Check for updates

RESEARCH ARTICLE

REVISED Revisiting inconsistency in large pharmacogenomic studies [version 3; referees: 2 approved, 1 approved with reservations]

Zhaleh Safikhani[1,2], Petr Smirnov[2], Mark Freeman[2], Nehme El-Hachem[3], Adrian She[2], Quevedo Rene[1,2], Anna Goldenberg[4,5], Nicolai J. Birkbak[6], Christos Hatzis [7,8], Leming Shi[9,10], Andrew H. Beck[11], Hugo J.W.L. Aerts[12,13], John Quackenbush[12,14], Benjamin Haibe-Kains[1,2,4,15]

[1]Department of Medical Biophysics, University of Toronto, Toronto, M5G 1L7, Canada
[2]Princess Margaret Cancer Centre, University Health Network, Toronto, M5G 1L7, Canada
[3]Institut de Recherches Cliniques de Montréal, Montréal, H2W 1R7, Canada
[4]Department of Computer Science, University of Toronto, Toronto, M5S 2E4, Canada
[5]Hospital for Sick Children, Toronto, M5G 1X8, Canada
[6]University College London, London, WC1E 6BT, UK
[7]Yale Cancer Center, Yale University, New Haven, CT, 06510, USA
[8]Section of Medical Oncology, Yale University School of Medicine, New Haven, CT, 06520, USA
[9]University of Arkansas for Medical Sciences, Little Rock, AR, 72205, USA
[10]Fudan University, Shanghai City, 200135, China
[11]Department of Pathology, Beth Israel Deaconess Medical Center and Harvard Medical School, Boston, MA, 02215, USA
[12]Department of Biostatistics and Computational Biology and Center for Cancer Computational Biology, Boston, MA, 02215, USA
[13]Department of Radiation Oncology and Radiology, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA, 02215, USA
[14]Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, MA, 02215, USA
[15]Ontario Institute of Cancer Research, Toronto, M5G 1L7, Canada

v3 **First published:** 16 Sep 2016, **5**:2333 (doi: 10.12688/f1000research.9611.1)
Second version: 25 Jul 2017, **5**:2333 (doi: 10.12688/f1000research.9611.2)
Latest published: 11 Aug 2017, **5**:2333 (doi: 10.12688/f1000research.9611.3)

Abstract

In 2013, we published a comparative analysis of mutation and gene expression profiles and drug sensitivity measurements for 15 drugs characterized in the 471 cancer cell lines screened in the Genomics of Drug Sensitivity in Cancer (GDSC) and Cancer Cell Line Encyclopedia (CCLE). While we found good concordance in gene expression profiles, there was substantial inconsistency in the drug responses reported by the GDSC and CCLE projects. We received extensive feedback on the comparisons that we performed. This feedback, along with the release of new data, prompted us to revisit our initial analysis. We present a new analysis using these expanded data, where we address the most significant suggestions for improvements on our published analysis — that targeted therapies and broad cytotoxic drugs should have been treated differently in assessing consistency, that consistency of both molecular profiles and drug sensitivity measurements should be compared across cell lines, and that the software analysis tools provided should have been easier to run, particularly as the GDSC and CCLE released additional data.

Open Peer Review

Referee Status: ✓ ? ✓

Invited Referees

Our re-analysis supports our previous finding that gene expression data are significantly more consistent than drug sensitivity measurements. Using new statistics to assess data consistency allowed identification of two broad effect drugs and three targeted drugs with moderate to good consistency in drug sensitivity data between GDSC and CCLE. For three other targeted drugs, there were not enough sensitive cell lines to assess the consistency of the pharmacological profiles. We found evidence of inconsistencies in pharmacological phenotypes for the remaining eight drugs.

Overall, our findings suggest that the drug sensitivity data in GDSC and CCLE continue to present challenges for robust biomarker discovery. This re-analysis provides additional support for the argument that experimental standardization and validation of pharmacogenomic response will be necessary to advance the broad use of large pharmacogenomic screens.

This article is included in the Preclinical Reproducibility and Robustness gateway.

	1	2	3
REVISED **version 3** published 11 Aug 2017	✓ report		
REVISED **version 2** published 25 Jul 2017		? report	
version 1 published 16 Sep 2016	✗ report	? report	✓ report

1 **Michael T. Hallett**, Concordia University, Canada

2 **Paul T. Spellman**, Oregon Health and Science University, USA

3 **David G. Covell**, National Cancer Institute, USA

Discuss this article

Comments (0)

Corresponding author: Benjamin Haibe-Kains (bhaibeka@uhnresearch.ca)

Author roles: Safikhani Z: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Project Administration, Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Smirnov P**: Data Curation, Formal Analysis, Methodology, Software, Validation; **Freeman M**: Formal Analysis, Methodology, Software, Writing – Review & Editing; **El-Hachem N**: Conceptualization, Data Curation; **She A**: Data Curation, Software; **Rene Q**: Data Curation, Formal Analysis, Methodology, Software, Validation, Writing – Review & Editing; **Goldenberg A**: Conceptualization, Methodology, Validation, Writing – Review & Editing; **Birkbak NJ**: Conceptualization, Investigation, Writing – Original Draft Preparation; **Hatzis C**: Conceptualization, Investigation, Writing – Original Draft Preparation; **Shi L**: Investigation, Validation, Writing – Original Draft Preparation; **Beck AH**: Conceptualization, Investigation, Writing – Original Draft Preparation; **Aerts HJWL**: Conceptualization, Investigation, Writing – Original Draft Preparation; **Quackenbush J**: Conceptualization, Investigation, Methodology, Writing – Original Draft Preparation; **Haibe-Kains B**: Conceptualization, Data Curation, Formal Analysis, Funding Acquisition, Investigation, Methodology, Project Administration, Resources, Software, Supervision, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

How to cite this article: Safikhani Z, Smirnov P, Freeman M *et al.* **Revisiting inconsistency in large pharmacogenomic studies [version 3; referees: 2 approved, 1 approved with reservations]** *F1000Research* 2017, **5**:2333 (doi: 10.12688/f1000research.9611.3)

Grant information: Z Safikhani was supported by the Cancer Research Society (Canada; grant #19271) and the Ontario Institute for Cancer Research through funding provided by the Government of Ontario. P Smirnov was supported by the Canadian Cancer Society Research Institute. C Hatzis was supported by Yale University. N Juul Birkbak was funded by The Villum Kann Rasmussen Foundation. L Shi was supported by the National High Technology Research and Development Program of China (2015AA020104), the National Natural Science Foundation of China (31471239), the 111 Project (B13016), and the National Supercomputer Center in Guangzhou, China. J Quackenbush was supported by grants from the NCI GAME-ON Cancer Post-GWAS initiative (5U19 CA148065) and the NHLBI (5R01HL111759). B Haibe-Kains was supported by the Canadian Institutes of Health Research, Cancer Research Society, Terry Fox Research Institute, and the Gattuso Slaight Personalized Cancer Medicine Fund at Princess Margaret Cancer Centre.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

First published: 16 Sep 2016, **5**:2333 (doi: 10.12688/f1000research.9611.1)

REVISED **Amendments from Version 2**

The paragraph reporting the limitations of the study has been updated to reflect the fact that, although it's challenging, determining drug-specific thresholds to discretize sensitivity data is possible and may affect consistency estimation.

The Discussion and Conclusion sections have been updated to reflect the fact that the noise within assays must be assessed and accounted for, and importantly, the complementarity across assays offer new opportunities to develop more robust biomarkers.

See referee reports

Box 1. Summary box

In 2013 we reported inconsistency in the drug sensitivity phenotypes measured by the Genomics of Drug Sensitivity in Cancer (GDSC) and the Cancer Cell Lines Encyclopedia (CCLE) studies. Here we revisit that analysis and address a number of potential concerns raised about our initial methodology:

- ***Different drugs should be compared based on the observed pattern of response.*** To address this concern, we considered drugs falling into three classes: (1) drugs with no observed activity in any of the cell lines; (2) drugs with sensitivity observed for only a small subset of cell lines; and (3) drugs producing a response in a large number of cell lines. For each class, we assessed the correlation in drug response between studies using a variety of metrics, selecting the metric that performed best in each individual comparison. While no metric identified any substantial consistency for the first class (sorafenib, erlotinib, and PHA−665752) due to no activity, judicious choice of metric found high consistency for three of eight highly targeted therapies in the second class (nilotinib, crizotinib, and PLX4720), but no metric found better than moderate correlation for two of four broad effect drugs in the third class (PD−0332901 and 17-AAG).
- ***Measure of consistency for targeted drugs.*** Beyond considering drug response profiles, targeted drugs should be treated differently when assessing consistency. We used six different statistics to test consistency, using both continuous and discretized drug sensitivity data. We confirmed that Spearman rank correlation, used in our 2013 study, does not detect consistency for the three targeted therapies profiled by GDSC and CCLE. Other statistics, such as Somers' Dxy or Matthews correlation coefficient, yielded moderate to high consistency for specific drugs, but there was no single metric that found good consistency for each of the targeted drugs.
- ***Consistency of molecular profiles across cell lines.*** In our initial published analysis, we reported correlations based on comparing drug response "across cell lines" while gene expression levels were compared "between cell lines." It has been suggested it would be more appropriate to compute correlations "across cell lines" for both molecular and pharmacological data. Here we report a number of statistical measures of consistency for both gene expression and drug response compared across cell lines and confirm our initial finding that gene expression is significantly more consistent than the reported drug phenotypes.
- ***Some published biomarkers are reproducible between studies.*** In our initial comparative study we found that the majority of known biomarkers predictive of drugs response are reproducible across studies. We extended the list of known biomarkers and found that seven out of 11 are significant in GDSC and CCLE. While one can find such anecdotal examples, they do not lead to a general process for discovering a new biomarker in one study that can be applied to another study.
- ***Research reproducibility.*** The code we provided with our original paper was incompatible with updated releases of the GDSC and CCLE datasets. We developed *PharmacoGx*, which is a flexible, open-source software package based on the statistical language R, and used it to derive the results reported here.

Introduction

The goal of precision medicine is the identification of the best therapy for each patient and their own unique manifestation of a disease. This is particularly important in oncology where multiple cytotoxic and targeted drugs are available, but their therapeutic benefits are often insufficient or limited to a subset of cancer patients. Large-scale pharmacogenomics studies in which experimental and approved drugs are screened against panels of molecularly characterized cancer cell lines, have been proposed as a means for identifying drugs effective against specific cancers and for developing genomic biomarkers predictive of drug response. The Genomics of Drug Sensitivity in Cancer project (GDSC, referred to as the Cancer Genome Project [CGP] in our initial study)[1], and the Cancer Cell Line Encyclopedia (CCLE)[2] have each reported results of such screens, providing data on drug sensitivities and molecular profiles for collections of representative cancer cell lines.

Presented with these two large studies, our hope was that we could use the data to identify new molecular biomarkers of drug response in one study that would predict response in the second. We[3] and others[4–6] reported difficulties in building and validating biomarkers of response using the GDSC and CCLE datasets, even when the analysis was limited to the drugs and cell lines screened in both studies. To understand the cause of this failure, we compared the gene expression profiles and the drug response data reported by the GDSC and CCLE[7,8]. We found that, although the gene expression data showed reasonable consistency between the two studies, the drug sensitivity measurements were surprisingly inconsistent. This inconsistency can be clearly seen by plotting drug response reported for each of the 15 drugs provided in both GDSC and CCLE for the 471 cell lines assayed by both studies[7–10]. Since the publication of our comparative analysis, we received a great deal of constructive feedback from the scientific community regarding multiple aspects of the analysis we reported, including suggestions for analytical methods that might uncover greater consistency between the studies. Moreover, both GDSC and CCLE have released new drug sensitivity and molecular profiling data, allowing us not only to revisit our initial analysis, but also to extend it using these new data.

To begin, we investigated alternative statistics to assess the inter-study consistency for drugs exhibiting different patterns of response across the collection of cell lines common to both studies. We then considered statistical methods for targeted drugs expected to be sensitive only in a subset of cell lines. We compared consistency estimates between continuous and discretized molecular features (gene expression, copy number variations and mutations) and drug sensitivity data, and importantly, assessed how potential discordance may affect the discovery of molecular features (biomarkers) predictive of drug response. We also revisited our analysis of consistency of molecular data between studies and evaluated "known biomarkers" of response expected to be predictive in these studies.

This extensive reanalysis found that by selecting specific statistical measures on a case-by-case basis, one can identify moderate to good consistency for two broad effect and three targeted therapies. However, overall, our results support our initial observations that drug sensitivity data in GDSC and CCLE are inconsistent for the majority of the drugs, even when considering metrics yielding the highest consistency for individual drugs. Our present analysis adds further evidence supporting the need for robust and standardized experimental pipelines to assure generation of comparable, biologically relevant measures of drug response as well as unbiased statistical and machine learning methods to better predict response. Failure to do so will continue to limit the potential for use of large-scale pharmacogenomic screens in reliable drug development and precision medicine applications.

Results

The overall analysis design of our study is represented in Figure 1.

Intersection between GDSC and CCLE

To identify the largest set of cell lines and drugs profiled by both GDSC and CCLE, we used the *PharmacoGx* computational platform[11] that is able to store, analyze, and compare curated pharmacogenomic datasets. We created curated datasets for the new releases of the GDSC (July 2015) and CCLE (February 2015) projects. The improved curation of new data using *PharmacoGx*[11] identified 15 drugs in common between GDSC and CCLE as well 698 cell lines, originating from 23 tissue types (Figure 2). This is the same number of shared drugs but the updated datasets contains a larger number of common cell lines than the 471 reported in our previous analysis[7].

Comparing single nucleotide polymorphism (SNP) fingerprints

To check the accuracy of cell line name matching, we compared single nucleotide polymorphism (SNP) fingerprints using data released in both studies. We first controlled for the quality of the SNP arrays and excluded 11 of 1,396 profiles due to low quality (see Methods). We then compared SNP fingerprints of cell lines with identical name using > 80% as threshold for concordance[12,13]. Consistent with the results reported by the CCLE[2], the vast majority of cell lines had highly concordant fingerprints (462 out of 470 cell lines with SNP profiles available in both GDSC and CCLE; Dataset 1). We found eight cell lines with same identifier but different SNP identity (Figure 3); these were removed from our subsequent analyses to avoid discrepancies due to the use of possibly mislabeled or contaminated cell lines.

Estimation and filtering of drug dose-response curves

We used the viability measures for each drug concentration in GDSC and CCLE to fit dose-response curves and assess their quality. An important factor influencing the fitting of drug dose-response curves is the range of concentration used for each cell line/drug combination. In CCLE, all dose-response curves were measured at eight concentrations: 2.5×10^{-3}, 8×10^{-3}, 2.5×10^{-2}, 8×10^{-2}, 2.5×10^{-1}, 8×10^{-1}, 2.5, and 8 μM. However, in GDSC response was measured at a different set of concentrations for each drug. The minimum concentrations for different drugs range from 3.125×10^{-5} to 15.625 μM. In each case, the concentrations tested by GDSC form a geometric sequence of nine terms with a common ratio of two between successive concentrations. Thus, the maximum concentration tested for each drug is 256 times the minimum concentration for that drug and ranges from 8×10^{-3} to 4000 μM.

To properly fit drug dose-response curves, one must make multiple assumptions regarding the cell viability measurements generated by the pharmacological platform used in a given study. For instance, one assumes that viability ranges between 0% and 100% after data normalization and that consecutive viability measurements remain stable or decrease monotonically reflecting response to the drug being tested. Quality controls were implemented to flag dose-response curves that strongly violate these assumptions (Supplementary Methods). We identified 2315 (2.9%) and 123 (1%) dose-response curves that failed to pass in GDSC and CCLE, respectively, as depicted in Figure 4 (all noisy curves are provided in Supplementary File 1. We excluded these cases to avoid erroneous curve fitting.

We used least squares optimization to fit a three-parameter sigmoid model (Methods) for the drug dose-response curves in GDSC and CCLE (Supplementary File 2). For each fitted curve, we computed the most widely used drug activity metrics, that are the area under the curve (AUC) and the drug concentration required to inhibit 50% of cell viability (IC_{50}).

Consistency of drug sensitivity data

We began by computing the area between the two drug dose-response curves (ABC) to assess consistency of cell viability data for each cell line combination screened in both GDSC and CCLE using the common concentration range. ABC measures the difference between two drug-dose response curves by estimating the absolute area between these curves, which ranges from 0% (perfect consistency) to 100% (perfect inconsistency). The ABC statistic identified highly consistent (Figure 5A, B) and highly inconsistent (Figure 5C, D) dose-response curves between GDSC and CCLE. The mean of the ABC estimates for all drug-cell

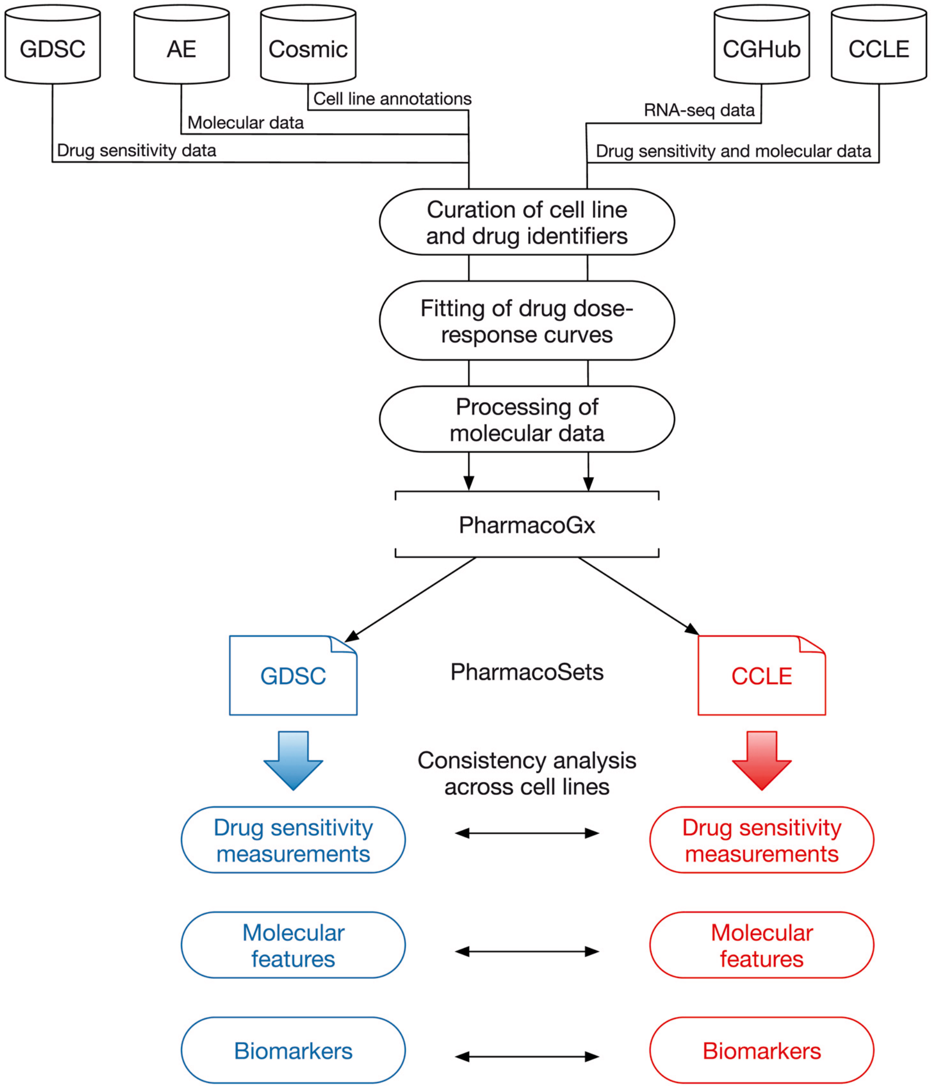
GDSC
AE
Cosmic
CGHub
CCLE
Cell line annotations
Molecular data
Drug sensitivity data
RNA-seq data
Drug sensitivity and molecular data
Curation of cell line and drug identifiers
Fitting of drug dose-response curves
Processing of molecular data
PharmacoGx
GDSC
PharmacoSets
CCLE
Consistency analysis across cell lines
Drug sensitivity measurements
Drug sensitivity measurements
Molecular features
Molecular features
Biomarkers
Biomarkers

Figure 1. Analysis design. GDSC: Genomics of Drug Sensitivity in Cancer; AE: ArrayExpress; Cosmic: Catalogue of Somatic Mutations in Cancer; CGHub: Cancer Genomics Hub; CCLE: Cancer Cell Line Encyclopedia.

line combinations was 10% (Supplementary Figure 1A), with paclitaxel yielding the highest discrepancies (Supplementary Figure 1B).

We compared biological replicates in GDSC, which were performed independently at the Massachusetts General Hospital (MGH) and the Wellcome Trust Sanger Institute (WTSI). These experiments are comprised of 577 cell lines treated with AZD6482, a PI3Kβ inhibitor screened in GDSC (Supplementary File 3). We computed the ABC of these biological replicates and observed both highly consistent and inconsistent cases (Supplementary Figure 2). We then computed the median ABC values for each pair of drugs in GDSC and used these as a distance metric for complete linkage hierarchical clustering. We found that the MGH- and WTSI-administered AZD6482 experiments clustered together, suggesting that the differences between dose-response curves of biological replicates were smaller than the differences observed between different drugs (Supplementary Figure 3A). We performed the same clustering analysis by computing the ABC-based distance between all the drugs in GDSC and CCLE and observed that only three out of the 15 common drugs clustered tightly (17-AAG, lapatinib, and PHA–665752; Supplementary Figure 3B). Despite the small number of cell lines exhibiting sensitivity to PHA–665752 and lapatinib, these drugs closely clustered between GDSC and CCLE; however this was not the case for other targeted therapies, such as AZD0530, nilotinib, crizotinib and TAE684 Supplementary Figure 3B).

Although the ABC values provide a measure of the degree of consistency between studies, it is the AUC and IC_{50} estimates, and their correlation with molecular features (such as mutational status and gene expression) that are commonly used to assess drug response.

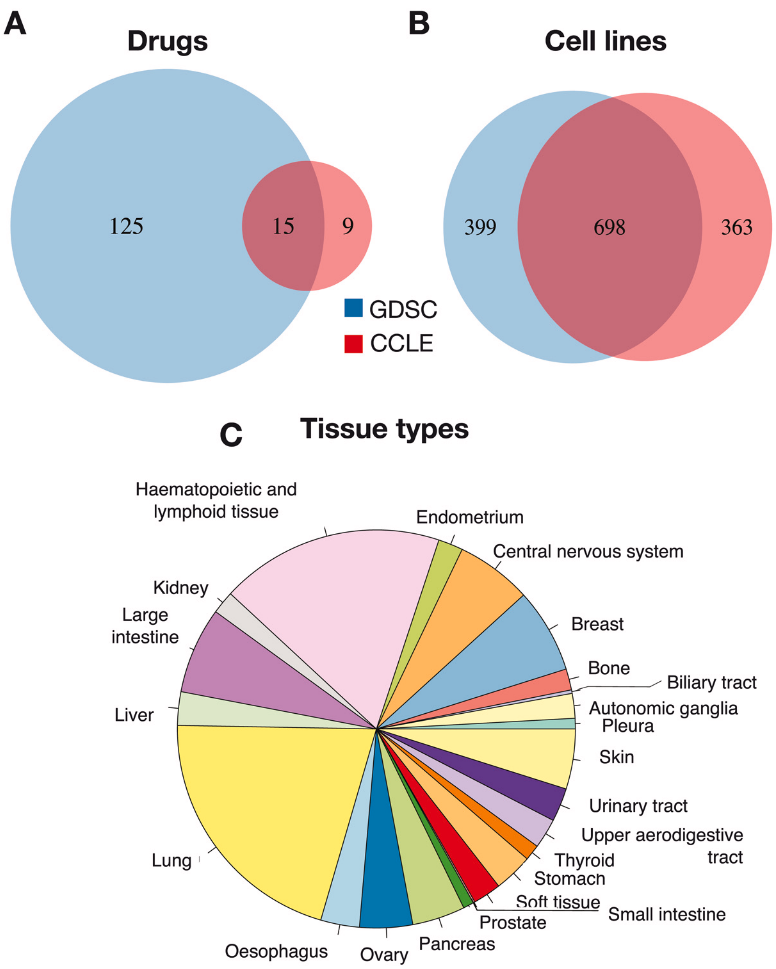

Figure 2. Intersection between GDSC and CCLE. Overlap of (**A**) drugs, (**B**) cell lines and (**C**) tissue types.

Therefore we revisited our comparative analysis of the drug sensitivity data using the expanded data and the standardized methods implemented in our *PharmacoGx* platform. Using the same three-parameter sigmoid model to fit drug dose-response curves in GDSC and CCLE (see Methods), we recomputed AUC and IC_{50} values and observed very high correlation between published and recomputed drug sensitivity values for each study individually (Spearman > 0.93; Figure 6; Dataset 2).

It has been suggested that some of the observed inconsistencies between the GDSC and CCLE may be due to the nature of targeted therapies, which are expected to have selective activity against some cell lines[10,14,15]. This is a reasonable assumption as the measured response in insensitive cell lines may represent random technical noise that one should not expect to be correlated between experiments. We therefore decided to clearly discriminate between targeted drugs with narrow growth inhibition effects and drugs with broader effects. We used the full GDSC and CCLE datasets to compare the variation of the drug sensitivity data of known targeted and cytotoxic therapies as classified in the original studies (Supplementary Figure 4). We observed that drugs can be classified in these two categories based on median absolute deviation (MAD) of the estimated AUC values (Youden's optimal cutoff[16] of AUC MAD > 0.13 for cytotoxic drugs). We then used this cutoff on the common drug-cell line combinations in GDSC and CCLE to define three classes of drugs (Supplementary Figure 5):

- **No/little effect**: Drugs with minimal observed activity (typically active in less than five sensitive cell lines with AUC > 0.2 or IC_{50} < 1 µM in either study). This class includes sorafenib, erlotinib and PHA−665752.
- **Narrow effect**: Targeted drugs with activity observed for only a small subset of cell lines (AUC MAD ≤ 0.13). This group includes nilotinib, lapatinib, nutlin-3, PLX4720, crizotinib, PD-0332991, AZD0530, and TAE684.
- **Broad effect:** Drugs producing a response in a large number of cell lines (AUC MAD > 0.13). This includes AZD6244, PD-0325901, 17-AAG and paclitaxel.

We then compared the AUC (Figure 7, Supplementary Figure 6 and Supplementary Figure 7 for published AUC, recomputed AUC and AUC computed over the common concentration range,

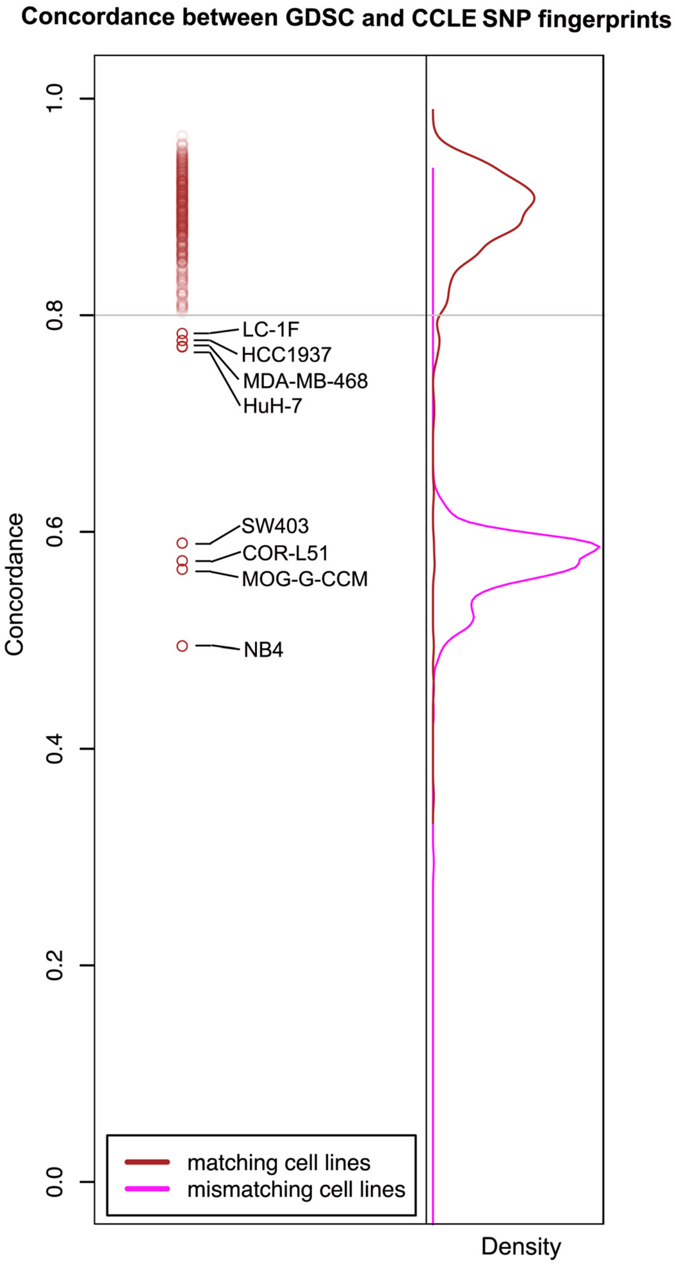

Figure 3. SNP fingerprinting between cancer cell lines screened in GDSC and CCLE.

respectively) and IC_{50} (Supplementary Figure 8 and Supplementary Figure 9) values and calculated the consistency of drug sensitivity data between studies using all common cases and only those that the data suggested were sensitive in at least one study (Figure 8 and Supplementary Figure 10 for AUC and IC_{50}, respectively, and Dataset 3). Given that no single metric can capture all forms of consistency, we extended our previous study by using the Pearson correlation[17], Spearman[18], and Somers' Dxy[19] rank correlation coefficients to quantify the consistency of continuous drug sensitivity measurements across studies (see Methods).

As expected, no consistency was observed for drugs with "no effect" (Figure 8A). For AUC of drugs with narrow and broad effects, Somers' Dxy was the most stringent, with consistency estimated to be < 0.4 except for two drugs (PD-0325901 and 17-AAG), which were also the two drugs identified as the most consistent using Spearman correlation (ρ ~ 0.6; Figure 8A). However, these statistics did not capture potential consistency for the most targeted therapies, nilotinib, crizotinib, and PLX4720, for which the Pearson correlation coefficient gave the best evidence of concordance, as this statistics is strongly influenced by a small number of highly sensitive cell lines (Figure 7). Our results concur with the

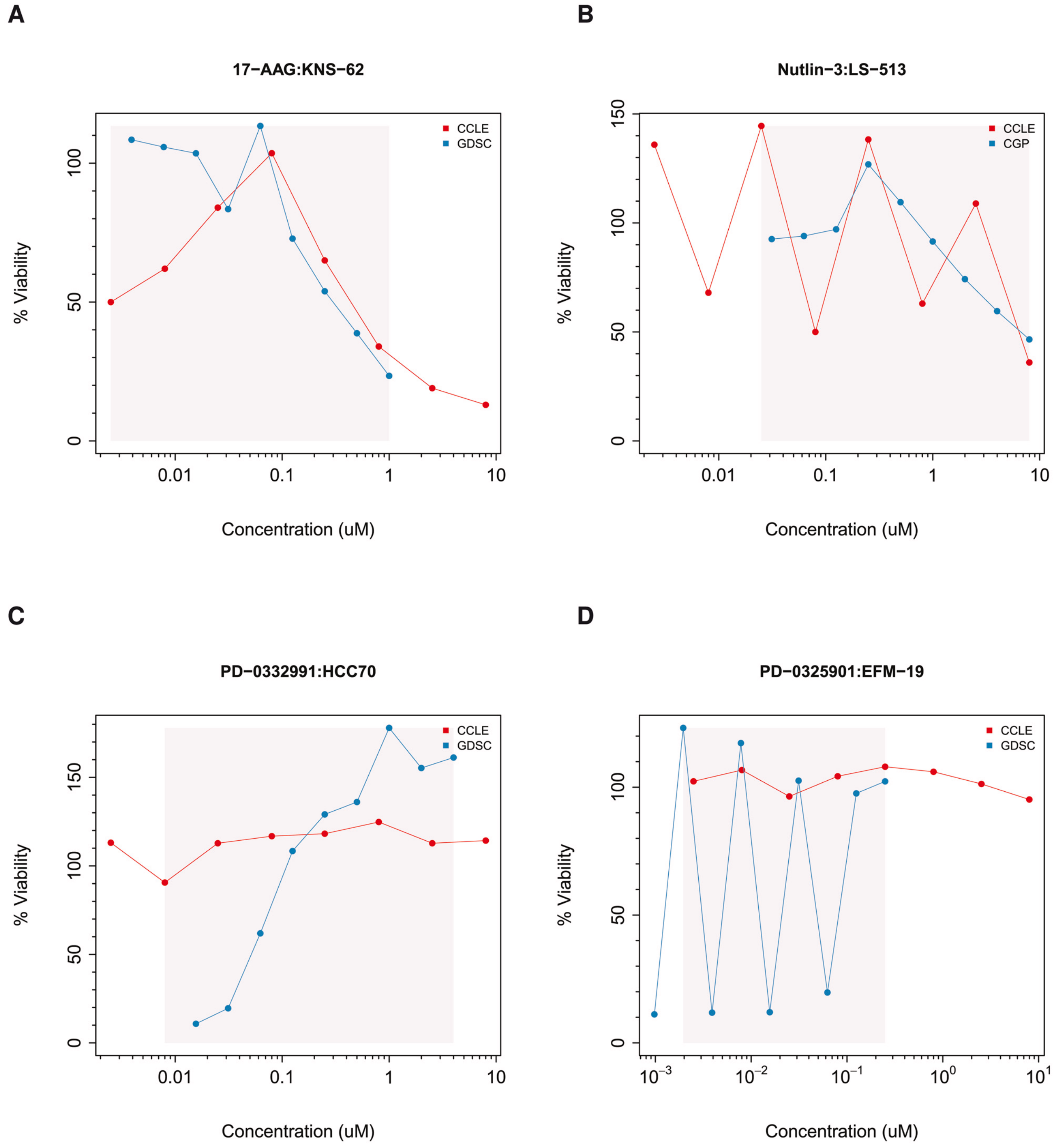

Figure 4. Examples of noisy drug dose-response curves identified during the filtering process in GDSC and CCLE. The grey area represents the common concentration range between studies. (**A**) JNS-62 cell line treated with 17-AAG; (**B**) LS-513 treated with nutlin-3; (**C**) HCC70 cell lines treated with PD-0332991; and (**D**) EFM-19 cell line treated with PD-0325901. Parameters have been set to ϵ = 25 and ρ = 0.80 (Supplementary methods). Red curve in (**A**) is the noisy due to violation of constraint 2, redcurve in (**B**) due to violation of constraint 1, blue curve in (**C**) is the noisy due to violation of constraint 2, blue curve in (**B**) due to violation of constraint 1 (Supplementary methods).

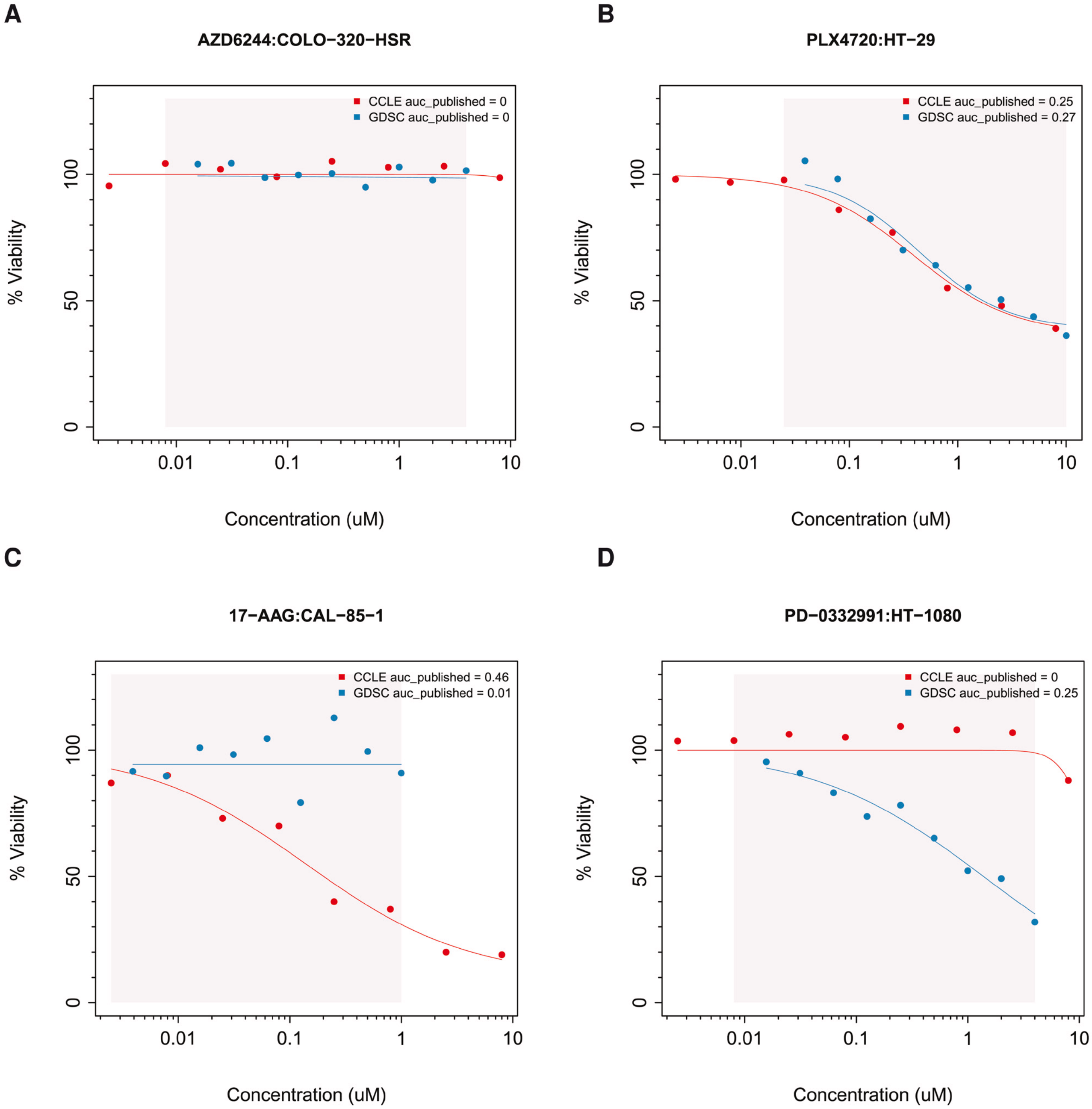

Figure 5. Examples of (**A**,**B**) consistent and (**C**,**D**) inconsistent drug dose-response curves in GDSC and CCLE. The grey area represents the common concentration range between studies. (**A**) COLO-320-HSR cell line treated with AZD6244; (**B**) HT-29 treated with PLX4720; (**C**) CAL-85-1 cell lines treated with 17-AAG; and (**D**) HT-1080 cell line treated with PD-0332991.

recent comparative study published by the GDSC and CCLE investigators[15].

We then restricted our analysis to the cell lines identified as sensitive in at least one study and computed the same consistency measures (Figure 8B). To our surprise, eliminating the insensitive cell lines resulted in decreased consistency for most drugs, which suggests a high level of inconsistency across sensitive cell lines, with the only exceptions of the targeted drugs nilotinib and crizotinib.

To test whether discretization of drug sensitivity data into binary calls (“insensitive” vs. “sensitive”; see Methods) improves consistency across studies, we used three association statistics, the Matthews correlation coefficient[20], Cramer’s V[21], and the informedness[22] statistics (Figure 8C). These statistics are designed for use with imbalanced classes, which is particularly relevant in large pharmacogenomic datasets where, for targeted therapies, there are often many more insensitive cell lines than sensitive ones. As expected, some of the targeted therapies, nilotinib and

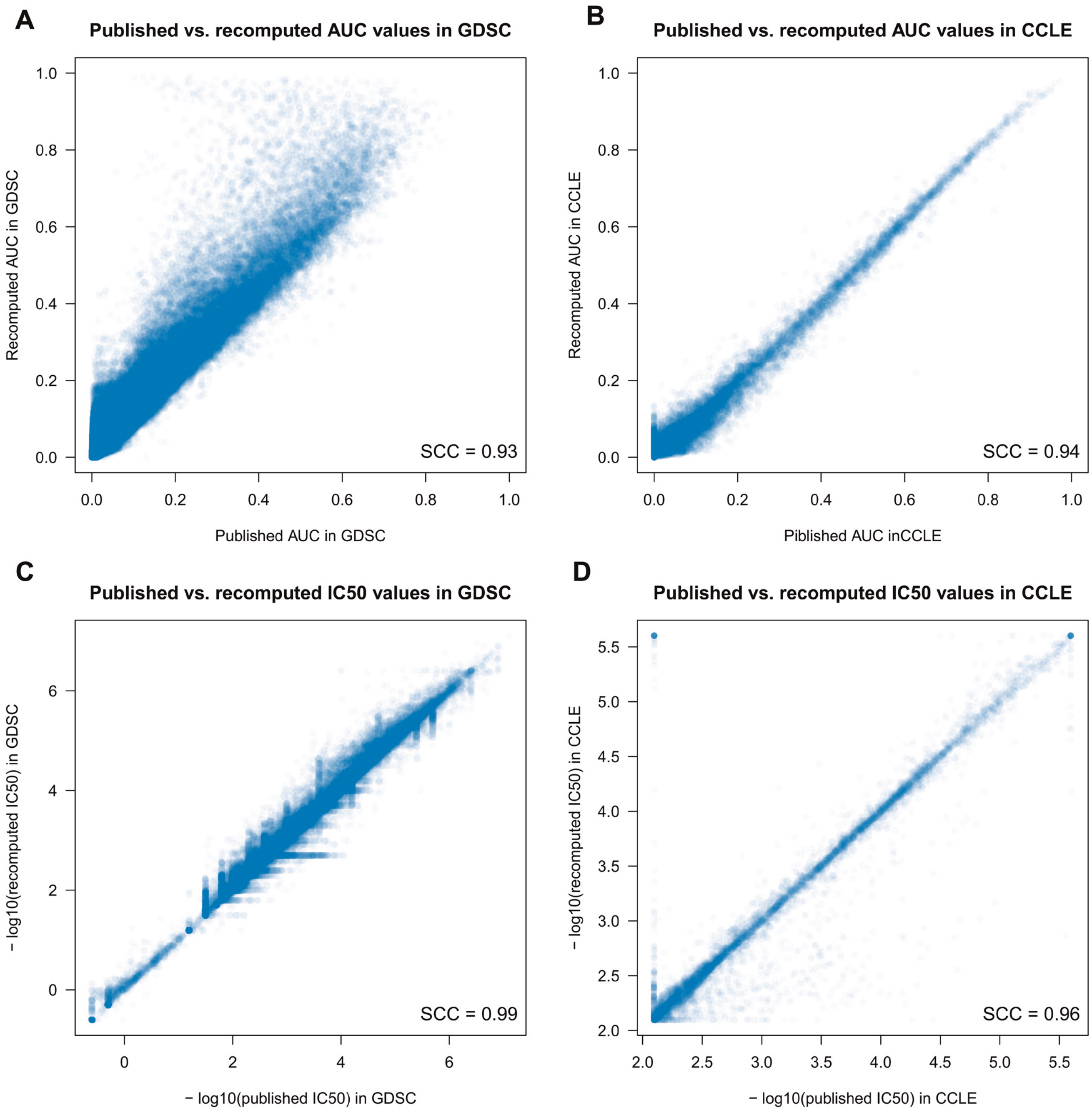

Figure 6. Comparison between published and recomputed drug sensitivity values between GDSC and CCLE. (**A**) AUC in GDSC; (**B**) AUC in CCLE; (**C**) IC_{50} in GDSC; and (**D**) IC_{50} in CCLE. SCC stands for Spearman correlation coefficient.

PLX4720 (and nutlin-3 using informedness), yielded high level of consistency, but this was not the case for the other targeted therapies. We also found that the drug sensitivity calls for drugs with broader inhibitory effects were also poorly correlated between studies (Figure 8C).

We performed the same analysis using IC_{50} values truncated to the maximum concentration used for each drug in each study separately. We observed similar patterns with nilotinib and crizotinib yielding moderate to high consistency across studies (Supplementary Figure 10). Note that Somers' Dxy rank correlation is biased in the presence of many repeated values in the datasets being analyzed, which is the case for truncated IC_{50}— pairs of cell line with identical IC_{50} values in one dataset but not in the other will not be taken into account as evidence of inconsistency — which explains the artifactual perfect consistency it suggests for both nilotinib and crizotinib.

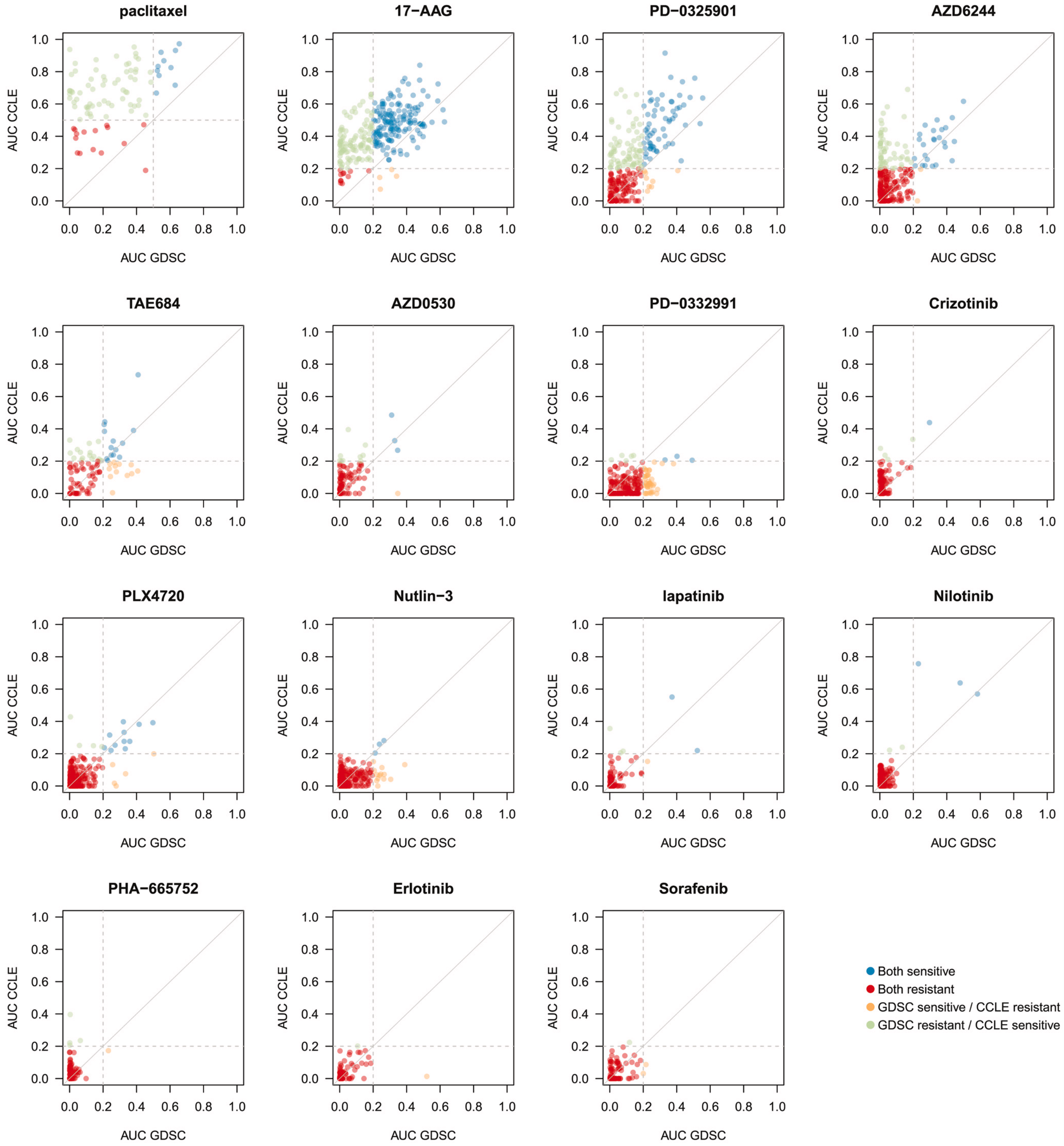

Figure 7. Comparison of AUC values as published in GDSC and CCLE. Cell lines with AUC >0.2 were considered as sensitive (AUC >0.4 for paclitaxel). In case of perfect consistency, all points would lie on the grey diagonal. The drugs are ranked based on their category: broad effect (AZD6244, PD–0325901, 17-AAG and paclitaxel), narrow effect (nilotinib, lapatinib, nutlin-3, PLX4720, crizotinib, PD-0332991, AZD0530, and TAE684) and no/little effect (sorafenib, erlotinib and PHA–665752).

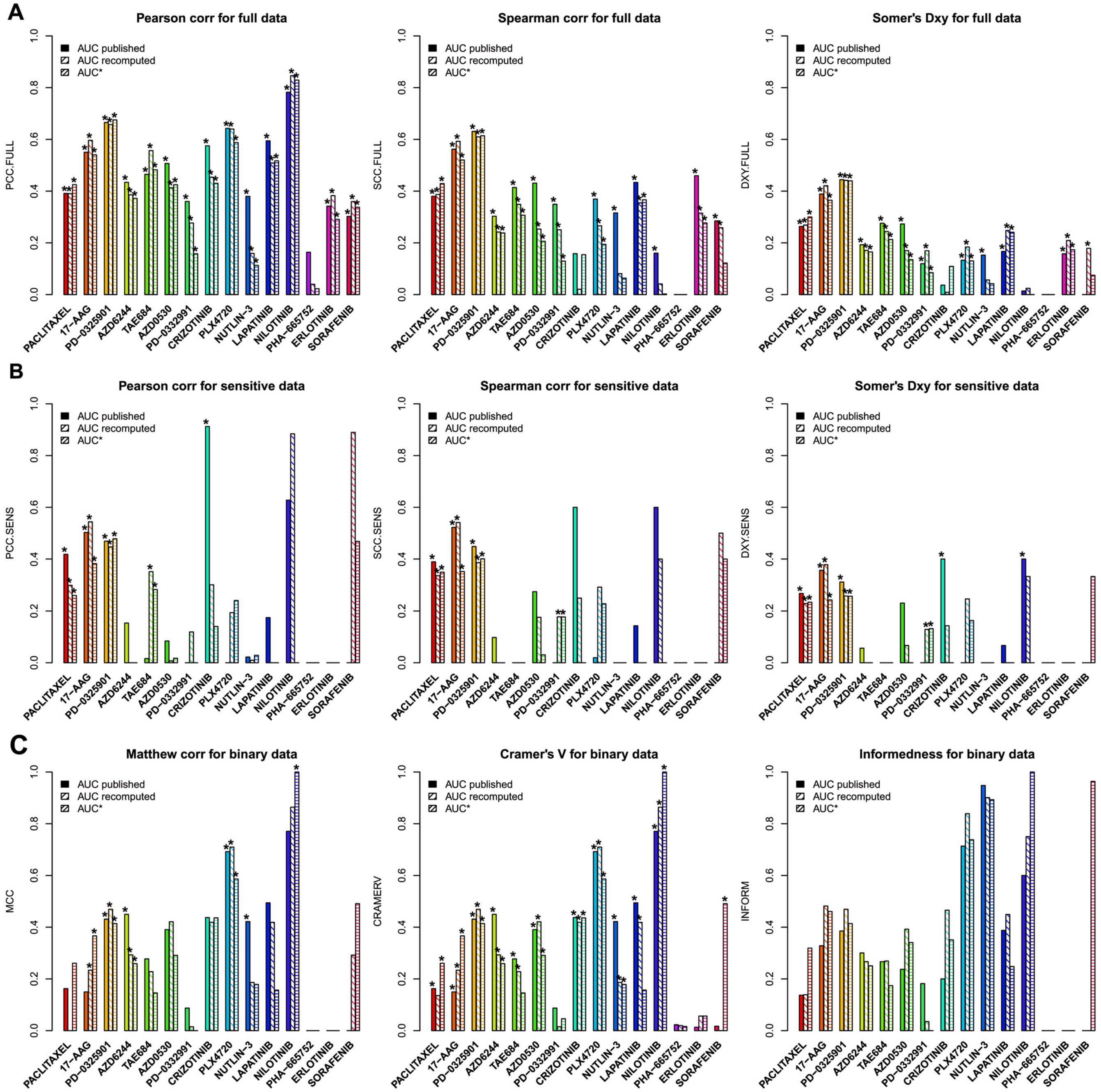

Figure 8. Consistency of AUC values as published and recomputed within *PharmacoGx*, with AUC* being computed using the common concentration range between GDSC and CCLE. The consistency is computed across cell lines, i.e., for each drug, a vector of drug sensitivity measures (AUC, IC_{50},...) is extracted from GDSC and CCLE and compared. (**A**) Consistency assessed using the full set of cancer cell lines screened in both studies. (**B**) Consistency assessed using only sensitive cell lines (AUC > 0.2 and AUC > 0.4 for targeted and cytotoxic drugs, respectively). (**C**) Consistently assessed by discretizing the drug sensitivity data using the aforementioned cutoffs for AUC. PCC: Pearson correlation coefficient; SCC: Spearman rank-based correlation coefficient; DXY: Somers' Dxy rank correlation; MCC: Matthews correlation coefficient; CRAMERV: Cramer's V statistic; INFORM: Informedness. The symbol '*' indicates whether the consistency is statistically significant (p<0.05).

Consistency of molecular profiles across cell lines

Discovering new biomarkers predictive of drug response requires both robust pharmacological data and molecular profiles. In our original study, we showed that the gene expression profiles for each cell line profiled by both GDSC and CCLE were highly consistent. However, we found that mutation profiles were only moderately consistent, a result that was later confirmed by Hudson *et al.*[23].

There have been questions as to whether the measures of consistency we reported for drug response should be compared to those we reported for gene expression. Specifically, we reported correlations based on comparing drug response "across cell lines," meaning that we examined the correlation of response of each cell line to a particular drug reported by the GDSC with the response of the same cell line to the same drug reported by the CCLE. In contrast we reported correlation of gene expression levels "between cell lines," meaning that we compared the expression of all genes within each cell line in the GDSC to the expression of all genes in the same cell line in the CCLE (see Supplementary Methods). It has been suggested that a more valid comparison would be to compare both drug response and gene expression across cell lines. We report the results of such an "across cell lines" analysis of gene expression here, computed using techniques analogous to those we used to compare drug response.

We began by comparing the distribution of gene expression measurements generated using the microarray Affymetrix HG-U219 platform in GDSC, the microarray Affymetrix HG-U133PLUS2 platform and the new Illumina RNA-seq data in CCLE (Supplementary Figure 11). We observed similar bimodal distributions, suggesting the presence of a natural cutoff to discriminate between lowly vs. highly expressed genes. We therefore fit a mixture of two gaussians and identified an expression cutoff for each platform separately (Supplementary Figure 11). We then compared the consistency of continuous and discretized gene expression values between (*i*) the microarray Affymetrix HG-U133PLUS2 and Illumina RNA-seq platforms within CCLE (intra-lab consistency); (*ii*) the microarray Affymetrix HG-U219 and HG-U133PLUS2 platforms used in GDSC and CCLE, respectively (microarray, inter-lab consistency); and (*iii*) the microarray Affymetrix HG-U219 and Illumina RNA-seq platforms used in GDSC and CCLE, respectively (inter-lab consistency). We performed a similar analysis for CNV log-ratios and observed high consistency across cell lines (Figure 9A). Supporting our previous observations, we found that CNV and gene expression measurements are significantly more consistent than drug sensitivity values when using all cell lines (Wilcoxon rank sum test p-value < 0.05; Figure 9A; Supplementary Figure 12A).

Similarly to the filtering we performed for drug sensitivity data, we subsequently restricted our analysis to the cell lines showing high expression of a given gene/cell line combination in at least one study. Again, CNV and gene expression measurements were significantly more consistent than drug sensitivity values in this case (Wilcoxon rank sum test p-value < 0.05; Figure 9B; Supplementary Figure 12B). When dichotomizing data into lowly/highly expressing, amplifications/deletions, and wild type/mutated cell lines and insensitive/sensitive cell lines, the CNV and gene expression data were still more consistent (Figure 9C) although the difference was not always significant (Supplementary Figure 12C). Concurring with the report of Hudson *et al.*[23], we observed low consistency for mutation calls across cell lines (Figure 9C).

Consistency of gene-drug associations

The primary goal of the GDSC and CCLE studies was to identify new genomic predictors of drug response for both targeted and cytotoxic therapies. We therefore evaluated whether the good consistency in drug sensitivity data observed for nilotinib, PLX4720 and crizotinib, and the moderate consistency observed for 17-AAG and PD-0332901 would translate in reproducible biomarkers. We estimated gene–drug associations by fitting, for each gene and drug, a linear regression model including gene expression, CNV and mutations as predictors of drug sensitivity, adjusted for tissue source (see Methods). As illustrated in Figure 1, we used the molecular and pharmacological data generated independently in GDSC and CCLE to identify and compare gene-drug associations. This approach prevents any information leak between the two datasets, which may lead to overoptimistic consistency between the studies, as in the recent comparative study published by the GDSC and CCLE investigators[9]. Given the high correlation between the published and recomputed AUC values in each study (Figure 6) and their similar consistency (Figure 9), all gene-drug associations were computed using published AUC for clarity.

We first computed the strength and significance of each gene in both datasets separately. Similarly to our initial study[7], the strength of a given gene-drug association is provided by the standardized coefficient associated to the corresponding gene profile in the linear model and its significance is provided by the p-value of this coefficient (see Methods). We then identified gene-drug associations that were reproducible in both datasets (same sign and False Discovery Rate [FDR] < 5%) or that were dataset-specific (different sign or significant in only one dataset) using continuous (Supplementary Figure 13 and Supplementary Figure 14 for common and all cell lines, respectively) and discretized (Supplementary Figure 15 and Supplementary Figure 15 for common and all cell lines, respectively) published AUC values as drug sensitivity data. We assessed the overlap of gene-drug associations discovered in both datasets using the Jaccard index[24]. All Jaccard indices were low, with nilotinib yielded the largest overlap of gene-drug associations (32%), followed by PD-0325901 and erlotinib (almost 20%), while the other drugs yielded less than 15% overlap (Supplementary Figure 17). Our results further indicate that larger overlap exists for gene-drug associations identified using the continuous drug sensitivity data compared with associations using discretized drug sensitivity calls (Wilcoxon signed rank test p-value of 4×10^{-2} and 2×10^{-3} for the common set and the full set of cell lines, respectively). We therefore focused our analyses on the gene-drug associations

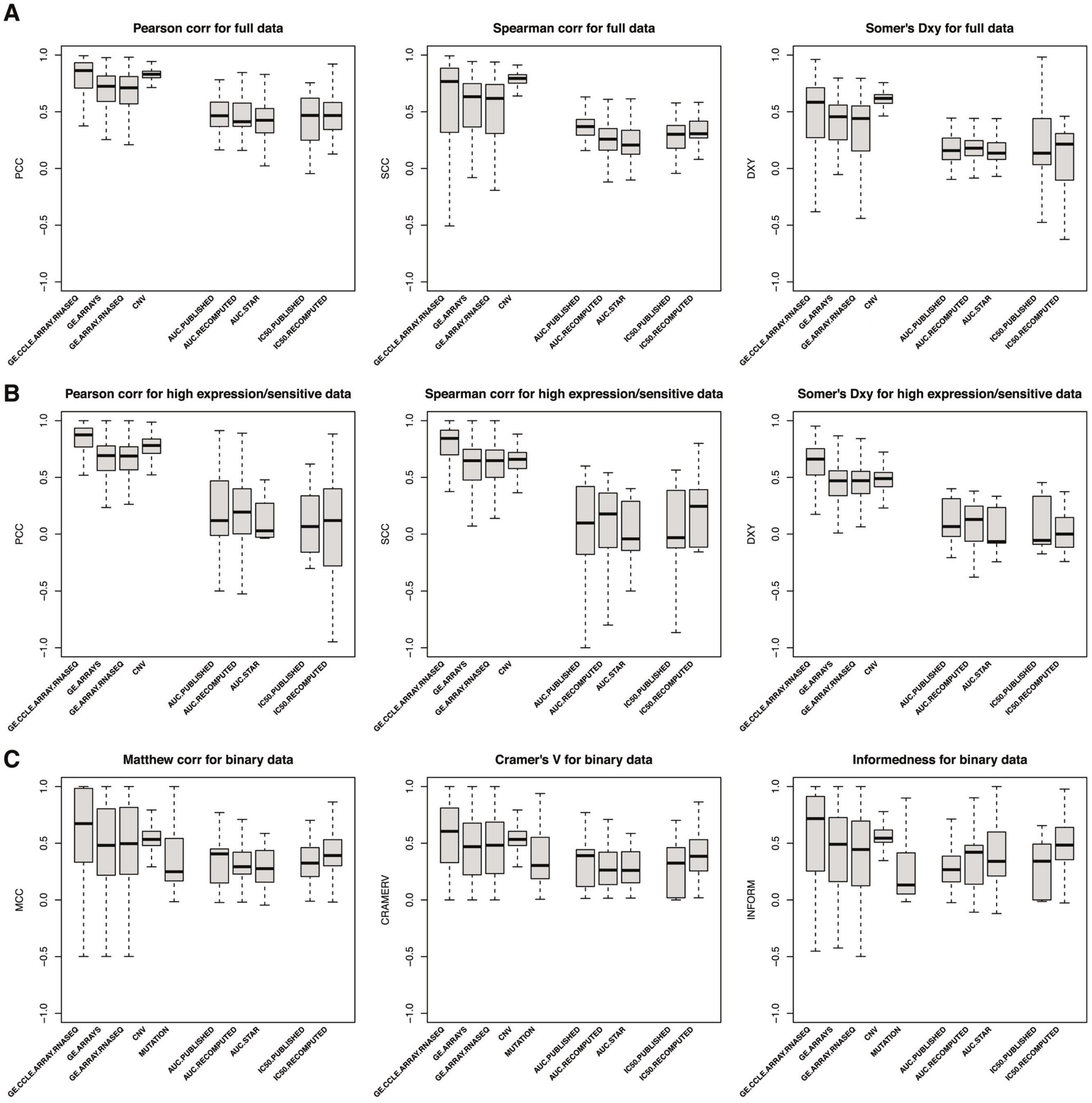

Figure 9. Consistency of molecular profiles (gene expression, copy number variation and mutation) and drug sensitivity data between GDSC and CCLE using multiple consistency measures. (**A**) Consistency assessed using the full set of cancer cell lines screened in both studies. (**B**) Consistency assessed using only sensitive cell lines (AUC >0.2 / IC_{50} <1 μM and AUC >0.4 / IC_{50} <10 μM for targeted and cytotoxic drugs, respectively). (**C**) Consistently assessed by discretizing the molecular and drug sensitivity data. GE.CCLE.ARRAY.RNASEQ: Consistency between gene expression data generated using Affymetrix HG-U133PLUS2 microarray and Illumina RNA-seq platforms within CCLE; GE.ARRAYS: Consistency between gene expression data generated using Affymetrix HG-U133A and HG-U133PLUS2 microarray platforms in GDSC and CCLE, respectively; GE.ARRAY.RNASEQ: Consistency between gene expression data generated using Affymetrix HG-U133A microarray and Illumina RNA-seq platforms in GDSC and CCLE, respectively; CNV: Consistency of copy number variation data in CCLE and GDSC, respectively; MUTATION: Consistency of mutation profiles in CCLE and GDSC, respectively. PCC: Pearson correlation coefficient; SCC: Spearman rank-based correlation coefficient; DXY: Somers' Dxy rank correlation; MCC: Matthews correlation coefficient; CRAMERV: Cramer's V statistic; INFORM: Informedness.

identified using continuous published AUC values. The number (and identity) of gene-drug associations computed using continuous published AUC values are provided in Supplementary Table 1 and Supplementary Table 2 (Dataset 5 and Dataset 6) for common and all cell lines, respectively.

Given that simply intersecting significant gene-drug associations identified in each dataset separately yielded poor reproducibility for all drugs, we sought to more closely mimic the biomarker discovery and validation process. We therefore used one dataset to discover significant gene-drug associations and test whether this subset of markers validated in an independent dataset. Using the discovery dataset, gene-drug associations are first ranked by nominal p-values and their FDR is computed. An association is selected if it is part of the top 100 markers and its FDR is less than 5%. This procedure ensure to control for both significance and number of selected biomarkers, which can vary with respect to the cell line panel used for the analysis (larger panels enable the identification of more significant biomarkers due to increased statistical power). A gene-drug association is validated in an independent dataset if its nominal p-value is less than 0.05 and its "direction", that is whether the marker is associated with sensitivity or resistance, is identical to the one estimated during the discovery process.

We computed the proportions of validated gene-drug associations for each drug using all available genomic molecular data profiles in GDSC as discovery set and CCLE as validation set, and *vice versa* (Figure 10). Overall, we found that biomarkers for PD-0325901, PLX4720 and nilotinib yielded a high validation rate (> 80%) with either dataset as discovery set using the common cell lines screened in GDSC and CCLE (Figure 10A). When using the entire cell line panels used in each study, two more drugs -- lapatinib and erlotinib -- yielded high validation rate (Figure 10B). 17-AAG, and TAE684 yielded validation rate between 60% and 80%, while the other drugs yielded a validation rate around 50% or lower. For ten out of the fifteen drugs, using the entire panel of cell lines screened in each study (Figure 10B) improved the validation rate compared to limiting the analysis to common cell lines (Figure 10A). However, validation rate decreased for three drugs, suggesting that using large, but different panels of cell lines may increase statistical power but could also introduce biases in the biomarker discovery process.

We then investigated whether higher validation rates would be obtained by using more stringent significance threshold and relaxing the constraint on the number of significant associations in the discovery set (Supplementary Figure 18 and Supplementary Figure 19). Using common cell lines, we found that proportion of validated gene-drug association monotonically increases with FDR stringency for six drugs, with very high validation rate for the most stringent FDR cutoff (validation rate > 80% for FDR < 0.1%) for 17-AAG, PD-0325901, PLX4720 and nilotinib using either dataset as discovery set (Supplementary Figure 18). Using the entire panel of cell lines in each study actually improved validation rate for six drugs, AZD6244, TAE684, AZD0530, lapatinib — and erlotinib and sorafenib, for which insufficient number of sensitive cell lines were screened in both GDSC and CCLE (Supplementary Figure 19). However, validation rate decreased for 17-AAG, crizotinib and PLX4720, which suggests again that large, but different panels of cell lines might introduce selection bias for some drugs.

Known biomarkers

As reported in the original GDSC (1) and CCLE (2) publications and in recent reports[10,14,15], several known biomarkers for targeted therapies have been shown to be predictive in both GDSC and CCLE. In our initial comparative study we also found the following known gene-drug associations:

- BRAF mutations were significantly associated with sensitivity to MEK inhibitors (AZD6244 and PD-0325901) and BRAF^{V600E} inhibitor (PLX4720) with nominal p-values < 0.01; see Supplementary File 10–Supplementary File 13 of our initial study.
- ERBB2 expression was significantly associated with sensitivity to lapatinib with nominal p-value = 0.04 and 8.4×10^{-15} for GDSC and CCLE, respectively; see Supplementary File 4 and Supplementary File 5 of our initial study.
- NQ01 expression was significantly associated with sensitivity to 17-AAG with nominal p-value = 2.4×10^{-13} and 6.2×10^{-14} for GDSC and CCLE, respectively; see Supplementary File 4 and Supplementary File 5 of our initial study.
- MDM2 expression was significantly associated with sensitivity to Nutlin-3 with nominal p-value = 7.7×10^{-18} and 7×10^{-8} for GDSC and CCLE, respectively; see Supplementary File 4 and Supplementary File 5 of our initial study.
- ALK expression was significantly associated with sensitivity to TAE684 with nominal p-value = 1.6×10^{-9} and 1.7×10^{-9} or GDSC and CCLE, respectively; see Supplementary File 4 and Supplementary File 5 of our initial study.

We revisited our biomarker analysis using the new data released by GDSC and CCLE to test whether additional known biomarkers can be identified. We recomputed all gene-drug associations based on expression, mutation, gene-fusion and amplification data using the common cell lines between studies Dataset 5, and entire panel of cell lines in each study (Dataset 6). We confirmed the reproducibility of the known associations reported in our initial study, but we were not able to find reproducible associations for EGFR mutations with response to AZD0530 and erlotinib, and HGF expression with response to crizotinib (Table 1). The reproducibility of the majority of these previously known associations attests to the relevance of the GDSC and CCLE datasets although our results demonstrated that the noise and inconsistency in drug sensitivity data render discovery of new biomarkers difficult for the majority of the drugs.

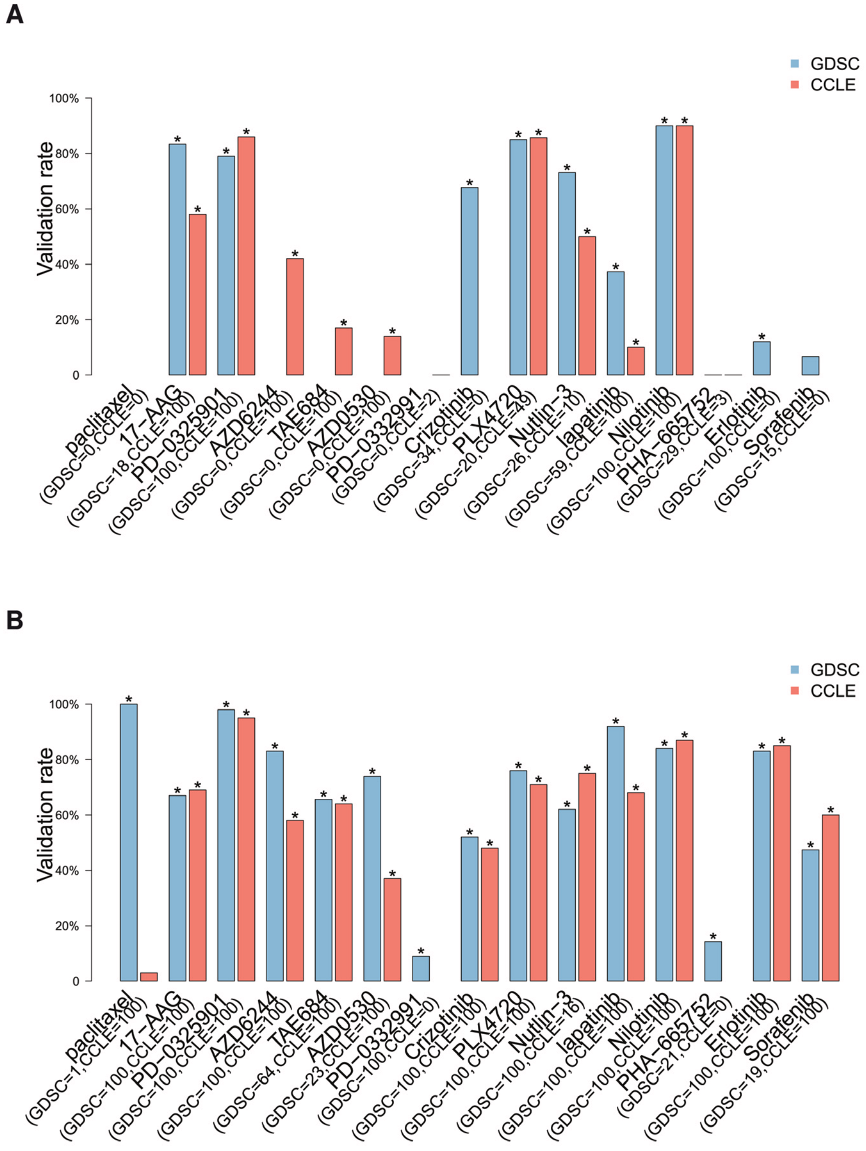

Figure 10. Proportion of gene-drug associations identified in a discovery set (top 100 gene-drug associations as ranked by p-values and FDR < 5%) and validated in an independent validation dataset. In blue and red are the gene-drug associations identified in GDSC and CCLE, respectively. Associations are identified using molecular profiles including gene expression, mutation and copy number variation data as input and (**A**) continuous published AUC values as output in a linear model using only common cell lines or (**B**) all cell lines. The number of selected gene-drugs associations in each datasets is provided in parentheses. The symbol '*' represents the significance of the proportion of validated gene-drug associations, computed as the frequency of 1000 random subsets of markers of the same size having equal or greater validation rate compared to the observed rate.

Table 1. List of known gene-drug associations with their effect size and significance in GDSC and CCLE. Gene-drug associations were estimated using the full panel of cell lines and AUC as measure of drug sensitivity.

Drug	Gene	Type	GDSC effect size	GDSC pvalue	CCLE effect size	CCLE pvalue	Reproducible
Nilotinib	BCR_ABL	fusion	6.13	1.10E-51	5.84	2.60E-28	YES
17-AAG	NQO1	expression	0.55	5.30E-39	0.6	4.70E-29	YES
	HSP90AA1	expression	0	9.00E-01	0.02	6.40E-01	NS
	HSP90AB1	expression	0.01	7.40E-01	0	9.40E-01	NS
PD-0325901	BRAF	mutation	0.83	6.40E-09	0.82	8.10E-10	YES
	MAP2K1	expression	-0.07	7.10E-02	-0.02	6.70E-01	NS
	MAP2K2	expression	0.02	5.60E-01	0.03	5.10E-01	NS
AZD6244	BRAF	mutation	0.93	6.10E-10	0.86	3.70E-10	YES
	MAP2K1	expression	-0.04	2.80E-01	-0.06	1.90E-01	NS
	MAP2K2	expression	0.01	8.40E-01	0.01	7.70E-01	NS
TAE684	ALK	expression	0.28	2.20E-07	0.26	1.10E-08	YES
AZD0530	EGFR	mutation	0.03	9.50E-01	0.51	8.20E-03	NO
	BCR_ABL	fusion	3.87	2.60E-18	3.35	3.50E-09	YES
	SRC	expression	0.07	2.80E-01	0.07	1.60E-01	NS
PD-0332991	CDK4	expression	0.03	5.10E-01	0	9.50E-01	NS
	CDK6	expression	0.08	7.50E-02	-0.02	6.60E-01	NS
Crizotinib	HGF	expression	-0.03	6.50E-01	0.28	1.30E-09	NO
	MET	amplification	0.1	8.10E-02	0.29	3.80E-09	NO
	ALK	expression	0.58	3.90E-33	0.13	6.80E-03	YES
PLX4720	BRAF	mutation	1.75	8.60E-46	1.38	2.20E-27	YES
Nutlin-3	MDM2	expression	0.39	2.00E-25	0.31	8.40E-12	YES
lapatinib	ERBB2	expression	0.42	1.10E-12	0.53	3.40E-33	YES
		amplification	0.24	8.40E-06	0.39	4.20E-19	YES
	EGFR	expression	0.26	1.20E-04	0.16	7.20E-03	YES
PHA-665752	HGF	expression	0.04	4.90E-01	0.06	2.00E-01	NS
	MET	amplification	0.22	2.20E-04	0.02	6.80E-01	NO
Erlotinib	EGFR	mutation	0.71	1.90E-01	1.27	2.40E-12	NO
Sorafenib	PDGFRA	expression	0.06	3.90E-01	0.13	7.90E-03	NO
	KDR	expression	0.01	8.70E-01	-0.02	6.90E-01	NS
	KIT	mutation			0.11	6.30E-01	
		expression	0.06	3.50E-01	-0.01	7.70E-01	NS
	FLT1	expression	-0.04	5.20E-01	0.01	8.30E-01	NS
	FLT3	expression	0.32	4.80E-08	0.33	1.10E-13	YES
	FLT4	expression	0.12	4.30E-02	-0.02	6.40E-01	NO
	RAF1	expression	0.03	6.40E-01	0.06	2.30E-01	NS
	BRAF	mutation	-0.34	1.70E-01	0.14	3.80E-01	NS
	BRAF	expression	-0.11	9.30E-02	0.02	7.40E-01	NS

Discussion

Our original motivation in analyzing the GDSC and CCLE data was to discover predictive genomic biomarkers of drug response. When we applied a number of methods using one study to select genomic features and to train a classifier, and then applied it to predict reported drug response in the second study, our predictive models failed to validate for half of the drugs tested[3]. Indeed, out of nine predictors yielding concordance index[25] ≥0.65 in cross-validation in the training set (GDSC), only four were validated in identical cell lines treated with the same drugs in the validation set (CCLE)[3].

As we explored the reasons for this failure, we first checked whether cell lines could have drifted and consequently exhibited different transcriptional profiles between GDSC and CCLE. We found that any genome-wide molecular profile in one study would almost always identify "itself" (its purported biological replica) as being most similar among the cell lines in the other study. In a way this is not surprising. When gene expression studies were in their infancy, there were many reports that compared the results from studies and found that they were inconsistent and unreproducible in new studies — as demonstrated by the countless microarray signatures that fail to reproduce beyond their initial publication. As a result, scientists involved in gene expression studies "circled the wagons" and developed both much more standardized laboratory protocols and "best practices" for reproducible analysis, including data normalization and batch corrections, that now mean that independent measurements from different laboratories are far more often consistent and so can be used for signature development and validation[26,27].

Unexpectedly, when we compared phenotypic measures of drug response that were released by the GDSC and CCLE projects, we found discrepancies in growth inhibition effects of multiple anticancer agents. What that means in practice is that, for some drugs, a molecular biomarker of drug response learned from one study would not likely be predictive of the reported response in the other. And consequently, neither of the studies might be useful in predicting response in patients as many had hoped when these large pharmacogenomic screens were published.

The feedback from the scientific community on our analysis, the availability of new data from the GDSC and CCLE, as well as improvements in the *PharmacoGx* software platform we developed to support this type of analyses[11], prompted us to revisit the question of consistency in these studies to see if we could find a principled way to identify correlated drug response phenotypes. By testing a variety of methods of classifying the data, and choosing the metric which gave the best consistency for each drug, we were able to find moderate to good consistency of sensitivity data for two broad effect and three targeted drugs. We also confirmed the overall lack of consistency between the studies for eight drugs, while there were not enough sensitive cell lines that had been screened by both GDSC and CCLE to properly assess consistency for the remaining three drugs. The summary box included with this paper briefly describes the most significant issues that people have raised in discussing our previous findings with us and summarizes what we have found in our reanalysis.

Some have suggested that one way to improve correlation would have been to compare the studies and throw out the most discordant data as noise and then compare the remaining concordant data. While this would certainly find concordance in the remaining data, the approach is equivalent to fitting data to a desired result, which is bad practice and certainly could not be extended to other data sets or to the classification of patient tumors as responsive or nonresponsive to a particular therapy. There is, however, merit in the suggestion that one would not expect to see correlation in noise. And noise is precisely what one would expect to see in drug response data from cell lines that are resistant to a particular drug or nonresponsive across the range of doses tested. As reported here, filtering the data in each study independently to classify cell lines in a binary fashion, and then comparing the binary classification between studies using a variety of metrics developed to handle the intricacies of this sort of response data, also failed to find simple correlations in the data, except for three of the targeted therapies, nilotinib, PLX4720 and crizotinib. What this ultimately means is that the most and the least sensitive cell lines would not appear to be the same when comparing the two studies.

There are many reasons for potential differences in measured phenotypes reported by the GDSC and CCLE, including substantial differences in doses used for each drug and in the methods used to both assay cell viability and to estimate drug response parameters. By comparing GDSC and CCLE with an independent pharmacogenomic dataset published by GlaxoSmithKline (GSK), we showed that higher consistency is achieved when the same pharmacological assay is used (GSK and CCLE used the CellTiter-Glo assay, while GDSC used Syto60)[7,8]. Genentech also used the CellTiter-Glo assay and observed higher consistency of drug sensitivity data with CCLE compared to GDSC[10]. The authors elegantly evaluated the impact of cell viability readout, growth medium, and seeding density. They observed only weak impact of the choice of pharmacological assay as their follow-up screen with the Syto60 assay clustered closer to their own CellTiter-Glo screen than GDSC, suggesting that other parameters might have driven the inconsistency observed with GDSC[10]. They further showed that increased fetal bovine serum and seeding cell density had a systematic effect on mean cell viability. Pozdeyev *et al.* showed that restricting the computation of AUC to the concentration range shared between GDSC and CCLE, the equivalent of our AUC* drug sensitivity measure, yielded a small, but statistically significant improvement in consistent of pharmacological profiles[28]. Ultimately what our analysis and these recent reports suggest is that not only drug sensitivity measurements must be carefully and appropriately compared, but also that there is a pressing need for more robust pharmacological assays and standardized computational methods for modeling drug response. However, in the absence of a "gold standard" screening platform demonstrated to accurately recapitulate drug response *in vivo*, the use of multiple assays is critical to probe different biological aspects of growth inhibition. Given that GDSC and CCLE used different pharmacological assays, it makes the release of these pharmacogenomic data even more valuable.

The primary goal of the GDSC and CCLE studies was to link molecular features of a large panel of cancer cell lines to their sensitivity to cytotoxic and targeted drugs. The reproducibility of most of the known gene-drug associations provides evidence that these large pharmacogenomic datasets are biologically relevant. When we investigated whether we could find significant gene-drug associations discovered in one dataset that validate in the other independent dataset, we observed over 75% validation rate for the most significant molecular biomarkers for eight of 15 drugs, which is a major improvement over our initial comparative study. However, this does not suggest that one can use these studies to find new, reproducible gene-drug associations for the rest of the drugs, as the majority of associations can be found in only one dataset but not in both. However, GDSC and CCLE could be jointly analyzed to identify biomarkers that are robust to the use of different biological assays, and are therefore more likely to work in new biological contexts[29].

This study has several potential limitations. First, while the raw drug sensitivity data are publicly available for GDSC, these data have not been released within the CCLE study. We could not fit the drug dose-response curves using the technical triplicates but rather relied on the published median sensitivity values. The lack of technical replicates in CCLE also prevented us to assess the level of noise of the drug sensitivity measurements. Second, we discretized drug sensitivity values by selecting a common threshold to discriminate between insensitive ($AUC \leq 0.2$ and $IC_{50} \geq 1$ μM) and the rest of the cell lines for all the targeted agents. However, it is clear that such a threshold could be optimized for each drug, which might have an impact on the consistency of drug phenotypes and gene-drug associations based on binary sensitivity calls, as was done in breast cancer[30] and in our response to the critic of Geeleher *et al.*[31,32]. Lastly, the current set of mutations assessed in both study is small (64 mutations), which drastically limits the search for mutation-based and other genomic aberrations associated with drug response. The exome-sequencing data available within the new GDSC1000 dataset will enable to better explore the genomic space of biomarkers in cancer cell lines, and their reproducibility across studies.

Conclusion

As is true of many scientists working in genomics and oncology, we were excited when the GDSC and CCLE released their initial data sets and were hopeful that these projects would help to accelerate drug discovery and further the development of precision medicine in oncology. However, what we found initially, and what the reanalysis presented here further indicates, is that there are inconsistencies between the measured phenotypic response to drugs in these studies. Even in our reanalysis, where we used methods specific to individual drugs and the response characteristics of the cell lines tested, we were only able to find new biomarkers consistently predictive of response for around half of the drugs screened in both studies. Consequently, it is challenging to use the data from these studies to develop general purpose classification rules for all drugs.

Our finding that molecular profiles are significantly more consistent than drug sensitivity data, indicates that the main barrier to biomarker development using these data is the discrepancy in the reported response phenotypes for many drugs. The experimental protocols and pharmacological assays used in the GDSC and CCLE studies are the state-of-the-art for high-throughput drug screening projects. Even though technical and biological replicates are necessary to assess and account for noise in drug sensitivity measurements, it is clear that the assays used in GDSC and CCLE probe different aspects of the biology underlying drug-induced growth inhibition. Without knowing which assay is more relevant for *in vivo* drug response, more research will be required to best leverage these complementary assays for robust biomarker discovery.

From having worked in large-scale genomic analyses, we recognize the challenges involved in planning and executing such studies and commend the GDSC and CCLE for their work and for making all the data available. However, we strongly encourage the GDSC, the CCLE, the pharmacogenomics and bioinformatics communities as a whole, to invest the necessary time and effort to account for the noise in drug response measurements and the complementary nature of different assays in order to assure that these studies are relevant for predicting response in patients. The recent report from Genentech is a significant step in this direction. Ultimately, that effort will help to assure that mammoth undertakings in drug characterization can deliver on their promise to identify better therapies and biomarkers predictive of response.

Methods

The PharmacoGx platform

The lack of standardization of cell line and drug identifiers hinders comparison of molecular and pharmacological data between large-scale pharmacogenomic studies, such as the GDSC and CCLE. To address this issue we developed *PharmacoGx*, a computational platform enabling users to download and interrogate large pharmacogenomic datasets that were extensively curated to ensure maximum overlap and consistency[11]. *PharmacoGx* provides (*i*) a new object class, called *PharmacoSet*, that acts as a container for the high-throughput pharmacological and molecular data generated in large pharmacogenomics studies (detailed structure provided in Supplementary Methods); and (*ii*) a set of parallelized functions to assess the reproducibility of pharmacological and molecular data and to identify molecular features associated with drug effects. The *PharmacoGx* package is open-source and publicly available on Bioconductor.

The GDSC (formerly CGP) dataset

Drug sensitivity data. We used the data release 5 (June 2014) with 6,734 new IC_{50} values for a total of 79,903 drug dose-response curves for 139 different drugs tested on a panel of up to 672 unique cell lines. The data are accessible from ftp://ftp.sanger.ac.uk/pub4/cancerrxgene/releases/release-5.0/.

Molecular profiles. Gene expression data were downloaded from ArrayExpress, accession number E-MTAB-3610. These new data were generated using Affymetrix HG-U219 microarray platform. We processed and normalized the CEL files using RMA[33] with BrainArray[34] chip description file based on Ensembl gene identifiers (version 19). This resulted in a matrix of normalized expression for 17,616 unique Ensembl gene ids. SNP array data for the Genome-Wide Human SNP Array 6.0 platform were downloaded from GEO with the accession number GSE36139. We processed the raw CEL data using Affymetrix Power Tools (APT) v1.16.1. Copy number segments were generated using HAPSEG v1.1.1[35] based on RMA-normalized signal intensities and Birdseed v2-called genotypes. These segments were further refined using ABSOLUTE v1.0.6[36] to identify allele-specificity within each segment. Mutation and gene fusion calls were downloaded from the GDSC website and processed as in our initial study[7].

The CCLE dataset

Drug sensitivity data. We used the drug sensitivity data available from the CCLE website (https://portals.broadinstitute.org/ccle/data/browseData) and updated on February 2015 with a total number of 11,670 dose-response curves for 24 drugs tested in a panel of up to 504 cell lines.

Molecular profiles. Gene expression data were downloaded from the CCLE website and CGHub[37] for the Affymetrix HG-U133PLUS2 and Illumina HiSeq 2500 platforms, respectively. SNP array data were downloaded from EMBL-EBI with the accession number EGAD00010000644. Normalization of microarray data (1036 cell lines) and SNP array data (1190 cell lines) was performed the same way than for GDSC. RNA-seq data (935 cell lines) were downloaded as BAM files previously aligned using TopHat[38] and the quantification of gene expression was performed using Cufflinks[38] based on Ensembl GrCh37 human reference genome. Mutation data were retrieved from the CCLE website and processed as in our initial study[7].

Curation of drug and cell line identifiers

The lack of standardization for cell line names and drug identifiers represents a major barrier for performing comparative analyses of large pharmacogenomics studies, such as GDSC and CCLE. We therefore curated these datasets to maximize the overlap in cell lines and drugs by assigning a unique identifier to each cell line and drug. Entities with the same unique identifier were matched. Manual search was then applied to match any remaining cell lines or drugs which were not matched based on string similarity; annotations were consistently extracted from Cellosaurus[39]. The cell line curation was validated by ensuring that the cell lines with matched name had a similar SNP fingerprint (see below). The drug curation was validated by examining the extended fingerprint of each of their SMILES strings[40] and ensuring that the Tanimoto similarity[41] between two drugs called as the same, as determined by this fingerprint, was above 0.95.

Cell line identity using SNP fingerprinting

To assess the identity of cell lines from GDSC and CCLE, data of low quality were first excluded from our analysis panel (detailed procedure described in Supplementary Methods). Of the 973 CEL files from GDSC, only 66 fell below the 0.4 threshold (6.88%) for contrast QC scores, indicating issues in resolving base calls. Additionally, five of the 1,190 CEL files from CCLE had an absolute difference between contrast QC scores for Nsp and Sty fragments greater than 2, thus indicating some issues with the efficacy of one enzyme set during sample preparation. CEL files with contrast QC scores indicative of some sort of issue with the assay that would affect the genotype call rate or birdseed accuracy were removed and genotype calling was conducted on the remaining CEL files using Birdseed version 2. The resulting files were then filtered to keep only the 1006 SNP fingerprints that originated from CEL files that had a common cell line annotation between GDSC and CCLE (503 CEL files from each). Finally, pairwise concordances of all SNP fingerprints were generated according to the method outlined by Hong *et al.*[12].

Drug dose-response curves

To identify artefactual drug dose-response curves due to experimental or normalization issues, we developed simple quality controls (QC; details in Supplementary Methods). Briefly, we checked whether normalized viability measurements range between 0% and 100% and that drug-response curve is monotically non-increasing as expected. The drug dose-response curves which did not pass these simple QC were flagged and removed from subsequent analyses as the curve fitting would have yielded erroneous results.

All dose-response curves were fitted to the equation

$$y(x) = E_\infty + \frac{1 - E_\infty}{(1 + \left(\frac{x}{EC_{50}}\right)^{HS})}$$

where $y = 0$ denotes death of all infected cells, $y = y(0) = 1$ denotes no effect of the drug dose, EC_{inf} is the viability observed in the presence of an arbitrarily large concentration of drug, EC_{50} is the concentration at which viability is reduced by half as much as it is in the presence of an arbitrarily large concentration of drug, and HS is a parameter describing the cooperativity of binding. $HS < 1$ denotes negative binding cooperativity, $HS = 1$ denotes noncooperative binding, and $HS > 1$ denotes positive binding cooperativity. The parameters of the curves were fitted using the least squares optimization framework. Comparison of our dose-response curve model with those used in the GDSC and CCLE publications is provided in Supplementary Methods.

Discretization of pharmacogenomic data

Drug sensitivity data. To discretize the drug sensitivity data, we used AUC ≤ 0.2 (IC_{50} ≥ 1 μM) and AUC ≤ 0.4 (IC_{50} ≥ 10 μM) to identify the "insensitive" cell lines for targeted and cytotoxic drugs, respectively, while the rest of the cell lines are classified as "sensitive". These reasonable, although somewhat arbitrary, cutoffs enabled to explore the potential of such binary drug sensitivity calls as new drug phenotypic measures to find consistency in drug sensitivity data and gene-drug associations.

Gene expression data. To discretize the drug sensitivity data into lowly vs. highly expressed genes, we fit a mixture of two Gaussians of unequal variance using the full distribution of expression values of the 17,401 genes in common between GDSC and CCLE datasets. We defined the expression threshold as the expression value for which the posterior probability of belonging to the left tail of the highly expression distribution is 10%.

Mutation data. Similarly to the GDSC and CCLE publications, we transformed the original mutation data into binary values that represent the absence (0) or presence (1) of any missense mutations in a given gene in a given cell line.

Gene-drug associations

We assessed the association, across cell lines, between a molecular feature and response to a given drug, referred to as gene-drug association, using a linear regression model adjusted for tissue source:

$$Y = \beta_0 + \beta_i G_i + \beta_t T$$

where Y denotes the drug sensitivity variable, G_i and T denote the expression of gene i and the tissue source respectively, and βs are the regression coefficients. The strength of gene-drug association is quantified by β_i, above and beyond the relationship between drug sensitivity and tissue source. The variables Y and G are scaled (standard deviation equals to 1) to estimate standardized coefficients from the linear model. Significance of the gene-drug association is estimated by the statistical significance of β_i (two-sided t test). When applicable, p-values were corrected for multiple testing using the FDR approach[42].

As we recognized that continuous drug sensitivity is not normally distributed, which violates one of the assumption of the linear regression model described above, we also assessed the consistency of gene-drug association using discretized (binary) drug sensitivity calls as the response variable in a logistic regression model adjusted for tissue source, similarly to the linear regression model.

Measure of consistency

Area between curves (ABC). To quantify the difference between two dose-response curves, we computed the area between curves (ABC). ABC is calculated by taking the unsigned area between the two curves over the intersection of the concentration range tested in the two experiments of interest, and normalizing that area by the length of the intersection interval. In the present study, we compared the curves fitted for the same drug-cell line combinations tested both in GDSC and CCLE. Further details are provided in Supplementary Methods.

Pearson correlation coefficient (PCC). PCC is a measure of the linear correlation between two variables, giving a value between +1 and −1 inclusive, where 1 represents total positive correlation, 0 represents no correlation, and −1 represents total negative correlation[17]. PCC is sensitive to the presence of outliers, like a few sensitive cell lines in the case of drug sensitivity data measured for targeted therapies or genes rarely expressed.

Spearman rank correlation coefficient (SCC). SCC is a nonparametric measure of statistical dependence between two variables and is defined as the Pearson correlation coefficient between the ranked variables[18]. It assesses how well the relationship between two variables can be described using a monotonic function. If there are no repeated data values, a perfect Spearman correlation of +1 or −1 occurs when each of the variables is a perfect monotone function of the other. Contrary to PCC, SCC can capture non linear relationship between variables but is insensitive to outliers, which is frequent for drug sensitivity data measured for targeted therapies or genes rarely expressed.

Somers' Dxy rank correlation (DXY). DXY is a non-parametric measure of association equivalent to (C - 0.5) * 2 where C represents the concordance index[25] that is the probability that two variables will rank a random pair of samples the same way[19].

Matthews correlation coefficient (MCC). MCC[20] is used in machine learning as a measure of the quality of classification predictions. It takes into account true and false positives and negatives, acting as a balanced measure which can be used when the classes are of different sizes. MCC is in essence a correlation coefficient between two binary classifications; it returns a value between −1 (perfect opposite classification) and +1 (identical classifications), with 0 representing association no better than random chance.

Cramer's V (CRAMERV). CRAMERV is a measure of association between two nominal variables, based on Pearson's chi-squared statistic, giving a value between 0 (no association) and +1 (perfect association)[21]. In the case of 2×2 contingency table, such as binary drug sensitivity or gene expression measurements, CRAMERV is equivalent to the Phi coefficient.

Informedness (INFORM). For a 2×2 contingency table comparing two binary classifications, INFORM can be defined as Specificity + Sensitivity - 1, which is equivalent to true positive rate - false positive rate[22]. The magnitude of INFORM gives the probability of an informed decision between the two classes, where INFORM > 0 represents appropriate use of information, 0 represents chance-level decision, < 0 represents perverse use of information.

Data and software availability

Open Science Framework: Dataset: Revisiting inconsistency in large pharmacogenomics studies, doi 10.17605/osf.io/xxxx[43]

Data: The list of all the pharmacogenomic datasets available through the *PharmacoGx* platform can be obtained from R using the *availablePSets()* function from the R/Bioconductor library *PharmacoGx*.

The GDSC and CCLE *PharmacoSets* used in this study are available from pmgenomics.ca/bhklab/sites/default/files/downloads/ using the *downloadPset()* function.

Code: The R code necessary to replicate all the results presented in this article is available from the cdrug2 GitHub repository.

Author contributions

Z Safikhani, P Smirnov, M Freeman, R Quevedo, N El-Hachem, and A She were responsible for downloading and curating the pharmacogenomic data. Z Safikhani wrote most of the analysis code with contributions of P Smirnov, M Freeman, and R Quevedo. Z Safikhani, J Quackenbush and B Haibe-Kains designed the study. B Haibe-Kains supervised the study. All authors participated in the interpretation of the results. Z Safikhani, A Goldenberg, N Juul Birkbak, C Hatzis, L Shi, A Beck, H Aerts, J Quackenbush and B Haibe-Kains participated in the manuscript writing.

Competing interests

No competing interests were disclosed.

Grant information

Z Safikhani was supported by the Cancer Research Society (Canada; grant #19271) and the Ontario Institute for Cancer Research through funding provided by the Government of Ontario. P Smirnov was supported by the Canadian Cancer Society Research Institute. C Hatzis was supported by Yale University. N Juul Birkbak was funded by The Villum Kann Rasmussen Foundation. L Shi was supported by the National High Technology Research and Development Program of China (2015AA020104), the National Natural Science Foundation of China (31471239), the 111 Project (B13016), and the National Supercomputer Center in Guangzhou, China. J Quackenbush was supported by grants from the NCI GAME-ON Cancer Post-GWAS initiative (5U19 CA148065) and the NHLBI (5R01HL111759). B Haibe-Kains was supported by the Canadian Institutes of Health Research, Cancer Research Society, Terry Fox Research Institute, and the Gattuso Slaight Personalized Cancer Medicine Fund at Princess Margaret Cancer Centre.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgements

The authors would like to thank the investigators of the Genomics of Drug Sensitivity in Cancer (GDSC) and the Cancer Cell Line Encyclopedia (CCLE) who have made their valuable data available to the scientific community. We thank Michael T. Hallet, Paul T. Spellman and David G. Covell for their constructive review of the initial version of the manuscript. We thank the MAQC/SEQC consortium and the scientific community for their constructive feedback.

Supplementary material

Supplementary file 1.

All the noisy curves identified in GDSC and CCLE.

Click here to access the data.

Supplementary file 2.

All drug dose-response curves in common between GDSC and CCLE.

Click here to access the data.

Supplementary file 3.

All drug dose-response curves for replicated experiments using AZD6482 in GDSC.

Click here to access the data.

Supplementary methods.

Click here to access the data.

Supplementary figures and tables.

Click here to access the data.

Supplementary Datasets

Dataset 1. SNP fingerprints of all the cell lines profiled with SNP arrays in GDSC and CCLE. Data used to generate Figure 3.

Click here to access the data.

Dataset 2. AUC and IC_{50} values as published and recomputed using PharmacoGx. Data used to generate Figure 6.

Click here to access the data.

Dataset 3. Consistency measures for AUC, AUC* (STAR) and IC_{50} values as published and recomputed using PharmacoGx, across cell lines. Data used to generate Figure 8.

Click here to access the data.

Dataset 4. Consistency measures for molecular profiles across cell lines. GE.CCLE.ARRAY.RNASEQ: Consistency between gene expression data generated using Affymetrix HG-U133PLUS2 microarray and Illumina RNA-seq platforms within CCLE; GE.ARRAYS: Consistency between gene expression data generated using Affymetrix HG-U133A and HG-U133PLUS2 microarray platforms in GDSC and CCLE, respectively; GE.ARRAY.RNASEQ: Consistency between gene expression data generated using Affymetrix HG-U133A microarray and Illumina RNA-seq platforms in GDSC and CCLE, respectively; CNV: Consistency of copy number variation data in CCLE and GDSC, respectively; MUTATION: Consistency of mutation profiles in CCLE and GDSC, respectively. Data used to generate Figure 12.

Click here to access the data.

Dataset 5. Spreadsheets reporting the statistics (effect size and significance) for all expression-based, mutation-based and amplification-based gene-drug associations for each drug using the common cell lines screened both in GDSC and CCLE. Gene-drug associations were estimated the relevant molecular profile of genes as input and continuous published AUC as output in a linear regression model adjusted for tissue source. Data used to generate Supplementary Table 1 and Supplementary Figure 13.

Click here to access the data.

Dataset 6. Spreadsheets reporting the statistics (effect size and significance) for all expression-based, mutation-based and amplification-based gene-drug associations for each drug using the entire panel of cell lines in GDSC and CCLE. Gene-drug associations were estimated using the relevant molecular profile of genes as input and continuous published AUC as output in a linear regression model adjusted for tissue source. Data used to generate Table 1, Figure 10, Supplementary Table 2 and Supplementary Figure 14.

Click here to access the data.

References

1. Garnett MJ, Edelman EJ, Heidorn SJ, *et al.*: **Systematic identification of genomic markers of drug sensitivity in cancer cells.** *Nature.* 2012; **483**(7391): 570–5.
PubMed Abstract | Publisher Full Text | Free Full Text

2. Barretina J, Caponigro G, Stransky N, *et al.*: **The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity.** *Nature.* 2012; **483**(7391): 603–7.
PubMed Abstract | Publisher Full Text | Free Full Text

3. Papillon-Cavanagh S, De Jay N, Hachem N, *et al.*: **Comparison and validation of genomic predictors for anticancer drug sensitivity.** *J Am Med Inform Assoc.* 2013; **20**(4): 597–602.
PubMed Abstract | Publisher Full Text | Free Full Text

4. Dong Z, Zhang N, Li C, *et al.*: **Anticancer drug sensitivity prediction in cell lines from baseline gene expression through recursive feature selection.** *BMC Cancer.* 2015; **15**: 489.
PubMed Abstract | Publisher Full Text | Free Full Text

5. Jang IS, Neto EC, Guinney J, *et al.*: **Systematic assessment of analytical methods for drug sensitivity prediction from cancer cell line data.** *Pac Symp Biocomput.* 2014; 63–74.
PubMed Abstract | Publisher Full Text | Free Full Text

6. Cortés-Ciriano I, van Westen GJ, Bouvier G, *et al.*: **Improved large-scale prediction of growth inhibition patterns using the NCI60 cancer cell line panel.** *Bioinformatics.* 2016; **32**(1): 85–95.
PubMed Abstract | Publisher Full Text | Free Full Text

7. Haibe-Kains B, El-Hachem N, Birkbak NJ, *et al.*: **Inconsistency in large pharmacogenomic studies.** *Nature.* 2013; **504**(7480): 389–93.
PubMed Abstract | Publisher Full Text | Free Full Text

8. Hatzis C, Bedard PL, Birkbak NJ, *et al.*: **Enhancing Reproducibility in Cancer Drug Screening: How Do We Move Forward?** *Cancer Res.* 2014; **74**(15): 4016–23.
PubMed Abstract | Publisher Full Text | Free Full Text

9. Safikhani Z, El-Hachem N, Quevedo R, *et al.*: **Assessment of pharmacogenomic agreement [version 1; referees: 3 approved].** *F1000 Res.* 2016; **5**: 825.
PubMed Abstract | Publisher Full Text | Free Full Text

10. Haverty PM, Lin E, Tan J, *et al.*: **Reproducible pharmacogenomic profiling of cancer cell line panels.** *Nature.* 2016; **533**(7603): 333–7.
PubMed Abstract | Publisher Full Text

11. Smirnov P, Safikhani Z, El-Hachem N, *et al.*: **PharmacoGx: an R package for analysis of large pharmacogenomic datasets.** *Bioinformatics.* 2016; **32**(8): 1244–6.
PubMed Abstract | Publisher Full Text

12. Hong H, Xu L, Liu J, *et al.*: **Technical reproducibility of genotyping SNP arrays used in genome-wide association studies.** *PLoS One.* 2012; **7**(9): e44483.
PubMed Abstract | Publisher Full Text | Free Full Text

13. Yu M, Selvaraj SK, Liang-Chu MM, *et al.*: **A resource for cell line authentication, annotation and quality control.** *Nature.* 2015; **520**(7547): 307–11.
PubMed Abstract | Publisher Full Text

14. Goodspeed A, Heiser LM, Gray JW, *et al.*: **Tumor-derived Cell Lines as Molecular Models of Cancer Pharmacogenomics.** *Mol Cancer Res.* 2016; **14**(1): 3–13.
PubMed Abstract | Publisher Full Text | Free Full Text

15. Cancer Cell Line Encyclopedia Consortium, Genomics of Drug Sensitivity in Cancer Consortium: **Pharmacogenomic agreement between two cancer cell line data sets.** *Nature.* 2015; **528**(7580): 84–7.
PubMed Abstract | Publisher Full Text

16. Youden WJ: **Index for rating diagnostic tests.** *Cancer.* 1950; **3**(1): 32–5.
PubMed Abstract | Publisher Full Text

17. Pearson K: **Note on Regression and Inheritance in the Case of Two Parents.** *Proc R Soc Lond.* 1895; **58**: 240–2.
Publisher Full Text

18. Spearman C: **The proof and measurement of association between two things. By C. Spearman, 1904.** *Am J Psychol.* 1987; **100**(3–4): 441–71.
PubMed Abstract

19. Somers RH: **A New Asymmetric Measure of Association for Ordinal Variables.** *Am Sociol Rev.* 1962; **27**(6): 799–811.
Publisher Full Text

20. Matthews BW: **Comparison of the predicted and observed secondary structure of T4 phage lysozyme.** *Biochim Biophys Acta.* 1975; **405**(2): 442–51.
PubMed Abstract | Publisher Full Text

21. Cramér H: **Mathematical Methods of Statistics.** (Princeton: Princeton University Press, 1946). CramérMathematical Methods of Statistics 1946.
Reference Source

22. Powers DM: **Evaluation: from Precision, Recall and F-measure to ROC, Informedness, Markedness and Correlation.** 2011.
Reference Source

23. Hudson AM, Yates T, Li Y, *et al.*: **Discrepancies in cancer genomic sequencing highlight opportunities for driver mutation discovery.** *Cancer Res.* 2014; **74**(22): 6390–6.
PubMed Abstract | Publisher Full Text | Free Full Text

24. Jaccard P: **Etude comparative de la distribution florale dans une portion des Alpes et du Jura.** Impr Corbaz; 1901; **37**: 547–579.
Publisher Full Text

25. Harrell FE Jr, Califf RM, Pryor DB, *et al.*: **Evaluating the yield of medical tests.** *JAMA.* 1982; **247**(18): 2543–6.
PubMed Abstract | Publisher Full Text

26. MAQC Consortium, Shi L, Reid LH, *et al.*: **The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements.** *Nat Biotechnol.* 2006; **24**(9): 1151–61.
PubMed Abstract | Publisher Full Text | Free Full Text

27. Shi L, Campbell G, Jones WD, *et al.*: **The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models.** *Nat Biotechnol.* 2010; **28**(8): 827–38.
PubMed Abstract | Publisher Full Text | Free Full Text

28. Pozdeyev N, Yoo M, Mackie R, *et al.*: **Integrating heterogeneous drug sensitivity data from cancer pharmacogenomic studies.** *Oncotarget.* 2016; **7**(32): 51619–51625.
PubMed Abstract | Publisher Full Text | Free Full Text

29. Safikhani Z, Thu KL, Silvester J, *et al.*: **Gene isoforms as expression-based biomarkers predictive of drug response *in vitro*.**
Publisher Full Text

30. Daemen A, Griffith OL, Heiser LM, *et al.*: **Modeling precision treatment of breast**

cancer. *Genome Biol.* 2013; **14**(10): R110. Erratum in: ***Genome Biol*. 2015;16:95**.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

31. Safikhani Z, El-Hachem N, Smirnov P, *et al.*: **Safikhani *et al*. reply.** *Nature.* 2016; **540**(7631): E2–E4.
PubMed Abstract | **Publisher Full Text**

32. Geeleher P, Gamazon ER, Seoighe C, *et al.*: **Consistency in large pharmacogenomic studies.** *Nature.* 2016; **540**(7631): E1–E2.
PubMed Abstract | **Publisher Full Text**

33. Irizarry RA, Hobbs B, Collin F, *et al.*: **Exploration, normalization, and summaries of high density oligonucleotide array probe level data.** *Biostatistics.* 2003; **4**(2): 249–64.
PubMed Abstract | **Publisher Full Text**

34. de Leeuw WC, Rauwerda H, Jonker MJ, *et al.*: **Salvaging Affymetrix probes after probe-level re-annotation.** *BMC Res Notes.* 2008; **1**: 66.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

35. Carter SL, Meyerson M, Getz G: **Accurate estimation of homologue-specific DNA concentration-ratios in cancer samples allows long-range haplotyping.** Scott L Carter, 2011; 59.
Reference Source

36. Carter SL, Cibulskis K, Helman E, *et al.*: **Absolute quantification of somatic DNA alterations in human cancer.** *Nat Biotechnol.* 2012; **30**(5): 413–21.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

37. Wilks C, Cline MS, Weiler E, *et al.*: **The Cancer Genomics Hub (CGHub): overcoming cancer through the power of torrential data.** *Database (Oxford).* 2014; **2014**: pii: bau093.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

38. Trapnell C, Roberts A, Goff L, *et al.*: **Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks.** *Nat Protoc.* 2012; **7**(3): 562–78.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

39. Bairoch A: **ExPASy - Cellosaurus [Internet].** *Cellosaurus.* 2015; [cited 2016 Jan 26].
Reference Source

40. Anderson E, Veith GD, Weininger D: **SMILES, a Line Notation and Computerized Interpreter for Chemical Structures.** 1987.
Reference Source

41. Tanimoto TT: **An Elementary Mathematical Theory of Classification and Prediction.** International Business Machines Corporation; 1958; (Internal Technical Report).
Reference Source

42. Benjamini Y, Hochberg Y: **Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.** *J R Stat Soc Series B Stat Methodol.* 1995; **57**(1): 289–300.
Reference Source

43. Safikhani Z, Smirnov P, Freeman M, *et al.*: **Dataset: Revisiting inconsistency in large pharmacogenomics studies.** *Open Science Framework.* 2016.
Data Source

Open Peer Review

Current Referee Status: ✓ ? ✓

Version 3

Referee Report 21 August 2017

doi:10.5256/f1000research.13399.r24961

✓

Michael T. Hallett
Centre for Structural and Functional Genomics, Department of Biology, Concordia University, Montréal, QC, Canada

My concerns have been addressed, and the quality of the presentation is now more than sufficient to understand the details of their methodology.

It is an important article, and the findings here contribute to (hopefully) a larger, quantitative discussion regarding pharmacogenomic studies. Differences (both technical and biological in nature) in methodologies between studies cause differences in results (eg which compounds are identified as affective). The methodological differences might be important: although they can "deflate" reproducibility, they also serve to probe a larger search space (that is, each assay may be exploring a subtly different part of a huge space of compounds with specific bioactivity).

Regardless, it is important to have the tools to have a quantitative discussion about whether two assays disagree due to "nuisances", or whether they disagree because they are fundamentally testing different hypothesis. This manuscript, and the software developed here, provide the community with important components to facilitate such discussions.

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 2

Referee Report 26 July 2017

doi:10.5256/f1000research.13143.r24479

?

Paul T. Spellman
Department of Molecular and Medical Genetics, Oregon Health and Science University, Portland, OR, USA

"Our finding that molecular profiles are significantly more consistent than drug sensitivity data, indicates that the main barrier to biomarker development using these data is the unreliability in the reported response phenotypes for many drugs. For studies such as these to realize their full potential, additional work must be done to develop robust and reproducible experimental and analytical protocols so that the same compound, tested on the same set of cell lines by different groups, yields consistent and comparable results. Barring this, a predictive biomarker of response developed from one study is unlikely to be able to reliably validated on another, and consequently, is unlikely to be useful in predicting patient response."

Again, I think the concern here is wrong for the same reasons I describe above. The main barrier is *not* the differences in quality between assays. The author's response to my initial concern shifts it to variability within an assay, but that is obvious -- if an assay doesn't work it can't be informative. Bad assays should be excluded. Period. The original premise of the work was that differences in results between assays was problematic but that is not the problem. Its that specific assays do not produce the high quality data necessary.

My further points remain. I agree it may be hard, but it is possible.

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reader Comment 07 Aug 2017

Benjamin Haibe-Kains,

We have now updated the discussion and conclusion to reflect the fact that the noise within assays must be assessed and accounted for, and the complementarity across assays offer new opportunities to develop more robust biomarkers. Updated parts are underlined and in italic.

“Our finding that molecular profiles are significantly more consistent than drug sensitivity data, indicates that the main barrier to biomarker development using these data is the discrepancy in the reported response phenotypes for many drugs. *The experimental protocols and pharmacological assays used in the GDSC and CCLE studies are the state-of-the-art for high-throughput drug screening projects. Even though technical and biological replicates are necessary to assess and account for noise in drug sensitivity measurements, it is clear that the assays used in GDSC and CCLE probe different aspects of the biology underlying drug-induced growth inhibition. Without knowing which assay is more relevant for in vivo drug response, more research will be required to best leverage these complementary assays for robust biomarker discovery.*

From having worked in large-scale genomic analyses, we recognize the challenges involved in planning and executing such studies and commend the GDSC and CCLE for their work and for making all the data available. However, we strongly encourage the GDSC, the CCLE, the pharmacogenomics and bioinformatics communities as a whole, t*o invest the necessary time and effort to account for the noise in drug response measurements and the complementary nature of different assays in order to assure that these studies are relevant for predicting response in patients.* The recent report from Genentech is a significant step in this direction. Ultimately, that effort will help to assure that mammoth undertakings in drug characterization can deliver on their promise to identify better therapies and biomarkers predictive of response.”

We agree with the reviewer regarding the reviewer's comments on drug-specific cutoffs, and have updated the paragraph about the limitations of our study to reflect this important point.

Competing Interests: None

Version 1

Referee Report 10 May 2017

doi:10.5256/f1000research.10354.r22599

✓ **David G. Covell**
Screening Technologies Branch, Developmental Therapeutics Program, National Cancer Institute, Frederick, MD, USA

The paper under review, 'Revisiting inconsistency in large pharmacogenomics studies' by Zhaleh Safikhani, Petr Smirnov, Mark Freeman, Nehme El-Hachem, Adrian She, Quevedo Rene, Anna Goldenberg, Nicolai J. Birkbak, Christos Hatzis, Leming Shi, Andrew H. Beck, Hugo J.W.L. Aerts, John Quackenbush, Benjamin Haibe-Kains, reports an updated analysis of results from two previously published systematic drug screening projects[1,2]. As explained in their introductory material, this report is motivated in part by the expansion of data from these earlier studies, and as a means to document alternative data analysis strategies that have been proposed for improving the original publication[3].

The authors address two highly important areas in basic and clinical research: data reproducibility and predictive (gene expression) biomarkers based on drug sensitivity data. The former issue represents a hallmark of basic science research; where results derived from different labs and measurement techniques serve to establish strong confidence in a proposed experimental protocol. The latter issue pertains best to highly confident (e.g. reproducible or consistent) experimental measurements; while data inconsistencies foreshadow a Pandora’s Box of alternatives in the search for the origins of these differences.

With respect to the issue of data reproducibility, I find no fault in the new manuscript. All of the results reported in the original 2013 paper and current paper under review can be obtained from their Supplementary R code. With a bit of diligence and tenacity, sequentially stepping through their R-code will yield the reported figures and tables. Towards that end, the original R-code is tedious, but their addition of an open-source R-package, PharmacoGx, relieves much of the tedium. In fact, the authors must be applauded for making their analysis completely reproducible, a feat rarely achieved with biological results.

Notwithstanding, the results remain largely the same; inconsistencies remain in the drug sensitivity profiles between the GCSC[2] and CCLE[1] groups. Data analysis based on alternative methods appear to be constructed around the arguments proposed in the pair of Brief Communications Arising from the original paper[4,5]. The general idea of this alternative data analysis is based on the limited role of weakly responsive tumor cells, and thus a failure to contribute to meaningful statistics. While segregating the data into three classes (drugs with no observed tumor cell activity, activity in a few tumor cells and activity in a large number of tumor cells) improves the statistics, the differences largely remain.

Speculations about the differences between the two datasets focus, naturally, on each measurement platform. The current manuscript's proposal of internal standardization may help identify the origin(s) of these differences, but this alone may not be sufficient. In this regard, I would recommend the Supplementary Information (all 47 pages) from the original article[3]. Specifically, Section 3, Comparison of experimental protocols, and the included Comparative table. The details of this section identify a number of platform differences that may underlie their measurement differences. Although not within the scope of the current article, a future study focused on these differences, combined with standardization, rather than looking for answers by segregating the data into three response classes, would be highly informative.

An alternative speculation regarding data inconsistencies considers the 'dated' possibility that each tumor cell's drug sensitivity and underlying phenotypic architecture (expression, mutation, snp, etc.) exemplifies a 'snowflake' phenomenon. Each tumor cell represents a unique circumstance, which can be modulated by any sort of environmental condition. Thus a drug response, even for the same tumor cell, may exhibit variation. Under these circumstances, the functional pathways, represented by groups of genes and their concordant expressions, become the focus, and derivation of pathway-based scoring schemes may significantly overcome inconsistencies between experimental groups. Clearly an appropriate pathway fitness scoring scheme has yet to be devised.

In summary, the analysis is sound, the results are clear, and the analyses of inconsistent data, as a means to obtain predictive biomarkers, remains a significant challenge.

References

1. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, Wilson CJ, Lehár J, Kryukov GV, Sonkin D, Reddy A, Liu M, Murray L, Berger MF, Monahan JE, Morais P, Meltzer J, Korejwa A, Jané-Valbuena J, Mapa FA, Thibault J, Bric-Furlong E, Raman P, Shipway A, Engels IH, Cheng J, Yu GK, Yu J, Aspesi P, de Silva M, Jagtap K, Jones MD, Wang L, Hatton C, Palescandolo E, Gupta S, Mahan S, Sougnez C, Onofrio RC, Liefeld T, MacConaill L, Winckler W, Reich M, Li N, Mesirov JP, Gabriel SB, Getz G, Ardlie K, Chan V, Myer VE, Weber BL, Porter J, Warmuth M, Finan P, Harris JL, Meyerson M, Golub TR, Morrissey MP, Sellers WR, Schlegel R, Garraway LA: The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity.*Nature*. 2012; **483** (7391): 603-7 PubMed Abstract | Publisher Full Text

2. Garnett MJ, Edelman EJ, Heidorn SJ, Greenman CD, Dastur A, Lau KW, Greninger P, Thompson IR, Luo X, Soares J, Liu Q, Iorio F, Surdez D, Chen L, Milano RJ, Bignell GR, Tam AT, Davies H, Stevenson JA, Barthorpe S, Lutz SR, Kogera F, Lawrence K, McLaren-Douglas A, Mitropoulos X, Mironenko T, Thi H, Richardson L, Zhou W, Jewitt F, Zhang T, O'Brien P, Boisvert JL, Price S, Hur W, Yang W, Deng X, Butler A, Choi HG, Chang JW, Baselga J, Stamenkovic I, Engelman JA, Sharma SV, Delattre O, Saez-Rodriguez J, Gray NS, Settleman J, Futreal PA, Haber DA, Stratton MR, Ramaswamy S, McDermott U, Benes CH: Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature*. 2012; **483** (7391): 570-5 PubMed Abstract | Publisher Full Text

3. Haibe-Kains B, El-Hachem N, Birkbak NJ, Jin AC, Beck AH, Aerts HJ, Quackenbush J: Inconsistency in large pharmacogenomic studies.*Nature*. 2013; **504** (7480): 389-93 PubMed Abstract | Publisher Full Text

4. Bouhaddou M, DiStefano MS, Riesel EA, Carrasco E, Holzapfel HY, Jones DC, Smith GR, Stern AD, Somani SS, Thompson TV, Birtwistle MR: Drug response consistency in CCLE and CGP.*Nature*. 2016; **540** (7631): E9-E10 Publisher Full Text

5. Safikhani Z, El-Hachem N, Smirnov P, Freeman M, Goldenberg A, Birkbak NJ, Beck AH, Aerts HJ,

Quackenbush J, Haibe-Kains B: Safikhani et al. reply.*Nature*. 2016; **540** (7631): E2-E4 Publisher Full Text

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 07 Jul 2017

Benjamin Haibe-Kains,

We thank Dr Covell for his constructive comments regarding our study. We are glad to hear that our PharmacoGx package is useful to reproduce our analysis results. The hope is that our package will enable other research groups to analyze and compare their own data with published large-scale pharmacogenomic datasets.

We agree with the reviewer that more investigation is required to better assess the technical vs biological variations for each of the pharmacological assays. Then biological variations could be leveraged, at the pathway level as the reviewer suggested, to define more robust biomarkers.

Competing Interests: None

Referee Report 21 December 2016

doi:10.5256/f1000research.10354.r17399

? **Paul T. Spellman**

Department of Molecular and Medical Genetics, Oregon Health and Science University, Portland, OR, USA

Safikhani *et al.* have updated their previous analysis of two of the largest systematic drug screening projects linked to genomics data. The previous findings indicated that there is a lack of concordance between the two datasets that makes finding biomarkers of response difficult. Updating these studies with a wider array of methods in response to comments about the original article leaves largely the same result. Drug sensitivity profiles show significant variation between the two groups, likely due to differences in assay condition.

Safikhani *et al.* follow this analysis up with a discussion on how to improve the situation and here I have some significant issues. The base argument is that the differences in platform are creating biases in the results and therefore the platforms need to be standardized. I think this is completely wrong. This makes sense if one platform were known to recapitulate *in vivo* response more accurately, but that is not true. We do not know if one platform is more physiologically relevant than another so the lack of standardization actually tells you something, it tells you when a predictor result is robust against biological context and is therefore more likely to work in new biological contexts. I would argue we need *more* variability in assays and platforms to broaden the scope of biological systems, not less.

Similarly, the statement is made that there is a 75% validation rate for eight drugs but that "this does not suggest that one can use these studies to find new, reproducible gene-drug associations...", I actually think it does, but perhaps I am missing a subtlety.

Finally, I think it is possible to set drug specific thresholds for each dataset. We have done this, I believe successfully, with datasets far smaller.

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 12 Jul 2017

Benjamin Haibe-Kains,

We thank the reviewer for his constructive comments. We agree with the reviewer that we need to update the discussion to reflect this important point. In the absence of “gold standard” screening platform, the best biomarkers are likely those that are robust to the use of different assays, as these assays assess different biological aspects of growth inhibition. However one must clearly distinguish between technical and biological variations. While biological variations might be interesting for biomarker discovery, assay variation must be kept as low as possible. Looking at the replicates performed for AZD6482 in GDSC, we found that drug sensitivity data lack consistency even when the same assay is used (see new Supplementary Figure 2E). In this setting, one cannot claim that the inconsistencies observed between GDSC and CCLE are solely due to differences in the type of assay used for drug screening. Even if we agree with the reviewer, we believe there is still work to be done to improve the robustness of each pharmacological assay. We have updated the discussion section of our manuscript accordingly.

Similarly, the statement is made that there is a 75% validation rate for eight drugs but that "this does not suggest that one can use these studies to find new, reproducible gene-drug associations...", I actually think it does, but perhaps I am missing a subtlety.

Although our new analysis revealed reasonable consistency for biomarker discovery for 8 out of 15 drugs, we could not get such a validation rate for the rest of the drugs where the biomarkers are only significant in one study but not the other.

Finally, I think it is possible to set drug specific thresholds for each dataset. We have done this, I believe successfully, with datasets far smaller.

Like this reviewer, we too tried to identify drug-specific cutoff that would allow us to binarize the drug sensitivity while optimising the consistency across datasets. Our best efforts were not successful though, except for Nilotinib (see Safikhani et al, Nature 2016; PMID: 27905430). Not to say that it is not feasible but we found it very challenging to define a cutoff within a dataset that would yield good concordance across datasets.

Competing Interests: None

Referee Report 03 November 2016

doi:10.5256/f1000research.10354.r16370

Michael T. Hallett
Centre for Structural and Functional Genomics, Department of Biology, Concordia University, Montréal, QC, Canada

This manuscript seeks to compare two large pharmacogenomics datasets (several hundred cancer cell lines screened against 15 common drugs) and evaluate their level of agreement via (1) the drug sensitivity values and (2) gene expression profiles of the cell lines. Broadly speaking, the value of these profiles is the discovery of (gene) biomarkers that could predict response of cells to the drugs. Previous efforts, including the authors' previous attempts, have had trouble with reproducibility. The authors have previously given harsh critiques regarding the reproducibility of the two datasets.

This manuscript is very important, and it has the potential to dissect sources of both agreement and disagreement that can be amplified or minimized in the future respectively. The reviewer also has little doubt that there is in fact disagreement between these two datasets and, moreover, it is significant enough as to interfere with the discovery of biomarkers. The reviewer also agrees with the authors that this is important to point out and understand, and the "call to arms" in the Discussion (the best written part of the manuscript) should certainly be listened to.

However, because this manuscript is very important , the cornerstones of the comparative analysis must be correct. The Supplemental Methods are near impossible to decipher and are littered with undefined terms, confusion in mathematical notation, poor equation formatting non-intuitive statements that do not assist the reader with understanding the numerous design choices (from throwing out poor quality data, to model fitting, to different measures of consistency). The Results section is not sufficient methodical to follow the argument of what does or does not represent ***statistically significant*** disagreement. Almost every paragraph until the conclusion presented serious challenges to this referee. They are included below.

This is an important effort and the authors should return with an improved manuscript. Many of the co-authors are skilled mathematicians and they are strongly recommended to revisit every line of this manuscript to ensure correctness and to present with craftmanship. This is especially true in the Supplementary Methods that actually provide the "meat" of the methodology: this I believe must have been an oversight with this submission.

This manuscript needs to be published and I believe there are important lessons to be learnt here but there has to be a more focused, tighter argument to establish where there is disagreement and hypotheses as to why (in the Discussion) and what can be done about it. But the main issue here is that the basics of the paper are not solid, or at least they cannot be evaluated. The authors should be commended for the effort to be reproducible (in the sense they give the list of R packages used and their code) but that is only one aspect. The mathematics and statistics requires clarity and correctness. Terms such as "metric" should be used properly, and novel equations that are derived (e.g their "E" parameter) must be done so in a careful correct manner, with attempts made to justify these parameters (e.g \epsilon, \rho, 2*\epsilon, the E parameter from the modified fit etc.

I would be happy to view a revised version of the manuscript and I hope that my comments aid in this important project.

pg5: What is “Dataset 1”? The link here doesn’t lead anywhere that I can tell.

Figure 3. I’m not really sure what the value is in plotting the density functions for the mismatched and matched cell lines. First, wouldn’t one density function suffice with a threshold I guess? Second, do you really need it at all?

In the Methods “Cell line identity….”, it is stated that 66 samples fell below threshold with a reference to the Supplementary methods. However I don’t see anything in the supplementary methods that discusses this. Moreover in the text, it seems that you threw 8 cases away. This is confusing.

pg 5. I’m not sure what you mean by “remain stable or decrease monotonically”?
Do you just mean “monotonically non-increasing”?

Please see comment regarding “Filtering of drug dose-response curves” form Supp Methods below. I think this really needs to be reworked, and I have to trust you guys here that you are doing the right thing.

“as exemplified…” depicted?

In Figure 4, is it possible to relate this back to the choice of \rho, \epsilon and 2 \cdot \epsilon from the Supp Methods, or perhaps integration a version of this figure (but annotated) into the Supp Methods.
In panel A, the grey area is a bit non-intuitive no? I would say that post 0.03, it’s looking pretty good, and it’s not measured in GDSC after that. However the first points are off.

I don’t know what \epsilon or \rho are so it’s hard to relate what is depicted in Figure 4 back to your model.

When I look through Supplemental Figure 1 (all the excluded comparisons), it seems like your criterion for excluding a comparison boils down cases where at least one of the curves has high variance, and the cutoff 2\epsilon I think is a constant independent of the distribution of points for either curve. I don’t see in your equations how you encode that the sequence is monotonically non-increasing, or how “order” along the left to right sweep is incorporated.

Supp Methods Filtering of drug dose-response curves

i +1^{st} -> (i+1)^{st}

It seems like there is a formatting problem here. In my pdf there is something like I”A squiggle after “in some large fraction ??? of the cases (1).” I guess that’s supposed to be \rho right?

I’m not sure I understand this sentence “Our quality control …” Are you saying that \Delta_{i, i+1} < \epsilon in some fraction \phi of the cases?

equation #2 below:
I also have never seen set notation such as \{ \Delta_{i,i+1} | \Delta_{i,i+1} < \epsilon \}.
Are you trying to say “given all the \Deltas that are less than \epsilon?
So the vertical | means cardinality here right?
But then the denominator has the cardinality of a value, or do you mean absolute value?
What is \rho? It’s undefined.

“Unfortunately …” The english is a bit rough - could be rephrased in terms of specificity and sensitive, I guess.

I don’t understand the significance of the sentence “Consider, for instance, …” But in the main text, you say that it remains stable or decreases monotonically. Here it increases monotonically for many successive points, so this violates your model, no?

I think that this subsection wants simply to spell out mathematically the thresholds and also provide some rationale for the parameters. I think the text doesn’t really do a good job of establishing this rationale and needs work. Perhaps define the parameters precisely and then phrase the exposition in standard terms e.g. specific and sensitivity for different \epislon, etc.

Your equation (2) is inconsistent. In the text you specify \Sigma_{\forall i,j} \Delta_{i,j} but below your criterion seems to change to (the correct) $i < j$. It is also sufficient to write \Sigma{ i < j } \Delta_{i,j} and avoid the double summation.

You should probably define D_i in the text and not make the reader deduce it from the figure below

The comments here w.r.t. the Supplementary Methods also apply to the associated subsection of the Methods. The mathematical correctness of some comments needs attention. For example, it is not quite correct to say that the “curve fitting would have yielded erroneous results”. The curve fitting is just that, curve fitting. It’s not really an error. Then in the Methods, you claim to use this equation but this is inconsistent with the discussion in the Supplemental Methods (where you have an equation with this undefined parameter E).

The least-squares method using a three-parameter sigmoid model. I understand the intuition for this, but when I look through Supplementary File 2, I think there are a lot cases where this is perhaps not the correct pattern to assume (e.g. straight lines). Moreover, there are some very strange fits, for example,
AZD6244:G−361
AZD6244:SK−MM−2

PLX4720:MDA–MB–175–VII
In some cases the curve is always above all of the observations.
Perhaps this is because the measurements of viability > 100%? In your model you have removed the \$E_0\$ from Barretina *et al.* for different reasons.

Supp Methods - Fitting of drug dose-response curves

I'm a bit lost with your choice of notation here. For example, you define y as an equation, not a function, and then write y(0) which I assume is supposed to be something like y(x). Ok but then you have y=0 and y = y(0) = 1. I'm not sure this is correct mathematically. (I think you mean to say that y(x) = … *where* y(0) =1.

"viability is reduced to half … concentration of the drug"… so the "Top", E_\infinity … I find this a bit wordy.

"The dose response equation now becomes …"

So I deduce that E is the new parameter?!?! Where has E_\infinity gone?!?!

(But then down below E is constrained to be in [0,1] and seems to related to the fitness of neoplastic cells. I'm not sure I understand this.)

There is no derivation of this formula whatsoever. In fact, I don't see how this could be correct any longer. Couldn't this be expressed as mixture of two cell types, and y would be then a sort of weighted average?

I really don't see how this was derived. This is a very central part of your paper (since the manuscript is measuring agreement) and therefore it needs to be bulletproof.

Please define "extant drug". Also HS is allowed to vary apparently but I don't see where it is then optimized in your analysis later on. This is confusing.

Consistency of Drug Sensitivity Data

Is the ABC method standard? Are there citations for this? You should probably define properly what you mean by the "insertion of the concentration range". Elsewhere it seems that you are referring to this as "common concentration range" e.g. SFig 2
More generally, isn't this a sort of (non-statistical) version of the Kolmogorov-Smirnoff test?

Figure 5A: Actually there are many such cases in your Supplementary Figures. Doesn't this just mean that the range of concentrations are not sufficient in both datasets?

pg 7 "We then computed the median …" I don't understand what your distance metric is here. If I understand correctly, you could the ABC for each pair of drugs in the GDSC dataset. From that I can imagine a distance matrix D where D_ij i the ABC between drugs i and j. But you said you take the median ABC? median over what? cell lines? repeats? Whatever the case, are you sure it's a distance metric?! Is it really true that distances derived from ABCs are metrics? I think this should be shown in the Supplemental Methods. Also as a minor comment, the caption in Supp Figure 3 says that you are using the mean ABC value but elsewhere it says median.

p8. I am not sure what the significance of Supplementary Figure 3 is: are any of these clusters significant? Why are two drugs coloured red in panel A? On page 8 it is claimed that the samples split by hospital (MGH vs WTSI) but I don't see how this is represented in Supplemental Figure 3.

pg 8. I have a very hard time estimating the significance of a statement like "…3 out of 15 common drugs clustered tightly". I am not sure what tight means here.
When I look at Supp Figure 3 there has been no effort to annotate the clusters with their reproducibility e.g. pvclust or measure their significance in some other way.
When I look at the figure I think there appears to be a lot of co-clustering of the drugs, at least given that the median ABC across a diverse collection of cell lines might not be such a great "distance" measure.

pg 9. What do you mean by "highly targeted therapies"?

pg 9. paragraph "Although the ABC values …" I think the ABC is interesting but it takes a very prominent role in your paper when there are other standard techniques already like AUC and IC_50. Perhaps the manuscript should have comments about why have chosen this approach that is not standard. Also I think you would need to make precise what the differences are between how GDSC and CCLE computed the AUC and IC_50 that are different than how PharmacoGX does. This is a very central concept in your comparison so it would have to have a very solid definition and analysis. Supplemental Figure 4 suggests in a round-about way that the only difference in in the number of cell lines (figure caption). This is a bit confusing.

Again, I am not sure what "Dataset 2" refers to. Perhaps the manuscript would benefit from the addition of some interpretation as to what you believe Supp Figure 4 means.

I don't understand the definition of your three classes of drugs (no effect (AUC > 0.2); narrow effect AUC \leq 0.13 or broad affect AUC > 0.13). I don't see how this definition clearly delineates between "no effect" and "narrow effect".

The bottom paragraph of the first column is one sentence that spans 8.5 lines. It references 2 main figures of the paper and 5 supplementary figures. To be honest, this is very frustrating. I have gone through the Supplemental Methods very closely and I don't see anywhere where the authors have distinguished between "recomputed AUC" and "AUC computed based on the common concentration range". Then "IC_50 (figure figure) values" ??

I'm not sure what to interpret re: Figure 7 for example. To me it looks like there is excellent agreement except for perhaps the first row. Only paclitaxel fits into this "cyotoxic drug' category but for the life of me, I don't see where this is defined. The authors just defined three types of drugs (no effect, narrow effect and broad effect) but that's not what they are using here. I don't understand this. To me it simply seems to be that at low AUCs there is high variance in the last 3 distributions of the first line (17-AAG PD-0325901 and AZD6244), but actually they look like they pretty well agree at higher AUCs. I'm not sure what that means

"and calculated the consistency of drug sensitivity data between studies using all common cases and only those that the data suggested were sensitive in at least on study.

Maybe a table would help, especially if each of these different objects were properly defined in the Methods/Supp Methods.

“Given that no single metric can capture all forms of consistency, …” So you add three more. I don’t see the point here. Why these three? and how is something like pearson \rho applied here. What is the vector? I would guess that Supp Figures should show the distribution of correlations for all three distributions so that we can look the different moments of these distributions (e.g. skew). In Figure 8, there is a use of a * but how where these p-values estimated? Are these empirical estimations of the p-value?!

I am totally confused here. You say in Figure 8 that panel A is “full data”. But panel B is “sensitive cell lines”. Where is this defined? the parentheses beside this in the figure caption? But why did you introduce this “broad, narrow, no effect” definitions only to redefine something else here?

I’m not sure I understand Supplemental Figure 11. Is this just all probe groups for the Affy arrays, or how were features chosen?
What is an “RNA-seq expression value”? how is this formed? rpm? Most importantly, I just don’t know what the message is here, and if there is any statistics
to support that statement.

I’m not sure I understand Figure 9 or what the take home message should be. I have a hard time understanding the labels along the x-axis in these figure. I just don’t really know statistically how one can conclude that gene expression is more “consistent” than the drug sensitivity values. There could be a million things going on in those arrays. There are so many more datapoint and you have literally a hundred thousand probes that probably don’t have an IQR > 1.5 on those arrays that “pump up” the correlation values, I would guess. What does this analysis mean?

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to state that I do not consider it to be of an acceptable scientific standard, for reasons outlined above.

Author Response 12 Jul 2017

Benjamin Haibe-Kains,

We thank the reviewer for his constructive comments. We too believe it is important for the community to be aware of the challenges for biomarker discovery stemming from the lack of consistency across large-scale pharmacogenomic datasets. We have addressed most of the reviewer’s comments, as detailed below.

pg5: What is “Dataset 1”? The link here doesn’t lead anywhere that I can tell.

We have updated the manuscript to add a description of each “Dataset”.

Figure 3. I’m not really sure what the value is in plotting the density functions for the mismatched and matched cell lines. First, wouldn’t one density function suffice with a threshold I guess? Second, do you really need it at all?

The representation of both mismatched and matched cell line density functions serves to illustrates two main points: to give context to the concordance scores, and to show that the mismatched cell lines have a concordance score that is distinct and separate from the matched cell lines. By representing only a single density function, a reader may not be able to appreciate the bimodal nature of matched/mismatched concordances, and that the distance between the two functions is *large enough to allow for a robust classification scheme.*

In the Methods “Cell line identity….”, it is stated that 66 samples fell below threshold with a reference to the Supplementary methods. However I don’t see anything in the supplementary methods that discusses this. Moreover in the text, it seems that you threw 8 cases away. This is confusing.

We analyzed the SNP array profiles of all 973 and 1190 CEL files available for GDSC and CCLE respectively. From these samples 66 and 5 with low quality SNP arrays in GDSC and CCLE have been removed from SNP fingerprinting pipeline. We continued the SNP fingerprinting pipeline with 503 cell lines with high quality SNP profiles available in common between CCLE and GDSC.

We compared the genotype concordance score for 503 out of 698 cell lines in common between CCLE and GDSC. We confirm that we removed 8 of 503 cell lines from analyses because their genotype concordance scores fell below the 0.8 threshold, as well as within the range of genotype concordances for cell lines with discordant genotypes. As such, we concluded that despite having the same annotations, these cell lines may have been contaminated or mislabelled in one of the two studies and further analysis on drug sensitivity cannot be compared between them. We have reported the quality scores and concordance scores in Dataset 1.

pg 5. I’m not sure what you mean by “remain stable or decrease monotonically”? Do you just mean “monotonically non-increasing”?

That is correct. We changed the manuscript accordingly.

Please see comment regarding “Filtering of drug dose-response curves” form Supp Methods below. I think this really needs to be reworked, and I have to trust you guys here that you are doing the right thing.

“as exemplified…” depicted?

We agree and updated the manuscript accordingly.

In Figure 4, is it possible to relate this back to the choice of \rho, \epsilon and 2 \cdot \epsilon from the Supp Methods, or perhaps integration a version of this figure (but annotated) into the Supp Methods.

Figure 4 legend is updated with the explanation of why these curves are identified as noisy.

In panel A, the grey area is a bit non-intuitive no? I would say that post 0.03, it’s looking pretty good, and it’s not measured in GDSC after that. However the first points are off.

In this panel the problematic curve is CCLE While the curve is monotonically decreasing for just 62% of points (5 out of 8) it is expected to be for at least 80% (the value considered for the \rho parameter) of points. However, GDSC curve is consistently decreasing monotonically for all points except for one. So we think it is a good example of how the constraints we considered are useful in identifying noisy curves like CCLE curve in this panel.

I don’t know what \epsilon or \rho are so it’s hard to relate what is depicted in Figure 4

back to your model.

Figure 4 legend is updated with the default values of \epsilon and \rho which have been used to flag noisy curves in our study.

When I look through Supplemental Figure 1 (all the excluded comparisons), it seems like your criterion for excluding a comparison boils down cases where at least one of the curves has high variance, and the cutoff 2\epsilon I think is a constant independent of the distribution of points for either curve. I don't see in your equations how you encode that the sequence is monotonically non-increasing, or how "order" along the left to right sweep is incorporated.

By definition, we expect to see two types of drug response curves. When the cell is resistant to the drug it is expected that the viability fluctuates slightly around 100% and when the drug is sensitive a monotonically decreasing manner is expected to be seen. However, noise is unavoidable in these experiments. So we assumed that the viability of each point on the curve may get higher than the viability of its immediate prior at most with the size of \epsilon. To filter the largely noisy experiments and keep the slightly noisy ones at the same time, we consider this constraint to be true for the majority of points which is defined by \rho. Applying these simple constraints will result in omitting all the curves in which viability is increasing monotonically or if it is fluctuating largely.

Supp Methods Filtering of drug dose-response curves

i +1^{st} -> (i+1)^{st}

Thanks for pointing this out, we corrected this in the revised manuscript.

It seems like there is a formatting problem here. In my pdf there is something like I"A squiggle after "in some large fraction ??? of the cases (1)." I guess that's supposed to be \rho right?

Thanks for pointing it out. You are right and we corrected this.

I'm not sure I understand this sentence "Our quality control …" Are you saying that \Delta_{i, i+1} < \epsilon in some fraction \phi of the cases?

You are correct and more explanation is presented in our previous comment.

equation #2 below:
I also have never seen set notation such as \{ \Delta_{i,i+1} | \Delta_{i,i+1} < \epsilon \}.
Are you trying to say "given all the \Deltas that are less than \epsilon?
So the vertical | means cardinality here right?
But then the denominator has the cardinality of a value, or do you mean absolute value?

Vertical bars are used as cardinality notation in both numerator and denominator of that equation.
What is \rho? It's undefined.

"Unfortunately …" The english is a bit rough - could be rephrased in terms of specificity and sensitivity, I guess.

We updated the manuscript accordingly.

I don’t understand the significance of the sentence “Consider, for instance, …” But in the main text, you say that it remains stable or decreases monotonically. Here it increases monotonically for many successive points, so this violates your model, no?

Apply only equation (1) will not filter all the noisy cases and the curve explained in this sentence is one of those cases. Hence we also applied equations (2) and (3) to filter these remaining noisy curves. We clarified this in the revised manuscript.

I think that this subsection wants simply to spell out mathematically the thresholds and also provide some rationale for the parameters. I think the text doesn’t really do a good job of establishing this rationale and needs work. Perhaps define the parameters precisely and then phrase the exposition in standard terms e.g. specific and sensitivity for different \epislon, etc.

We thank the reviewer for his suggestion. We improved the clarity of our equations and clearly state the parameters we used. We agree that the selection of the parameter value is arbitrarily, although reasonable. It would be possible to create a set of manually curated curves as a gold standard set and tune our parameters accordingly. Although the idea is appealing, it would require a large set of curators as manual classification tends to be unstable too. Such an analysis is definitely of interest and we will pursue this in future studies.

Your equation (2) is inconsistent. In the text you specify \Sigma_{\forall i,j} \Delta_{i,j} but below your criterion seems to change to (the correct) $i < j$. It is also sufficient to write \Sigma{ i < j } \Delta_{i,j} and avoid the double summation.

Corrected.

You should probably define D_i in the text and not make the reader deduce it from the figure below

Corrected.

The comments here w.r.t. the Supplementary Methods also apply to the associated subsection of the Methods. The mathematical correctness of some comments needs attention. For example, it is not quite correct to say that the “curve fitting would have yielded erroneous results”. The curve fitting is just that, curve fitting. It’s not really an error. Then in the Methods, you claim to use this equation but this is inconsistent with the discussion in the Supplemental Methods (where you have an equation with this undefined parameter E).

We agree with the reviewer and updated the manuscript accordingly.

The least-squares method using a three-parameter sigmoid model. I understand the intuition for this, but when I look through Supplementary File 2, I think there are a lot cases where this is perhaps not the correct pattern to assume (e.g. straight lines). Moreover, there are some very strange fits, for example,

AZD6244:G–361
AZD6244:SK–MM–2
PLX4720:MDA–MB–175–VII
In somecases the curve is always above all of the observations.
Perhaps this isbecause the measurements of viability > 100%? In your model you have removed the E_0 from Barretina et al. for different reasons.

Data points where the measured viability exceeds 100% always lie above their respective curves of best fit, since the functional form of the equation forces predicted viability to lie between 0 and 100% for all drug concentrations. The three-parameter sigmoid model's intuition and flexibility makes it an attractive choice for the majority of cases, and for ease and uniformity of analysis, we felt it prudent to choose the same model to fit all curves, even if it may not have fit well in a few anomalous cases.

Supp Methods - Fitting of drug dose-response curves

I'm a bit lost with your choice of notation here. For example, you define y as an equation, not a function, and then write y(0) which I assume is supposed to be something like y(x). Ok but then you have y=0 and y = y(0) = 1. I'm not sure this is correct mathematically. (I think you mean to say that y(x) = … *where* y(0) =1.

"viability is reduced to half … concentration of the drug"… so the "Top", E_\infinity … I find this a bit wordy.

"The dose response equation now becomes …"

So I deduce that E is the new parameter?!?! Where has E_\infinity gone?!?!

(But then down below E is constrained to be in [0,1] and seems to related to the fitness of neoplastic cells. I'm not sure I understand this.)

There is no derivation of this formula whatsoever. In fact, I don't see how this could be correct any longer. Couldn't this be expressed asmixture of two cell types, and y would be then a sort of weighted average?

I really don't see how this was derived. This is a very central part of your paper (since the manuscript is measuring agreement) and therefore it needs to be bulletproof.

Please define "extant drug".Also HS is allowed to vary apparently but I don't see where it is then optimized in your analysis later on. This is confusing.

We thank the reviewer for his comments and we apologise for the lack of clarity of our Supplementary Methods. We have now rewritten this section to improve clarity and address all the reviewer's comments.

Consistency of Drug Sensitivity Data

Is the ABC method standard? Are there citations for this? You should probably define properly what you mean by the "insertion of the concentration range". Elsewhere it seems

that you are referring to this as “common concentration range” e.g. SFig 2

More generally, isn’t this a sort of (non-statistical) version of the Kolmogorov-Smirnoff test?

We created and used the ABC method as a convenient and intuitive way of quantifying the agreement of analogous dose-response curves in different datasets over the intersection of the concentration ranges tested by them (henceforth referred to as their “common concentration range”). This method is inspired from two recent publications where the authors restricted the analysis to the common concentration range between datasets (Pozdeyev et al, Oncotarget 2016) and compared two curves directly (Yadav et al, Scientific Reports 2014).

While ABC does have some similarities to Kolmogoroff-Smirnoff, it evaluates the area between curves rather than the maximum vertical linear distance between them. Furthermore, it takes into account the behaviour of the fitted dose-response curves over their common concentration range only, rather than across their entire domains. Since the behaviour of fitted dose-response curves at concentrations far outside the concentration ranges over which they were fitted tends not to be robust to noise, we felt that ABC was a more appropriate test than Kolmogorov-Smirnoff for assessing accordance of dose-response curves in this study.

We updated the manuscript wit these references and clarifications.

Figure 5A: Actually there are many such cases in your Supplementary Figures. Doesn’t this just mean that the range of concentrations are not sufficient in both datasets?

Given the limited concentration range tested in high-throughput in vitro drug screening studies, such as GDSC nd CCLE, it is not possible to rule out that drug yielding no effect on cell viability could actually yield substantial effect at higher dose. However, these higher doses are likely to be clinically irrelevant.

pg 7 “We then computed the median …” I don’t understand what your distance metric is here. If I understand correctly, you could the ABC for each pair of drugs in the GDSC dataset. From that I can imagine a distance matrix D where D_ij i the ABC between drugs i and j. But you said you take the median ABC? median over what? cell lines? repeats? Whatever the case, are you sure it’s a distance metric?! Is it really true that distances derived from ABCs are metrics? I think this should be shown in the Supplemental Methods.

The ABC is computed for each cell lines and the median of ABC values was used as a measure of “distance” between two drugs.

Also as a minor comment, the caption in Supp Figure 3 says that you are using the mean ABC value but elsewhere it says median.

The caption has been corrected to read ‘median’ rather than ‘mean’.

p8. I am not sure what the significance of Supplementary Figure 3 is: are any of these clusters significant? Why are two drugs coloured red in panel A? On page 8 it is claimed that the samples split by hospital (MGH vs WTSI) but I don’t see how this is represented in

Supplemental Figure 3.

As described in the manuscript, GDSC drug sensitivity experiments have been performed in two centers (MGH and MSTI) separately. The only drug has been tested by both centres is AZD6482 (the ones in red in panel A of Supplementary figure 3). The aim of that figure is to illustrate how well the biological replicates have been clustered together. However the other drugs are not expected to cluster according to their corresponding center thus there is not such a labeling in this figure.

pg 8. I have a very hard time estimating the significance of a statement like "…3 out of 15 common drugs clustered tightly". I am not sure what tight means here.
When I look at Supp Figure 3 there has been no effort to annotate the clusters with their reproducibility e.g.pvclust or measure their significance in some other way.
When I look at the figure I think there appears to be a lot of co-clustering of the drugs, at least given that the median ABC across a diverse collection of cell lines might not be such a great "distance" measure.

We refer to the closest neighbor for each drug. We agree with the reviewer that our statement should be more quantitative. We therefore compared the ABC values between common drugs and different drugs and observed a significant differences (one-sided Wilcoxon test p-value = 0.004). We agree that median ABC might not be the best distance measure, this is why we updated Supplementary Figure 3 with other distance measures for completeness.

pg 9. What do you mean by "highly targeted therapies"?

It meant drugs for which there is a few sensitive cell lines in the CCLE and GDSC (narrow effect). However, we changed it to targeted to avoid any confusion.

pg 9. paragraph "Although the ABC values …" I think the ABC is interesting but it takes a very prominent role in your paper when there are other standard techniques already like AUC and IC_50. Perhaps the manuscript should have comments about why have chosen this approach that is not standard. Also I think you would need to make precise what the differences are between how GDSC and CCLE computed the AUC and IC_50 that are different than how PharmacoGX does.

GDSC and CCLE fit a different family of curves to their dose-response data, as described in Haibe-Kains et al, Nature 2013. To eliminate this source of heterogeneity, we fitted the same three-parameter model for all the CCLE and GDSC curves, as implemented in PharmacoGx does. Once the curve is fitted, GDSC, CCLE, and PharmacoGx agree on how to calculate its AUC and IC_50.

This is a very central concept in your comparison so it would have to have a very solid definition and analysis. Supplemental Figure 4 suggests in a round-about way that the only difference in in the number of cell lines (figure caption). This is a bit confusing. Again, I am not sure what "Dataset 2" refers to. Perhaps the manuscript would benefit from the addition of some interpretation as to what you believe Supp Figure 4 means.

We have updated the manuscript with a clear interpretation of Suppl Figure 4, which shows that drugs listed as targeted therapies exhibit less variation (as estimated by the median absolute deviation) in drug sensitivity (AUC) than cytotoxic therapies. Although expected, these results *allowed us to define a cutoff fro MAD(AUC) to classify drug into broad vs narrow effect, as described in the manuscript.*

We have also updated the manuscript to add a description of each "Dataset", a denomination required by F1000Research formatting guidelines.

I don't understand the definition of your three classes of drugs (no effect (AUC > 0.2); narrow effect AUC \leq 0.13 or broad affect AUC > 0.13). I don't see how this definition clearly delineates between "no effect" and "narrow effect".

We apologize for the confusion. We corrected the manuscript with the following definitions. Drugs with "no effect": all AUC values < 0.2 (no sensitive cell lines); "narrow effect": MAD(AUC) <= 0.13 (see Suppl Figure 4); "broad effect": MAD(AUC) > 0.13 (see Supp Figure 4).

The bottom paragraph of the first column is one sentence that spans 8.5 lines. It references 2 main figures of the paper and 5 supplementary figures. To be honest, this is very frustrating. I have gone through the Supplemental Methods very closely and I don't see anywhere where the authors have distinguished between "recomputed AUC" and "AUC computed based on the common concentration range". Then "IC_50 (figure figure) values" ??

We have now clearly stated these definitions in the manuscript. AUC recomputed and AUC computed based on common concentration range both are computed by our PharmacoGx package by fitting the sigmoid model described in the Supplemental Methods. The only difference between these metrics is that the former is computed over the whole concentration range for each study while the former one is computed over the common concentration range between CCLE and GDSC. We updated the manuscript to reflect the fact that recomputed IC_50 values have been used in Supplementary Figure 8. Recomputed IC_50 values are inferred from the sigmoid model fitted to the data by the aim of PharmacoGx package.

I'm not sure what to interpret re: Figure 7 for example. To me it looks like there is excellent agreement except for perhaps the first row. Only paclitaxel fits into this "cyotoxic drug' category but for the life of me, I don't see where this is defined. The authors just defined three types of drugs (no effect, narrow effect and broad effect) but that's not what they are using here. I don't understand this. To me it simply seems to be that at low AUCs there is high variance in the last 3 distributions of the first line (17-AAG PD-0325901 and AZD6244), but actually they look like they pretty well agree at higher AUCs.

We agree with the reviewer that we have not been consistent in our definition of drugs with no, narrow and broad effect. This is now fixed in the updated manuscript. In Figure 7 (and all other figures) we have ordered the drugs by their "status" (no, narrow and broad effect). For the ease of interpretation, we also choose to color each AUC based on a standard cutoff for sensitivity of AUC > 0.2 (and therefore cell lines with AUC <= 0.2 are called "insensitive"). Although paclitaxel is the only drugs that is referred to as cytotoxic in the literature, we observed that 17-AAG, PD-0325901, and AZD6244 decreases cell viability for a large number of cell lines. As their MAD(AUC) > 0.13 (Supp Figure 4), we classified these drugs as "broad effect". In this case, the consistency of drug sensitivity data (AUC) seems to be poor, with CCLE having much more sensitive cell lines than GDSC. Drugs with narrow effect (MAD(AUC) <= 0.13) (2 middle rows) yield better consistency for some drugs (e.g., crizotinib, PLX4720, lapatinib, lapatinib) but there are still cell lines with AUC >

0.2 ("sensitive") that are far off the diagonal. The last row include all the drugs with "no effect", i.e., the vast majority of cell lines yielded AUC <= 0.2, where no consistency is expected due to low signal / noise ratio.

I'm not sure what that means "and calculated the consistency of drug sensitivity data between studies using all common cases and only those that the data suggested were sensitive in at least on study.

The consistency of drug sensitivity data was performed twice. First using all the cell lines in common between the two studies. Second, using only the cell lines that are "sensitive" (AUC > 0.2) in at least one dataset. The second analysis aims to address a criticism we received from the community that only sensitive cell lines should be compared. We rephrased this part in the updated version of the manuscript.

Maybe a table would help, especially if each of these different objectswere properly defined in the Methods/Supp Methods.

We enriched the "acronym table" in Supplementary Methods to add definitions of the additional objects and concepts used in our paper.

"Given that no single metric can capture all forms of consistency, …" So you add three more. I don't see the point here. Why these three? and how is something likepearson \rho applied here. What is the vector? I would guess that Supp Figures should show the distribution of correlations for all three distributions so that we can look the different moments of these distributions (e.g. skew). In Figure 8, there is a use of a * but how where these p-values estimated? Are these empirical estimations of the p-value?!

In the absence of gold standard measure of consistency for drug sensitivity data, we decided to include other measures that could be used as alternative to Pearson and Spearman correlation already used in previous publications. The consistency measures are computed across cell lines. For each drug, a vector of drug sensitivity measurements (AUC' IC_50,...) is extracted from GDSC and CCLE and then compared. P-values were computed analytically, as described in the updated Supplemental Methods. We updated the caption of Figure 8 to state these important points.

I am totally confused here. You say in Figure 8 that panel A is "full data". But panel B is "sensitive cell lines". Where is this defined? the parentheses beside this in the figure caption? But why did you introduce this "broad, narrow, no effect" definitions only to redefine something else here?

We apologise for the lack of definition and inconsistency. We have updated the figure to use a consistent classification of drugs and now clearly define the restriction to "sensitive data" (now renamed as "sensitive cell lines" for clarity).

**I'm not sure I understand Supplemental Figure 11. Is this just all probe groups for the Affy arrays, or how were features chosen?
What is an "RNA-seq expression value"? how is this formed? rpm? Most importantly, I just don't know what the message is here, and if there is any statistics to support that statement.**

Brainarray probe gene mapping cdf files have been used to quantify the expression value for each gene represented on the Affymetrix arrays. FPKM values for genes annotated by Gencode V.19 annotation were normalized by transforming them using log2(FPKM+1). The aim of Supplementary Figure 11 is to show the distribution of expression data for each platform and how well a mixture of 2 gaussians could help define a cutoff to binarize the data. The caption has been updated to clearly state how the cutoffs have been determined.

I'm not sure I understand Figure 9 or what the take home message should be. I have a hard time understanding the labels along the x-axis in these figure. I just don't really know statistically how one can conclude that gene expression is more "consistent" than the drug sensitivity values. There could be a million things going on in those arrays. There are so many more datapoint and you have literally a hundred thousand probes that probably don't have an IQR > 1.5 on those arrays that "pump up" the correlation values, I would guess. What does this analysis mean?

We have now clearly defined the labels of the x-axis in the caption. We agree with the reviewer that sensitivity data and gene expression data have very different properties. However, as we looked at univariate biomarkers, one gene at a time, we sought to assess whether expression of each individual gene suffers from the same level of inconsistency than drug sensitivity data across cell lines. We have added a word of caution in the text to reflect on the limitation of this analysis.

Competing Interests: None