|\delta_j\right|\right) + \lambda_2(\|\boldsymbol{\beta}\|_2^2 + r\|\boldsymbol{\delta}\|_2^2),\end{aligned} \tag{2}$$

where $\lambda_1, \lambda_2 > 0$ are tuning parameters. The elastic net estimators were suggested to be used to construct the weights, where $\omega_j = \left|\hat{\beta}_j^{enet}\right|^{-1}$ and $\psi_j = \left|\hat{\delta}_j^{enet}\right|^{-1}$, because they are $\sqrt{n}$-consistent estimators and can guarantee variable selection consistency in (2), i.e. the true covariates in the model can be identified[16]. We let $r = 1/\sqrt{2}$ as recommended[12]. Ten-fold cross validation is used to optimize λ_1 and λ_2 from a pre-specified set of values.

To avoid a lack of a potential invariance due to the choice of the reference group in (1), an alternate formulation of the model can be made as represented in Formulation 2:

Formulation 2:

$$\begin{aligned} y_i &= (\boldsymbol{\beta} + \boldsymbol{\delta}_G)^T\mathbf{x}_i + \varepsilon_i,\ i \in G \\ y_i &= (\boldsymbol{\beta} + \boldsymbol{\delta}_C)^T\mathbf{x}_i + \varepsilon_i,\ i \in C, \end{aligned} \tag{3}$$

where $\boldsymbol{\delta}_G \in \mathbb{R}^p$ are discrepancy parameters of G, $\boldsymbol{\delta}_C \in \mathbb{R}^p$ are discrepancy parameters of C. Here $\boldsymbol{\beta}$ represents a baseline effect that is shared across datasets. Formulations 1 and 2 are two different ways to model discrepancies between datasets. However, the two models are potentially interchangeable. More specifically, denote as $\boldsymbol{\beta}_1$ the $\boldsymbol{\beta}$ in Formulation 1 and $\boldsymbol{\beta}_2$ the $\boldsymbol{\beta}$ in Formulation 2. Then one can show that $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 + \boldsymbol{\delta}_G$ and $\boldsymbol{\delta} = \boldsymbol{\delta}_C - \boldsymbol{\delta}_G$. We will restrict most of our attention to Formulation 1 since fewer parameters need to be estimated and fitted models are generally more stable as a result especially when sample size is small.

Simulation Design. In order to study performance of our proposed approach, we first conduct a simulation study where we know what the true level of dataset discordancy is and compare it against alternative strategies. Specifically, we consider the following simulation design.

Training data were generated from model (1) where $\mathbf{x}_i$ were i.i.d. from $N_p(\mathbf{0}, \boldsymbol{\Sigma})$ with $p = 1000$ and $\Sigma(i, j) = 0.75^{|i-j|}$, and ε_i was i.i.d. from $N(0, 1)$. The autoregressive covariance structure is to mimic high correlations between neighboring genes (i.e. this could be thought of as genes on the same biological pathway for instance). Also since not all 1000 features are likely to be correlated with the response variable, we assumed both $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ were sparse. To mimic grouped effects in genomic data, non-zero elements in $\boldsymbol{\beta}$ were distributed in 16 blocks with block size 3, and the non-zero values were randomly drawn from $N(1.5, 1)$. Moreover, we assumed that there are discordances in both non-zero effect genes and zero-effect genes. Twenty indices of non-zero entries and 10 indices of zero entries of $\boldsymbol{\beta}$ were randomly selected which constituted the indices of non-zero elements of $\boldsymbol{\delta}$, and the non-zero values were randomly drawn from $N(0, 0.5)$.

In the training dataset, we let the sample size of dataset G increase from 100 to 400 and the sample size of dataset C be fixed at 300. We also drew a test dataset with sample size 100 from the upper equation of model (1), i.e. from the same distribution as G dataset. We applied model (1) to training datasets G and C, and calculated the mean prediction error on test data using the shared models. The prediction performances were compared to other candidate models: individual model built on dataset G, individual model built on dataset C and individual model built on merged datasets of G and C. The simulation was repeated 200 times.

Note that the same input training datasets G and C can also be fitted by Formulation 2 and test set mean prediction error can be calculated. Moreover, as mentioned above, model parameters from Formulation 2 can be transformed into parameters in Formulation 1. We will not present the results here.

Analysis of GDSC and CCLE Datasets. *Data Preprocessing.* We used the drug measure AUC as a response variable. There are 15 drugs in common between the GDSC and CCLE datasets. While 706 cell lines are also shared between the two datasets and have associated genomic feature data, only a subset of these cell lines have AUC values in both datasets. By further removing cell lines with different SNP identity[11] and cell lines of which the number of missing values in feature data is greater than 10000, we get a range of 77–274 cell lines per drug (median = 88; mean = 162 cell lines).

The shared genomic feature data includes mutation status of 63 cancer genes, copy number variations of 24960 genes, 16150 gene expression profiles and 23 tissue types. Genomic features whose missing rate was larger than

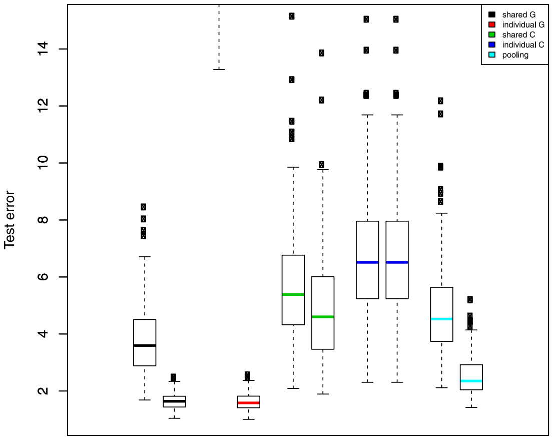

Figure 2. Simulation test set error boxplot pairs by analysis method. Left hand member of the pair represents a training set sample size of 100 and the right hand member, a training set sample size of 400. Test set sample sizes are always set a 300. Black represents predictions generated by using $\hat{\boldsymbol{\beta}}$ only estimated from the data sharing model across G and C; red represents predictions generated from a model using only dataset G; green represents predictions using $\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\delta}}$ only from the data sharing model across G and C; blue represents predictions generated from a model using only dataset C; and finally light blue, predictions from a model pooling datasets G and C.

0.3 were filtered. The remaining feature data had less than <0.1% missing values which were imputed by feature means. We standardized each covariate vector and drug response vector to have mean 0 and variance 1.

Data Analysis Strategy. In analysis of prediction accuracy, we applied model (1) to the shared portions of GDSC and CCLE datasets. Because the shared subset data only have a limited number of cell lines and many cell lines miss the copy number variation data, we excluded copy number variations from the feature set of the model. The data also went though a quality control by removing a few number of outliers.

A random sample of 20% GDSC dataset were withheld as test dataset. Similar to the design in simulation study, we let sample size of the GDSC training dataset vary to see effects of increased sharing. Thus in the first scenario, the CCLE dataset and a random sample of 70% of the 80% GDSC dataset were used as training data. In the second scenario, the CCLE dataset and the 80% GDSC dataset were used as training data. In each scenario, we used sure independence screening (SIS)[17] to reduce the number of covariates to a moderate size, where 2000 features most correlated with the response variable were selected respectively in *each* training CCLE and training GDSC dataset. A union of the two sets of selected features were used as input variables. Then model (1) was fitted to the screened training dataset. Mean prediction error for the 20% GDSC test dataset using either the shared GDSC model or shared CCLE model was calculated. This process was repeated 200 times to account for the variability of random sampling. We compared the mean prediction errors from the shared models against following methods: i) using the training GDSC dataset alone (this represents a gold standard subject to sample size limitations when discordancies exist between the datasets), ii) using the training CCLE dataset alone, iii) merging the training GDSC and training CCLE datasets as one common training dataset.

Then we switched the roles of GDSC and CCLE datasets and conducted a similar prediction accuracy analysis. More specifically, a random sample of 20% CCLE dataset were withheld as test dataset. We let sample size of the CCLE training dataset vary. In the first scenario, a random sample of 70% of the 80% CCLE dataset and the GDSC dataset were used as training data. In the second scenario, the 80% CCLE dataset and the GDSC dataset were used as training data. The rest are similar to previous prediction analysis where 20% GDSC dataset were withheld as test dataset.

For signature validation analysis, we ran a shared analysis using both datasets in full for the 15 drugs in common. All cell lines in GDSC and CCLE were used, and copy number variations were included into the feature set of the model because of the increased sample size available for analysis. Sure independence screening was used to reduce the number of covariates to a moderate size in a same way as in previous prediction accuracy analysis. Then model (1) was fitted to the screened GDSC and CCLE datasets. We examined the estimated $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ values in order to determine how many genomic markers reliably reproduced their effects across datasets, and how many effects were *washed away* because of dataset discordancies.

Covariates with non-zero estimates were determined to be predictors associated with the drug response variable. This procedure was repeated 200 times for each drug to assess the stability of covariates when applying ten-fold cross validation for tuning parameters selection in equation (2). A variable list was built for each drug. It consists of all covariates that appeared in any of the 200 runs along with a frequency that a feature appears and an average coefficient given to that feature over the runs it appears. The average coefficient was used to assess effect size of a feature to the drug response variable. Here, the most significant predictors associated with the drug response variable are defined as those with an effect size ±2 s.d. from the mean and a frequency rate ≥80%. We also fit Formulation 2 to the same screened GDSC and CCLE datasets. Estimates of $\boldsymbol{\beta} + \boldsymbol{\delta}_G$ and $\boldsymbol{\beta} + \boldsymbol{\delta}_C$ were examined.

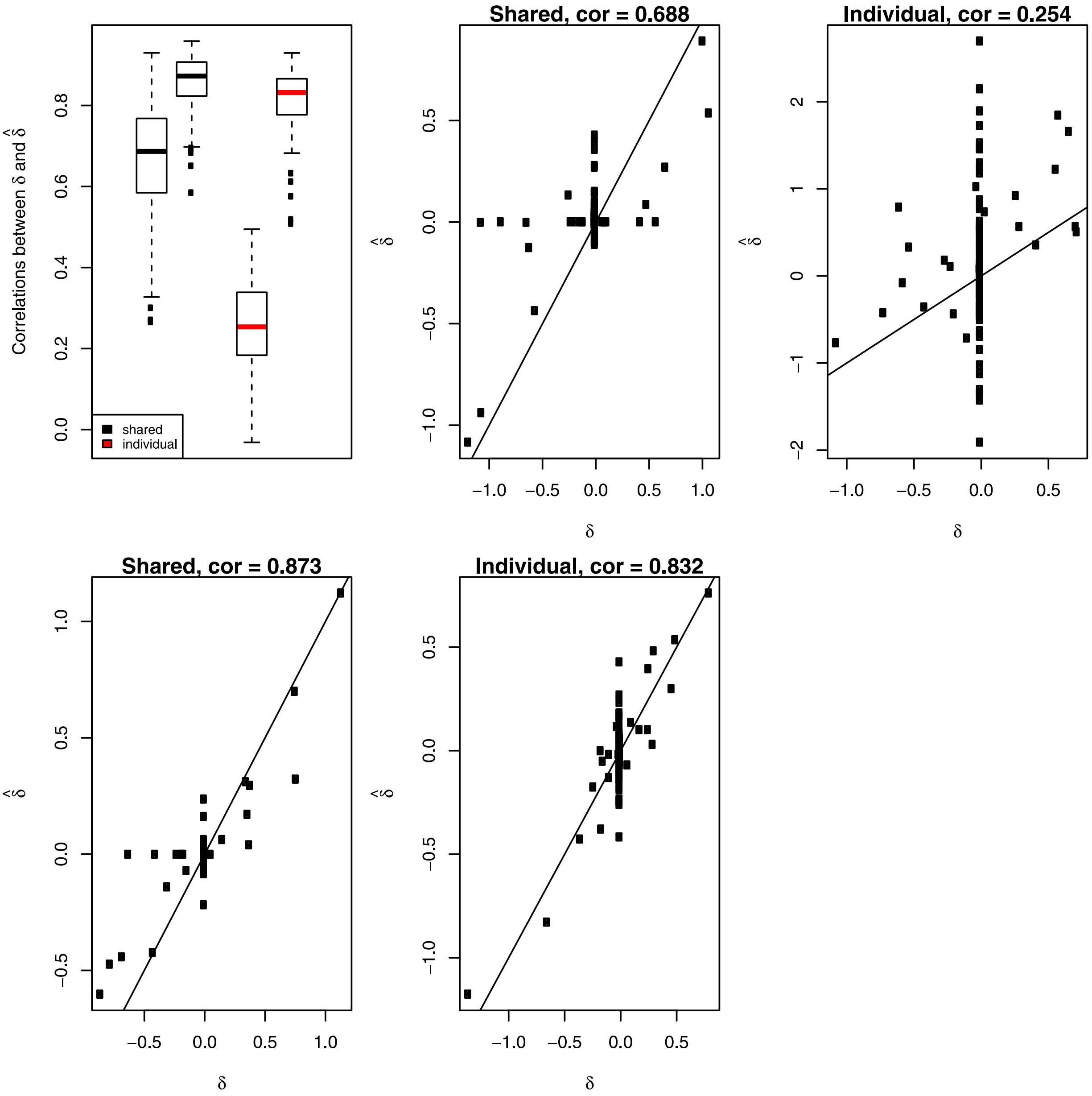

Figure 3. Top left panel: Simulation correlations between estimated $\boldsymbol{\delta}$ and the true δ across the 200 runs of the simulation for the shared model (black) and extracted from fitted individual models (red). The side by side boxplots represents the two different sample sizes in the simulation. All other panels show the scatterplots of estimated values of $\boldsymbol{\delta}$ versus true values at the median correlations from the boxplots in the top left panel.

Results

Simulation. Prediction results in the simulation study are summarized in Fig. 2. When sample sizes are small, the shared G model outperformed all other methods and individual G model has much larger test errors than other methods do (which exceeds the range of the plot). Note that non-overlapping interquartile ranges of boxplots signify clearly statistically significant differences. Also, interquartile ranges that may overlap but do not overlap with another boxplot's median value also indicate statistically significant differences in prediction error performance. As sample size increases, the individual G model serves as a gold standard (as the test data were from this model), and more importantly the shared G model performs clearly as well as this gold standard. Since there are non-negligible disparities between data G and C due to the $\boldsymbol{\delta}$ in our simulations, the shared/individual C model and individual aggregated model never beat the shared G model in all cases.

Correlations of $\hat{\boldsymbol{\delta}}$ from model (1) with the true $\boldsymbol{\delta}$ in 200 simulations were calculated and present in boxplot of Fig. 3. For comparison, we subtracted $\hat{\boldsymbol{\beta}}_G$ (estimate of $\boldsymbol{\beta}$ from individual G model) from $\hat{\boldsymbol{\beta}}_C$ (estimate of $\boldsymbol{\beta}$ from individual C model), then calculated its correlation with the true $\boldsymbol{\delta}$. Not surprisingly, the shared method designed for estimating dataset discrepancies produced much higher correlations than alternative method did, see Fig. 3. We also calculated correlations of $\hat{\boldsymbol{\beta}}$ from model (1) with the true $\boldsymbol{\beta}$ in 200 simulations and compared them to the

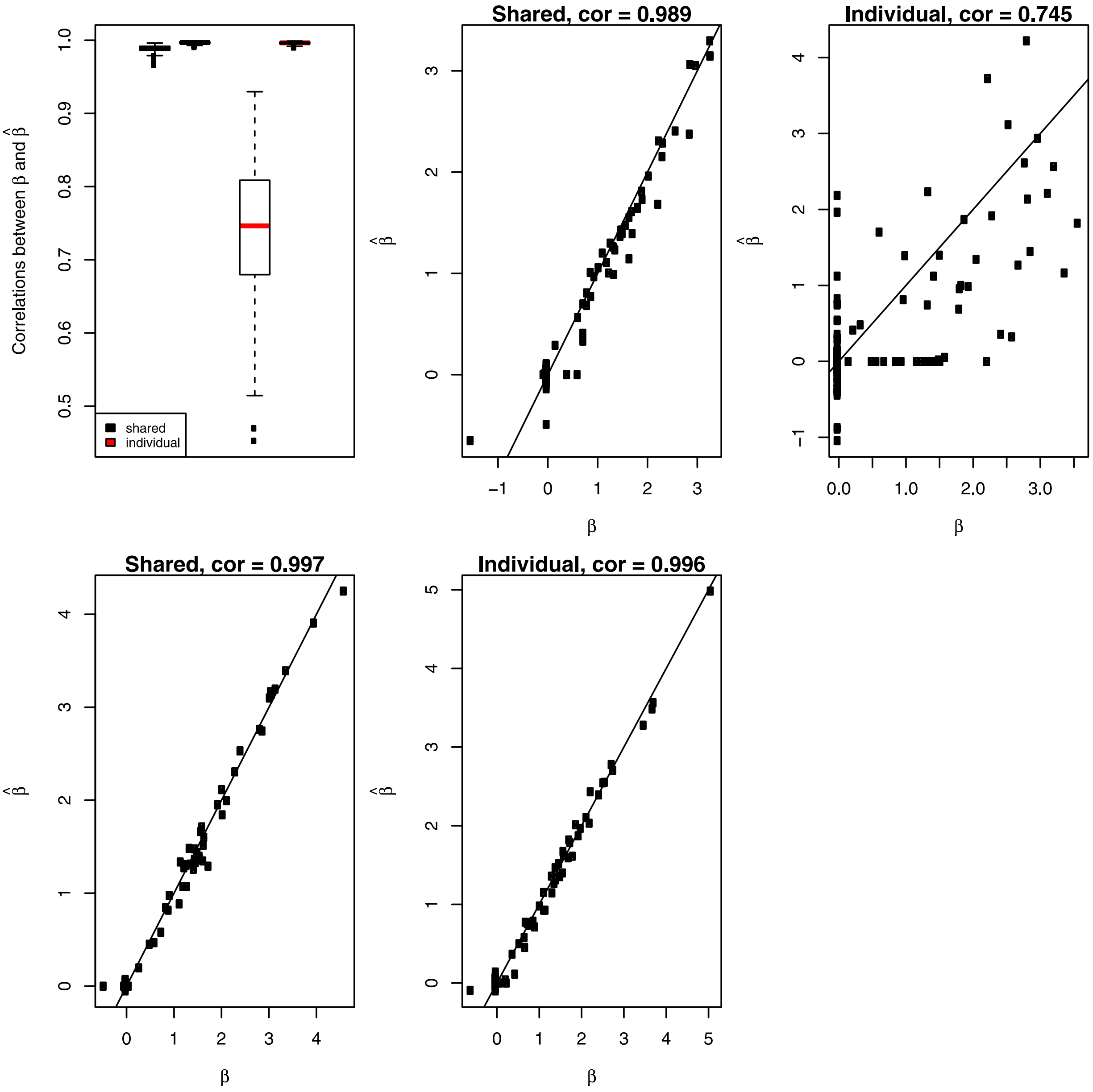

Figure 4. Top left panel: Simulation correlations between estimated $\boldsymbol{\beta}$ and the true β across the 200 runs of the simulation for the shared model (black) and extracted from fitted individual models (red). The side by side boxplots represents the two different sample sizes in the simulation. All other panels show the scatterplots of estimated values of $\boldsymbol{\beta}$ versus true values at the median correlations from the boxplots in the top left panel.

correlations between $\hat{\beta}_G$ and the true $\boldsymbol{\beta}$. Results were summarized in Fig. 4. Again, model (1) produced much higher correlations than individual G model did when sample size is small. As sample size largely increases, individual G model becomes the gold standard and model (1) is as good as this gold standard.

GDSC and CCLE Datasets. Figure 5 shows side-by-side boxplots (over the 200 random splits of the GDSC dataset) of test set error rates for the common 15 drugs when 70% of 80% GDSC dataset (scenario 1) was used in training data. Figure 6 show results when all 80% GDSC dataset (scenario 2) was used in training data. The changes, however, are negligible between the two scenarios. This is reasonable given the limited change in sample size. More specifically, given a median sample size of 88, the sample size was only increased by 21 in scenario 2. Highlighted in yellow, are the names of drugs where other groups had found reasonable concordance between the two drug response datasets[4] (Pearson correlation > 0.45).

We used following criteria to compare the side-by-side boxplots. If interquartile ranges of two boxplots do not overlap, there is a statistically meaningful difference in measures of central location between distributions. When two interquartile ranges overlap, we calculate the distance between the medians (DBM) as a percentage of overall visible spread (OVS), i.e. BDM/OVS, where OVS is the distance between the lower quarter of one box and the

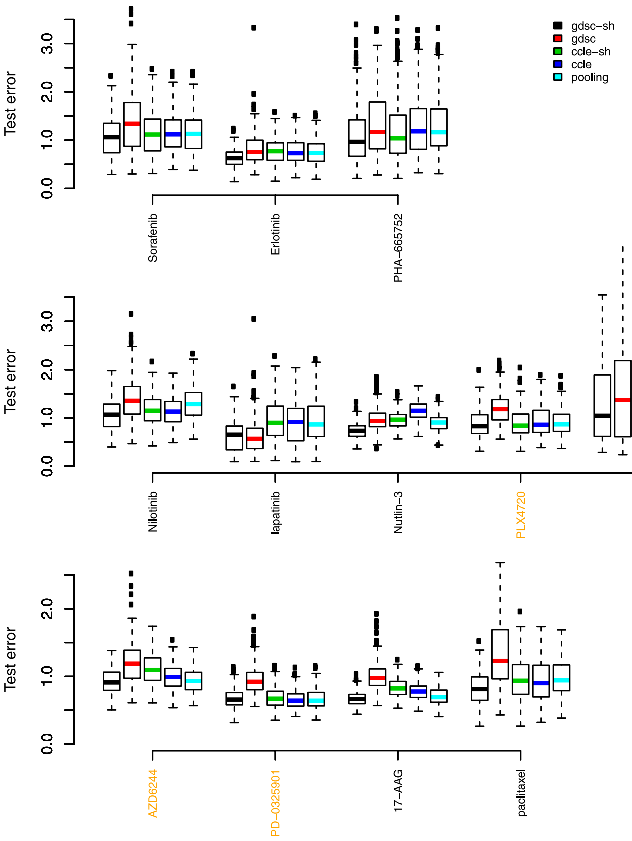

Figure 5. Test error rate boxplots by drug when 70% of 80% GDSC dataset were used in training data and 20% GDSC dataset were reserved as test data. Black represents predictions based on data sharing estimates for $\boldsymbol{\beta}$ alone; red for predictions from a model using GDSC alone; green for predictions from a data sharing model using estimates of $\boldsymbol{\beta}+\boldsymbol{\delta}$; blue for predictions from a model using CCLE alone, and light blue, predictions from a pooled GDSC + CCLE model.

higher quarter of another box. For a sample size of 100, if this percentage is over 20%, there is likely a statistically meaningful difference between the two boxplots. We opt for this conservative approach rather than reporting multiple p-values and then having to correct for multiple testing because it's a more intuitive way to relate inference to the boxplots themselves.

Some striking conclusions are evident. Among the compared methods, it is easy to see that the reduction in sample size by using the GDSC training data alone (the red boxplots) has a clear deleterious effect. Typical modeling would use the CCLE dataset alone (the dark blue boxplots). Table 1 shows the BDM/OVS values for boxplots with respect to the shared GDSC model and CCLE alone model for scenario 2. For drugs with reasonable concordance, the shared method does not provide much in way of test set error reductions. This is exactly what should happen. However, for 7 out of the other 10 drugs, the shared method clearly produces the most accurate models in terms of test set error rates. The 7 drugs are Erlotinib, PHA-665752, lapatinib, Nutlin-3, PD-0332991, TAE684 and 17-AAG. For the rest 3 drugs, the shared method is as good as or slightly better than the CCLE alone method.

Figures 7 and 8 show results where 20% CCLE dataset were reserved as test data, and Table 2 shows the BDM/OVS values for boxplots with respect to the shared CCLE model and GDSC alone model for scenario 2. Conclusions are similar to the analysis where 20% GDSC dataset were reserved as test data. Now, the shared method clearly produces the most accurate models in terms of test set error rates for 2 out of the 5 drugs with reasonable concordance and for 7 out of the other 10 discordant drugs. They include Sorafenib, Nilotinib, lapatinib, Nutlin-3, PD-0332991, AZD6244, PD-0325901, 17-AAG and paclitaxel.

Figure 9 shows a very interesting plot of signature validation for the 15 drugs in common. The y-axis shows results from the fit of model (1) on the GDSC data (lowercase g) and the CCLE data (lowercase c). The x-axis depicts a subset of genomic effects with at least one highly significant non-zero effect across the 15 drugs. The body of the plot shows the estimated biomarker effect sizes. For g-drugs, the $\hat{\boldsymbol{\beta}}$ values are plotted. For c-drugs, the $(\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\delta}})$ values are plotted. Blue colors indicate markers which predict drug sensitivity, red the opposite - the darker the color, the more intense the effect. Other genomic predictors not plotted showed no highly significant non-zero estimated effects for any of the 15 drugs. Shown in color are the known drug-marker associations[6]. Clearly we do an excellent job of recovering those. We find all of the ones previously found except for the Nutlin-3-MDM2 expression interaction. The association pattern of Crizotinib-HGF expression interaction was

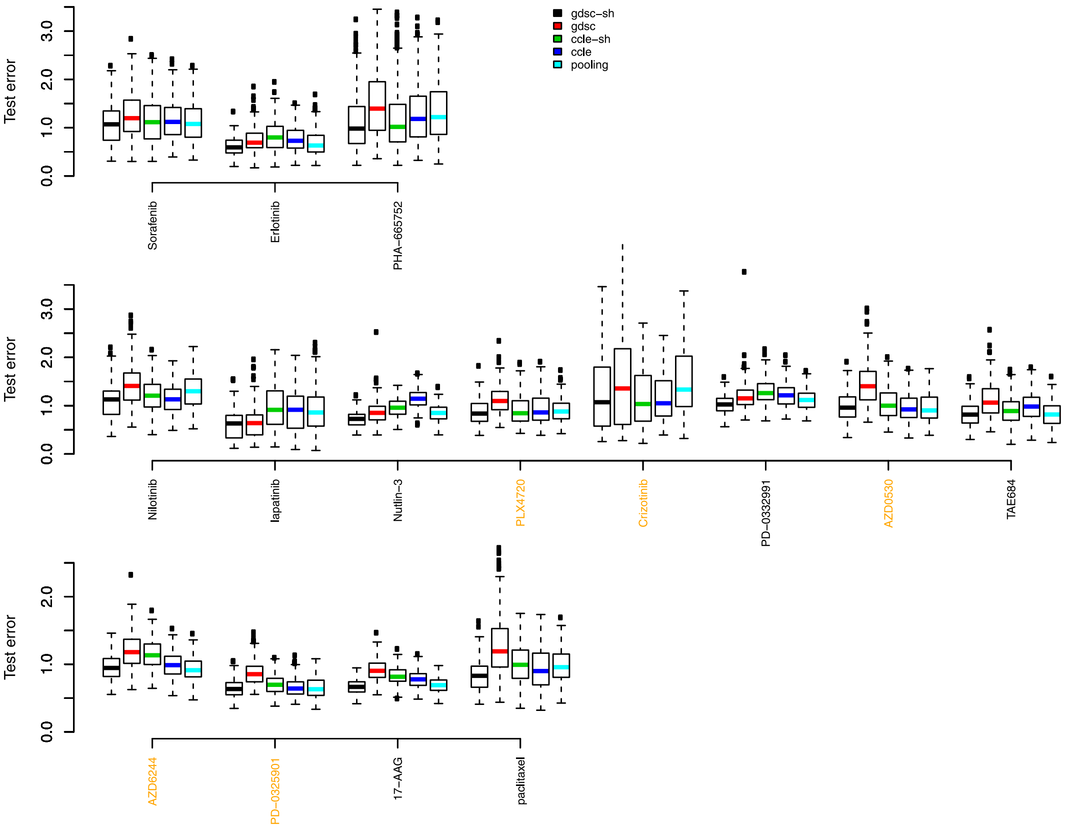

Figure 6. Test error rates by drug when 80% GDSC were used in training data and 20% GDSC dataset were reserved as test data. Black represents predictions based on data sharing estimates for β alone; red for predictions from a model using GDSC alone; green for predictions from a data sharing model using estimates of $\beta + \delta$; blue for predictions from a model using CCLE alone, and light blue, predictions from a pooled GDSC + CCLE model.

Drug	min	lower quarter	median	upper quarter	max	BDM/OVS
Sorafenib	$\frac{0.307}{0.394}$	$\frac{0.741}{0.857}$	$\frac{1.07}{1.12}$	$\frac{1.347}{1.419}$	$\frac{2.18}{2.2}$	0.075
Erlotinib	$\frac{0.196}{0.221}$	$\frac{0.477}{0.578}$	$\frac{0.593}{0.729}$	$\frac{0.742}{0.945}$	$\frac{1.042}{1.467}$	0.292
PHA-665752	$\frac{0.221}{0.324}$	$\frac{0.672}{0.81}$	$\frac{0.982}{1.182}$	$\frac{1.438}{1.652}$	$\frac{2.546}{2.881}$	0.204
Nilotinib	$\frac{0.362}{0.488}$	$\frac{0.819}{0.92}$	$\frac{1.129}{1.131}$	$\frac{1.302}{1.34}$	$\frac{2.027}{1.928}$	0.004
lapatinib	$\frac{0.117}{0.092}$	$\frac{0.332}{0.532}$	$\frac{0.632}{0.915}$	$\frac{0.799}{1.196}$	$\frac{1.298}{2.041}$	0.328
Nutlin-3	$\frac{0.392}{0.747}$	$\frac{0.602}{1.016}$	$\frac{0.726}{1.144}$	$\frac{0.82}{1.271}$	$\frac{1.12}{1.644}$	0.624
PLX4720	$\frac{0.384}{0.386}$	$\frac{0.676}{0.699}$	$\frac{0.836}{0.86}$	$\frac{1.045}{1.155}$	$\frac{1.488}{1.805}$	0.049
Crizotinib	$\frac{0.256}{0.392}$	$\frac{0.576}{0.785}$	$\frac{1.071}{1.05}$	$\frac{1.798}{1.516}$	$\frac{3.464}{2.453}$	−0.023
PD-0332991	$\frac{0.563}{0.722}$	$\frac{0.894}{1.036}$	$\frac{1.025}{1.214}$	$\frac{1.151}{1.372}$	$\frac{1.486}{1.815}$	0.395
AZD0530	$\frac{0.339}{0.33}$	$\frac{0.764}{0.755}$	$\frac{0.959}{0.922}$	$\frac{1.182}{1.155}$	$\frac{1.727}{1.734}$	−0.093
TAE684	$\frac{0.3}{0.285}$	$\frac{0.641}{0.778}$	$\frac{0.815}{0.981}$	$\frac{0.987}{1.174}$	$\frac{1.453}{1.744}$	0.312
AZD6244	$\frac{0.554}{0.536}$	$\frac{0.819}{0.857}$	$\frac{0.946}{0.987}$	$\frac{1.085}{1.119}$	$\frac{1.46}{1.436}$	0.138
PD-0325901	$\frac{0.347}{0.408}$	$\frac{0.549}{0.56}$	$\frac{0.634}{0.641}$	$\frac{0.73}{0.74}$	$\frac{0.981}{0.996}$	0.037
17-AAG	$\frac{0.417}{0.485}$	$\frac{0.59}{0.689}$	$\frac{0.666}{0.777}$	$\frac{0.741}{0.861}$	$\frac{0.947}{1.115}$	0.411
paclitaxel	$\frac{0.409}{0.321}$	$\frac{0.661}{0.696}$	$\frac{0.83}{0.9}$	$\frac{0.972}{1.165}$	$\frac{1.407}{1.736}$	0.14

Table 1. Statistical metrics to compare test error rates from shared GDSC model (numerator) and CCLE alone model (denominator) in scenario 2 when portion of GDSC dataset was withheld as test data.

recovered. It however was not identified as highly significant based on previously described inclusion rule and thus is not shown in Fig. 9. What is more interesting is how many new reproducible markers the data sharing strategy finds. For each drug, new bands of markers are discovered. What's also noticeable is that the bands do not entirely overlap between g and c. These areas without overlap are markers where the dataset discordancies were

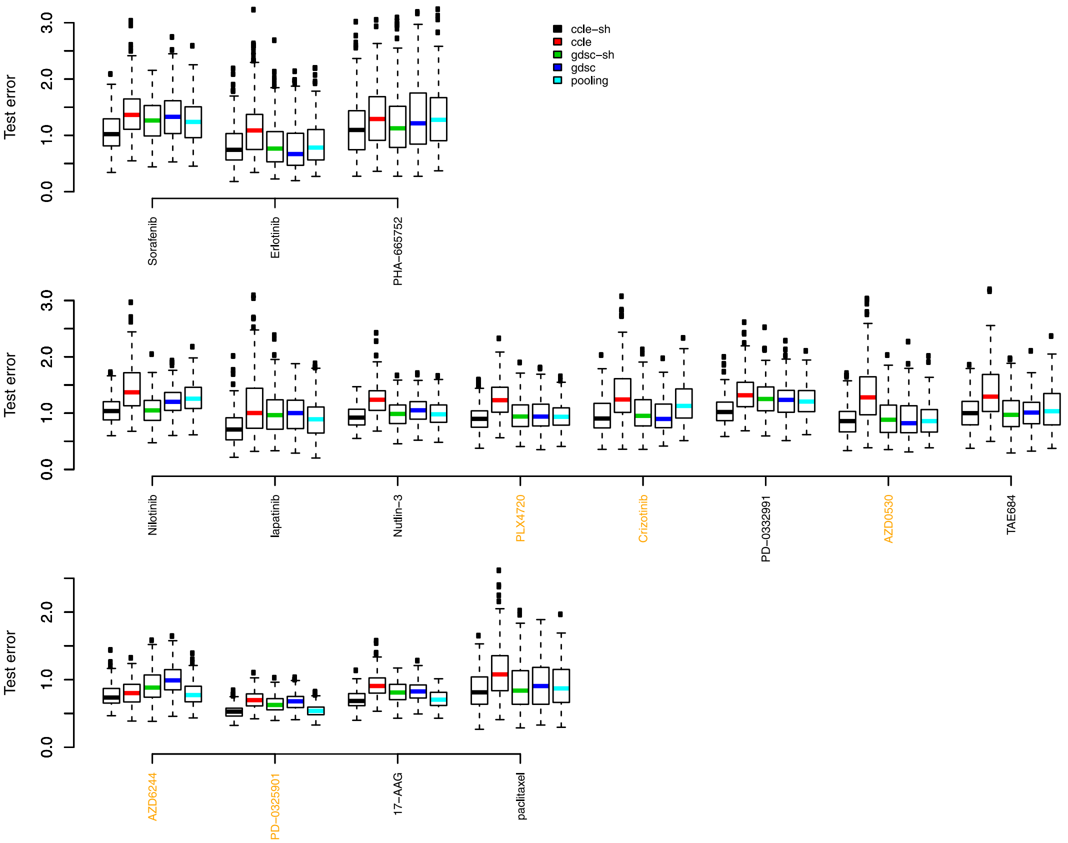

Figure 7. Test error rate boxplots by drug when 70% of 80% CCLE dataset were used in training data and 20% CCLE dataset were reserved as test data. Black represents predictions based on data sharing estimates for $\boldsymbol{\beta}$ alone; red for predictions from a model using GDSC alone; green for predictions from a data sharing model using estimates of $\boldsymbol{\beta}+\boldsymbol{\delta}$; blue for predictions from a model using CCLE alone, and light blue, predictions from a pooled GDSC + CCLE model.

Drug	min	lower quarter	median	upper quarter	max	BDM/OVS
Sorafenib	$\frac{0.387}{0.528}$	$\frac{0.8234}{1.032}$	$\frac{1.025}{1.329}$	$\frac{1.372}{1.615}$	$\frac{2.074}{2.449}$	0.384
Erlotinib	$\frac{0.150}{0.196}$	$\frac{0.569}{0.468}$	$\frac{0.736}{0.668}$	$\frac{0.970}{1.037}$	$\frac{1.493}{1.882}$	−0.145
PHA-665752	$\frac{0.273}{0.273}$	$\frac{0.770}{0.857}$	$\frac{1.141}{1.214}$	$\frac{1.486}{1.745}$	$\frac{2.498}{2.967}$	0.075
Nilotinib	$\frac{0.548}{0.603}$	$\frac{0.899}{1.048}$	$\frac{1.104}{1.200}$	$\frac{1.267}{1.367}$	$\frac{1.794}{1.761}$	0.205
lapatinib	$\frac{0.211}{0.290}$	$\frac{0.514}{0.726}$	$\frac{0.690}{1.001}$	$\frac{0.923}{1.230}$	$\frac{1.457}{1.878}$	0.435
Nutlin-3	$\frac{0.508}{0.520}$	$\frac{0.786}{0.892}$	$\frac{0.936}{1.047}$	$\frac{1.097}{1.204}$	$\frac{1.554}{1.582}$	0.266
PLX4720	$\frac{0.444}{0.352}$	$\frac{0.807}{0.772}$	$\frac{0.925}{0.938}$	$\frac{1.055}{1.159}$	$\frac{1.418}{1.691}$	0.038
Crizotinib	$\frac{0.389}{0.415}$	$\frac{0.716}{0.749}$	$\frac{0.910}{0.897}$	$\frac{1.194}{1.154}$	$\frac{1.907}{1.726}$	−0.029
PD-0332991	$\frac{0.511}{0.512}$	$\frac{0.849}{1.016}$	$\frac{0.974}{1.235}$	$\frac{1.112}{1.405}$	$\frac{1.448}{1.960}$	0.47
AZD0530	$\frac{0.332}{0.311}$	$\frac{0.687}{0.653}$	$\frac{0.887}{0.821}$	$\frac{1.102}{1.135}$	$\frac{1.689}{1.795}$	−0.146
TAE684	$\frac{0.333}{0.326}$	$\frac{0.788}{0.812}$	$\frac{0.955}{1.01}$	$\frac{1.135}{1.192}$	$\frac{1.569}{1.719}$	0.136
AZD6244	$\frac{0.435}{0.457}$	$\frac{0.638}{0.853}$	$\frac{0.706}{0.997}$	$\frac{0.818}{1.149}$	$\frac{1.062}{1.587}$	0.568
PD-0325901	$\frac{0.3}{0.408}$	$\frac{0.446}{0.587}$	$\frac{0.496}{0.68}$	$\frac{0.566}{0.752}$	$\frac{0.702}{0.984}$	0.601
17-AAG	$\frac{0.425}{0.492}$	$\frac{0.604}{0.728}$	$\frac{0.686}{0.826}$	$\frac{0.76}{0.922}$	$\frac{0.967}{1.209}$	0.44
paclitaxel	$\frac{0.217}{0.329}$	$\frac{0.586}{0.636}$	$\frac{0.761}{0.906}$	$\frac{1.025}{1.182}$	$\frac{1.487}{1.889}$	0.242

Table 2. Statistical metrics to compare test error rates from shared CCLE model (numerator) and GDSC alone model (denominator) in scenario 2 when portion of CCLE dataset was withheld as test data.

large enough to wash away true effects ($|\delta_j| \gg |\beta_j|$) such that they were not detected as reproducible. Closer examination of the particular drugs reveals that the washing out effect is happening in those drugs where discordancies were previously established and much less so where concordance was found. *This is exactly what the theory would have predicted.*

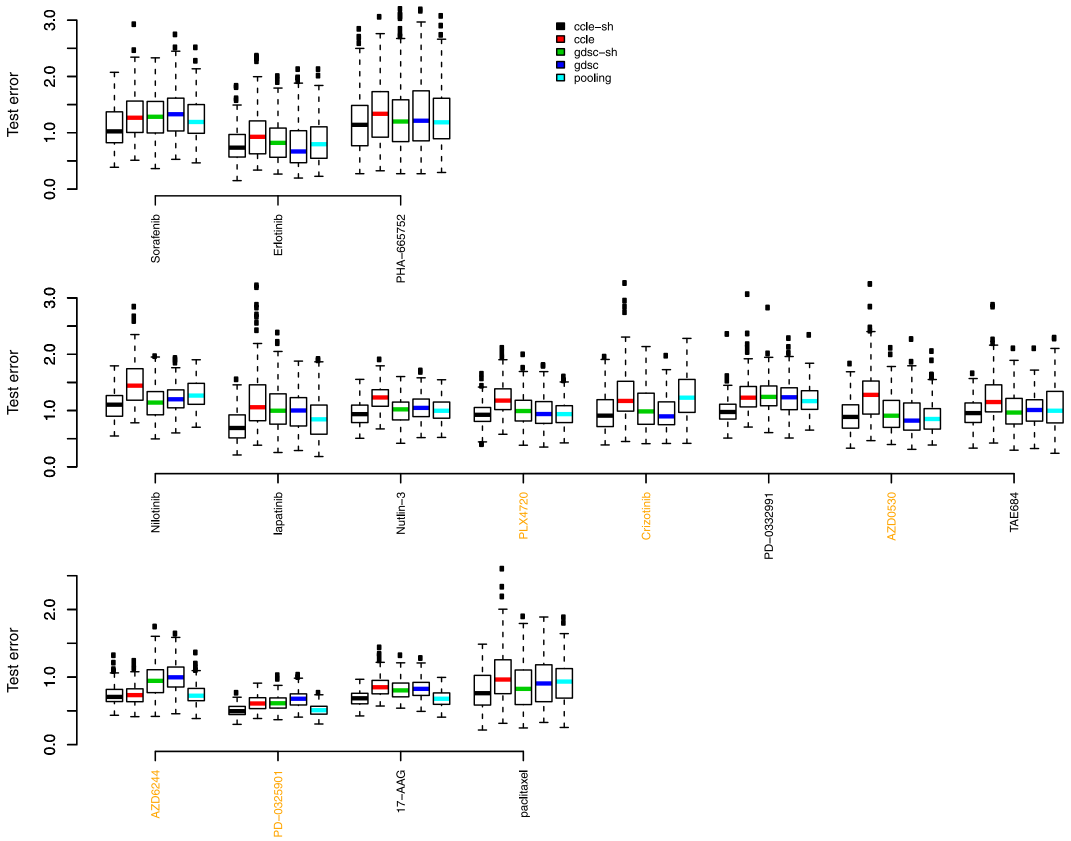

Figure 8. Test error rates by drug when 80% CCLE dataset were used in training data and 20% CCLE dataset were reserved as test data. Black represents predictions based on data sharing estimates for β alone; red for predictions from a model using GDSC alone; green for predictions from a data sharing model using estimates of $\beta + \delta$; blue for predictions from a model using CCLE alone, and light blue, predictions from a pooled GDSC + CCLE model.

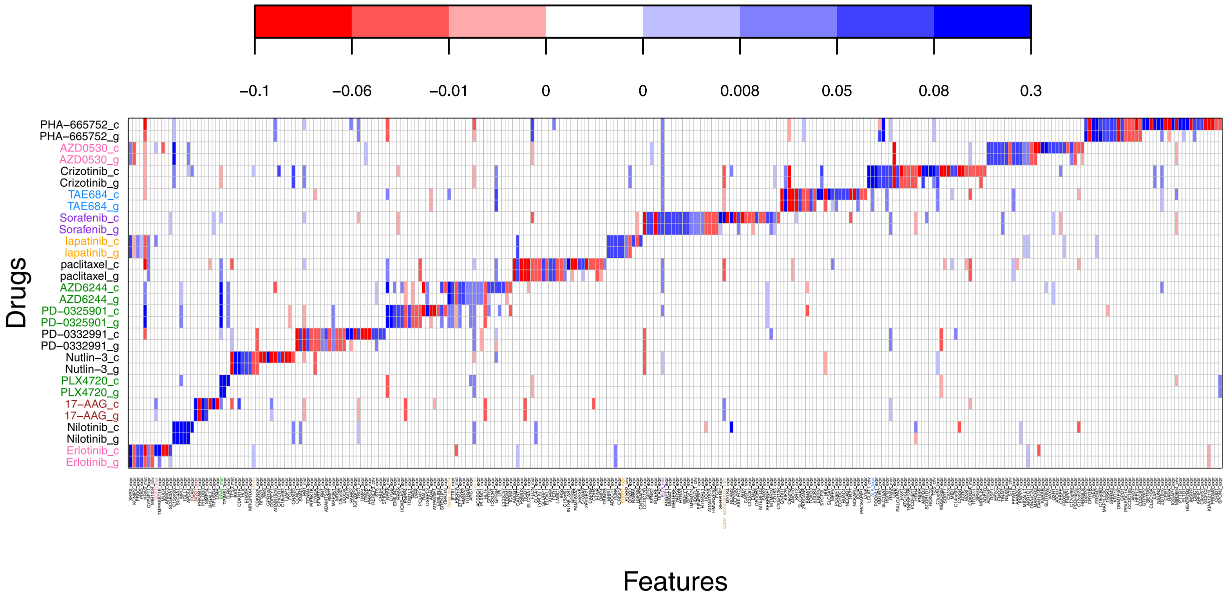

Figure 9. Signature validation plot based on GDSC (g) and CCLE (c) datasets using Formulation 1. The y-axis shows the different drugs under study. The x-axis shows the markers which had significant effects. Blue coloring indicates a positive estimated effect; red a negative estimated effect. White indicates an estimated null effect.

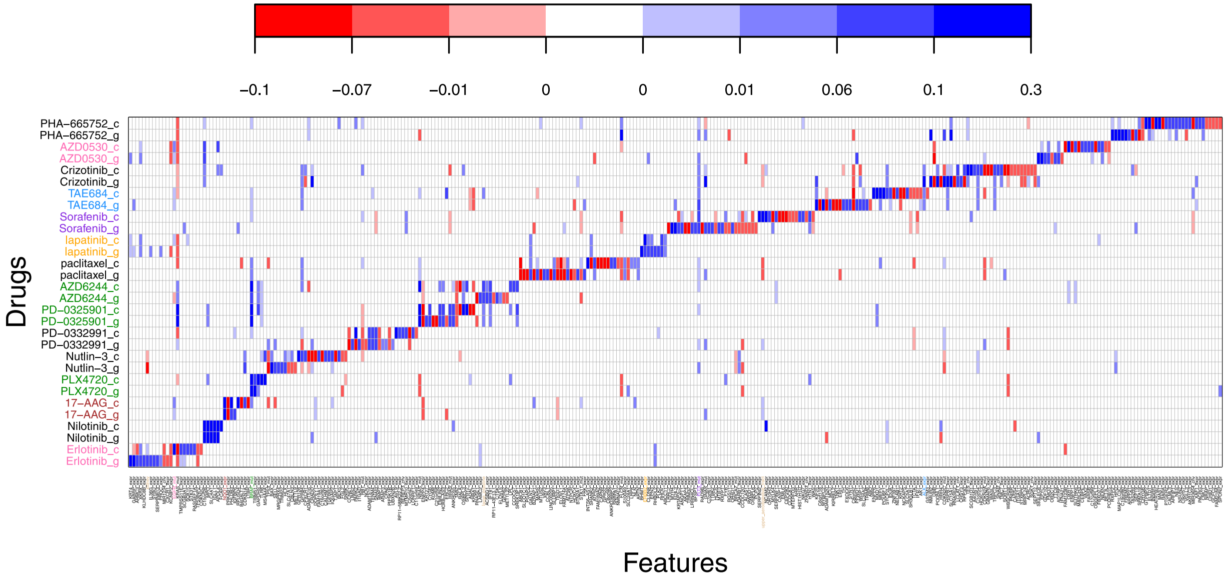

Figure 10. Signature validation plot based on GDSC (g) and CCLE (c) datasets using Formulation 2. The y-axis shows the different drugs under study. The x-axis shows the markers which had significant effects. Blue coloring indicates a positive estimated effect; red a negative estimated effect. White indicates an estimated null effect.

One can dig further into the signature validation plot. For instance, population stratifications can be made to better identify only those patients who may provide a higher likelihood of a favorable drug responses. Take nilotnib as an example. Here, having increased gene expression in C15orf26, SLC4A1, MPO, IGLL1, and APOL4 (i.e. the reproducible markers shaded dark blue) would be predictive of a favorable drug response. In addition, this was accentuated for thyroid cell lines as compared to the other cancer cell lines. On the other hand, for PLX4720, having the BRAF mutation and TRIM51 expression present, would predict a favorable response *as long as no NRAS mutation was present*. Other interesting patterns like this can be gleaned for each drug.

The horizontal banding is clearly what is most visible initially. However, one can also look vertically at specific genomic markers and find interesting information. For example, with the KRAS mutation, the only drug which shows strong reproducible predicted sensitivity to its presence is PD-0325901 - a MEK inhibitor. Many other drugs are predicted to encounter a resistance effect as indicated by the red c-g pairs along the KRAS column. In fact, recent evidence suggests that PD-0325901 used in conjunction with dacomitinib may be of use when KRAS mutations are present[18].

Figure 10 shows a signature validation plot based on model Formulation 2. Focusing on a specific discoveries illuminates how to reconcile Figs 9 and 10. Take the TP53-mut biomarker and the drug Nutlin-3 for instance. Both Formulation 1 and Formulation 2 fits (Figs 9 and 10 respectively), show bright red colorations for both datasets CCLE and GDSC. This would indicate a gene-drug interaction that validated with both model Formulations. Take the neighboring biomarker BAX-expr for the same drug: in Formulation 1, the interaction appears solid blue for both datasets. For Formulation 2, it's slightly darker blue for GDSC than CCLE. This simply implies that by Formulation 2, the fitted effects across the two datasets are very similar with GDSC having a slightly stronger difference from baseline than CCLE. Translating back to Formulation 1, this means that the concordancy effect estimate dominates the discordancy estimate which is exactly what we see. Other similar inferences can be made looking at different gene-drug interactions between the two Formulation plots. However, contrast this to the BRCA2 mutation. It does not show up at all on Fig. 9. Thus one may conclude that no detectable signal was associated with this marker for any drug. This can be confirmed by examining Fig. 10 which shows the signature validation plot using Formulation 2. Here BRCA2 does show and does indicate a smaller reproducible effect for Nutlin-3. However the coloring is identical indicating that the estimated values of $\boldsymbol{\delta}_C$ and $\boldsymbol{\delta}_g$ are either estimated very similarly in magnitude (or both estimated near zero). The difference between the two model Formulations is attributable to differences in model tuning and to the stringency of the inclusion rules that are used.

Discussion

We have demonstrated that reproducible signal can in fact be partitioned from signal due to dataset discordancy using data sharing strategies. The most exciting thing may in fact be that we found new reproducible biomarkers for *every* drug in the analysis, whereas others have only been able to do so for about half of the drugs[11]. Furthermore, patient populations can be partitioned based on their reproducible biomarker profiles to more precisely predict a favorable response to a drug. Additionally, we have shown that test set prediction error rates are markedly lower when using discordancy partitioning models and very much follow established theory. Thus while experimental challenges may still exist in order to better standardize protocols, it's likely datasets will never be completely concordant. Discordancy partitioning approaches like what we have presented can adapt easily to varying degrees of discordancy to produce more accurate assessments of validation.

As for the methodology itself, a few more remarks are in order. With regards to estimation, it should be noted that we are not saying that the combined elastic net/lasso estimation technique is the only one that could be used here. In fact we developed our own new modeling strategy based on a generalized finite mixture of regressions model which can be used to find interesting new patterns of therapeutic biomarkers[19]. However, even this new modeling strategy can be embedded within a data sharing strategy for validation purposes. The reason we chose to illustrate results with the elastic net/lasso approach is because it's a more widely known approach for estimating sparse genomic models.

References

1. Yang, W. *et al*. Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic acids research* **41**(D1), D955–D961 (2013).
2. Garnett, M. J. *et al*. Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature* **483**(7391), 570–575 (2012).
3. Barretina, J. *et al*. The Cancer Cell Line Encyclopedia enables predictive modeling of anticancer drug sensitivity. *Nature* **483**(7391), 603–607 (2012).
4. Cancer Cell Line Encyclopedia Consortium and Genomics of Drug Sensitivity in Cancer Consortium. Pharmacogenomic agreement between two cancer cell line data sets. *Nature* **528**(**7580**), 84–87 (2015).
5. Haibe-Kains, B. *et al*. Inconsistency in large pharmacogenomic studies. *Nature* **504**(7480), 389–393 (2013).
6. Papillon-Cavanagh, S. *et al*. Comparison and validation of genomic predictors for anticancer drug sensitivity. *Journal of the American Medical Informatics Association* **20**(4), 597–602 (2013).
7. Ishwaran, H. & Rao, J. S. Geometry and properties of generalized ridge regression in high dimensions. *Contemp. Math* **622**, 81–93 (2014).
8. Greshock, J. *et al*. Molecular target class is predictive of *in vitro* response profile. *Cancer Res.* **70**(9), 3677–3686 (2010).
9. Mpindi, J. P. *et al*. Consistency in drug response profiling. *Nature* **540**(7631), E5–E6 (2016).
10. Haverty, PeterM. *et al*. Reproducible pharmacogenomic profiling of cancer cell line panels. *Nature* **533**(7603), 333–337 (2016).
11. Safikhani, Z. *et al*. Assessment of pharmacogenomic agreement. *F1000 Research* **5**, 825 (2016).
12. Gross, S. M. & Tibshirani, R. Data Shared Lasso: A novel tool to discover uplift. *Computational Statistics and Data Analysis* **101**, 226–235 (2016).
13. Chen, A., Owen, A. B. & Shi, M. Data enriched linear regression. *Electronic Journal of Statistics* **9**(1), 1078–1112 (2015).
14. Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*. *Series B* (*Methodological*) 267–288 (1996).
15. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**(2), 301–320 (2005).
16. Zou, H. & Zhang, H. H. On the adaptive elastic-net with a diverging number of parameters. *Annals of statistics* **37**(4), 1733 (2009).
17. Fan, J. & Lv, J. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70**(5), 849–911 (2008).
18. Hamidi, H. *et al*. KRAS mutational subtype and copy number predict *in vitro* response of human pancreatic cancer cell lines to MEK inhibition. *British Journal of Cancer* **111**, 1788–1801 (2014).
19. Liu, H. & Rao, J. S. Precision Therapeutic Biomarker Identification with Application to the Cancer Genome Project. *arXiv preprint* arXiv:1702.02264 (2017).

Acknowledgements

J.S.R. was supported in part by grant R01 CA160593A1 and NSF grant DMS 1513266. H.L. was supported in part by a graduate student fellowship from the Sylvester Comprehensive Cancer Center at the University of Miami.

Author Contributions

J.S.R. developed the methodology and wrote the main manuscript text. H.L. analyzed the data, performed the simulations, prepared all figures and assisted in writing the manuscript text.

Additional Information

Competing Interests: The authors declare that they have no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.