Drug response consistency in CCLE and CGP

ARISING FROM B. Haibe-Kains *et al. Nature* **504,** 389–393 (2013); doi:10.1038/nature12831

The Cancer Cell Line Encyclopedia[1] (CCLE) and Cancer Genome Project[2] (CGP) are two independent large-scale efforts to characterize genomes, mRNA expression, and anti-cancer drug dose–responses across cell lines, providing a public resource relating cellular biochemical context to drug sensitivity. A recent study[3] analysed correlations between reported dose–response metrics and found inconsistency between CCLE and CGP, thus questioning the validity of not only these, but also other current and future costly large-scale studies. Here, we examine two metrics of drug responsiveness (slope and area under the curve) that we derive from the original CCLE and CGP data, and find reasonable and statistically significant consistency. Our results revive confidence that the CCLE and CGP drug dose–response data are of sufficient quality for meaningful analyses. There is a Reply to this Comment by Safikhani, Z. *et al. Nature* **540,** http://dx.doi.org/10.1038/nature20581 (2016).

CCLE and CGP share 2,520 dose–responses across 285 cell lines and 15 drugs, but cells were treated with different dose ranges. To compare

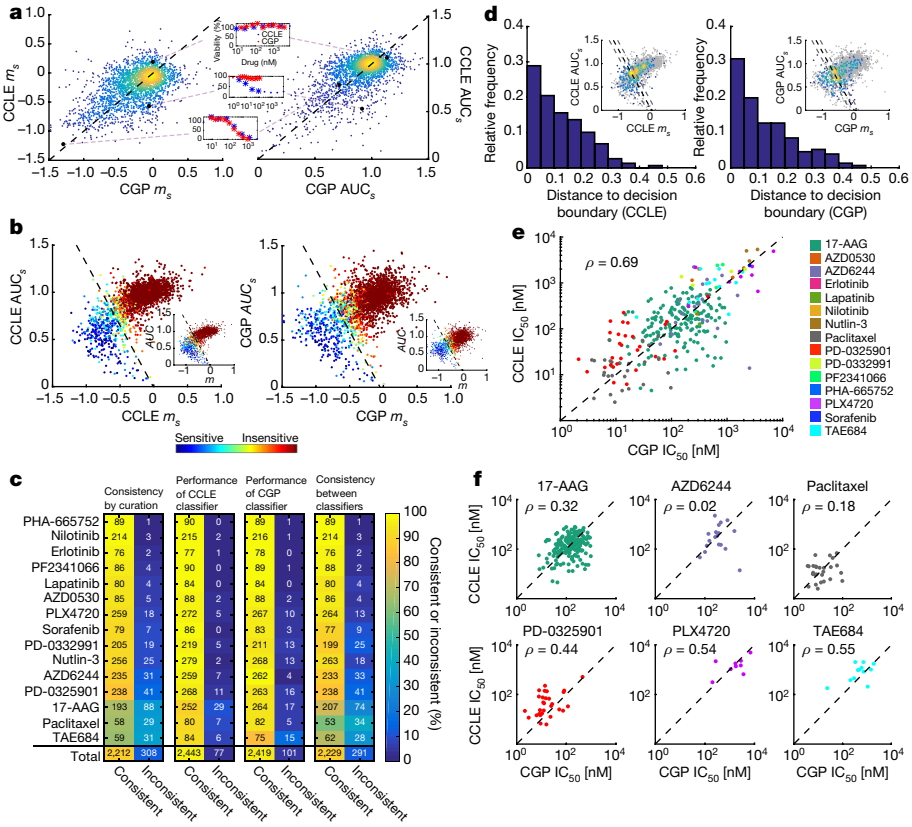


Figure 1 | Consistency between pharmacological data in CCLE and CGP. a, Slope (m_s; left) or area under the curve (AUC$_s$; right) of the dose–response curves for all overlapping drug/cell line pairs (2,520) in CCLE and CGP, considering only the shared dose range (denoted by subscript *s*). All m_s and AUC$_s$ values were normalized based on the respective drug dose range, to facilitate comparison across drugs (see Supplementary Methods). Colour indicates density of dots. The black dashed line is $x = y$. In example dose–response curves, stars represent the shared dose range. **b**, Relationship between m_s and AUC$_s$ for each database (inset *m* and AUC defined with the entire dose range as opposed to the shared dose range). The SVM classifier decision boundary divides the plot into sensitive and insensitive drug/cell line pairs, as indicated by the black dashed line. Slope and *y*-intercept of boundary line for CCLE$_s$: $m = -1.32$, $b = -0.01$; CGP$_s$: $m = -1.31$, $b = -0.06$. Colour of dots indicates the mean of the binary classifications from eight manual curators; blue indicates a unanimous sensitivity rating, green a very uncertain rating, and red a unanimous insensitivity rating. **c**, Consistency (left) and inconsistency (right) of classification methods broken down by drug. Far left plot shows manual curation consistency between CCLE and CGP. Middle left plot shows consistency between the manual curation data from CCLE and the CCLE SVM classifier. Middle right plot shows consistency between the manual curation data from CGP and the CGP SVM classifier. Far right plot shows consistency between the CCLE SVM classifier used to classify CGP data and the CGP SVM classifier used to classify CCLE data. Colour indicates percentage consistency as denoted by the colour bar. Numbers denote number of observations, black for consistent, white for inconsistent. **d**, Inconsistent drug/cell line pairs based on manual curation results. Histograms bin the Euclidian distance between each discrepantly classified drug/cell line pair (that is, called sensitive in one database and insensitive in the other) and the decision boundary (black dashed line) in the AUC$_s$ versus m_s plots for CCLE (left) or CGP (right). In inset, coloured dots indicate drug/cell line pairs that were classified discrepantly in CCLE and CGP. Colour corresponds to density of dots. Black dashed line indicates the decision boundary for the SVM classifier. Grey dashed lines indicate a Euclidian distance of 0.1 from the decision boundary in either direction **e**, IC$_{50}$ values from all sensitive cell line/drug combinations as determined by SVM classifier in CCLE or CGP. The black dashed line is $x = y$. **f**, IC$_{50}$ values from all sensitive cell line/drug pairs (same as in Fig. 1e) stratified by drug, for drugs having at least 5 points. All correlation coefficients are Pearson.

CCLE and CGP dose–responses, we calculated a common viability metric (0–100%) across a shared $\log_{10}$-dose range, and computed slope (m_s) and area under the curve (AUC_s) values (in which subscript 's' denotes the shared dose range) (Fig. 1a). This analysis revealed surprisingly good quantitative agreement between the two studies (m_s: population Pearson correlation coefficient (ρ) = 0.52, $P < 10^{-16}$; AUC_s: $\rho = 0.61$, $P < 10^{-16}$). Furthermore, since a small m_s or large AUC_s value indicates insensitivity, these data suggest that most cell lines are insensitive to the majority of tested drugs (~85%, Fig. 1a, b). Characterizing such insensitive trends with a sigmoid model meant for sensitive cell lines (that is, half-maximum inhibitory concentration, IC_{50}) may lead to incorrect dataset consistency conclusions.

To evaluate consistency of sensitivity classification between the two studies, we first asked eight people to curate binary sensitivity manually (all dose–response curves and their manually curated classification results are provided in the Supplementary Information). For manual curation, only data from a single database within the shared dose range was presented on each plot, and the order of plot presentation was randomized with respect to the study, the drug, and the cell line for each curator (see Extended Data Figs 1 and 2). Using the manual curation results, we built a separate support vector machine (SVM) classifier for each study with m_s and AUC_s as predictors (Fig. 1b). Both SVMs performed well (Fig. 1c, middle two plots), and the decision boundaries are independently similar for CCLE and CGP (Fig. 1b, black dashed line). These SVM classifiers also seem to parse data derived from the full (not shared) range of drug doses effectively (Fig. 1b; insets; m and AUC without subscript s), which may be important for future, database-specific analyses.

The manual curation data along with the SVM classifiers allowed evaluation of consistency between CCLE and CGP in terms of binary sensitivity classification (Fig. 1c). Comparison of manual curation results shows high (~88%) and statistically significant consistency between the two studies overall (Cohen's kappa (κ) = 0.53 ± 0.025), and for most individual drugs (Fig. 1c, far left). Using the CCLE SVM to classify CGP data, and vice versa (Fig. 1c, far right), also yielded high and statistically significant consistency (88%, $\kappa = 0.55 \pm 0.025$). These results strongly suggest that drug dose–response data in the CCLE and CGP can be considered consistent when used to classify binary sensitivity.

The drugs 17-AAG, paclitaxel and TAE684 account for 48% of the inconsistent drug/cell line pairs. We hypothesized that most of these and other inconsistent drug/cell line pairs would be located near the SVM decision boundary. The primary reason is because this boundary necessarily travels through the region of AUC_s–m_s space where determining binary sensitivity is the most challenging for manual curators (Fig. 1b, cyan to yellow dots denote uncertainty among curators). If true, then this would imply that a main factor driving the observed inconsistency is self-induced: imposing a strict cutoff. Indeed, most such inconsistent points are located close to the decision boundary; for CCLE 53% of the inconsistent points are within 0.1 distance units from the decision boundary, and 51% for CGP (Fig. 1d). Manual inspection of these inconsistent binary classification cases also supports this interpretation (Supplementary Data 1). We do observe some strongly inconsistent drug cell/line pairs (for example, Fig. 1a inset-middle, and Supplementary Data 1), but these are relatively rare, and are highly likely to be located far from a decision boundary. These results suggest that inconsistency between the two studies on the level of binary classification is, to a large extent, a result of the information loss associated with collapsing a two-dimensional continuous description of drug sensitivity onto a single binary variable. Thus, we propose that drug sensitivity is better described as a spectrum (AUC and m) than as a binary classification.

We next re-calculated and compared IC_{50} data only from drug/cell line pairs determined to be sensitive in either CCLE or CGP by the SVM classifier (another requirement was the existence of a non-extrapolated IC_{50} value). We found good correlation between the two studies overall (Fig. 1e; $\rho = 0.69$, $P < 0.0001$). However, stratification by drug generally yields poor IC_{50} correlations (Fig. 1f). Thus, caution should be taken for inference of IC_{50} values for specific cell line/drug combinations from CCLE and CGP, despite consistency on the level of slope, area under the curve, and binary sensitivity classification. Haibe-Kains *et al.*[3] stratified IC_{50} by drug for sensitive and insensitive lines (IC_{50} values for insensitive lines are unreliable), which contributed to their conclusion of inconsistency.

We conclude that the drug dose–response data in CCLE and CGP are acceptably consistent for most cases. Furthermore, we made no attempts to remove potentially suspect dose–response data, but doing so in future efforts could further facilitate data usability. That the two studies are this consistent is quite remarkable, given the different viability assays used, as well as inescapable confounding factors such as cell confluency, clonal variations, genomic drift, different drug suppliers/batches, laboratories/equipment and serum composition. This suggests that the measured genomic and gene expression parameters may provide a robust cellular context that dictates drug sensitivity.

Methods

For each drug/cell line pair found in both CCLE and CGP, we calculated the slope and AUC of each dose–response curve (percentage cell viability versus $\log_{10}$ drug dose) only in the shared dose range. These values were normalized to account for different dose ranges used by each drug. One CCLE point and one CGP point defined boundaries of the shared dose range to maximize data coverage. IC_{50} values were calculated as the drug concentration needed to reach 50% cell viability (using a fit to a sigmoid response model) if within the shared dose range (see Supplementary Methods). All scripts and data needed to reproduce the figures, including the MATLAB code, are provided in Supplementary Data 2.

Mehdi Bouhaddou[1], Matthew S. DiStefano[1], Eric A. Riesel[1], Emilce Carrasco[1], Hadassa Y. Holzapfel[1], DeAnalisa C. Jones[1], Gregory R. Smith[1], Alan D. Stern[1], Sulaiman S. Somani[1], T. Victoria Thompson[1] & Marc R. Birtwistle[1,2,3]

[1]Department of Pharmacological Sciences, Icahn School of Medicine at Mount Sinai, New York, New York 10029, USA.
email: marc.birtwistle@mssm.edu
[2]DToXS LINCS Center, Icahn School of Medicine at Mount Sinai, New York, New York 10029, USA.
[3]Systems Biology Center New York, Icahn School of Medicine at Mount Sinai, New York, New York 10029, USA.

Received 15 November 2014; accepted 13 October 2016.

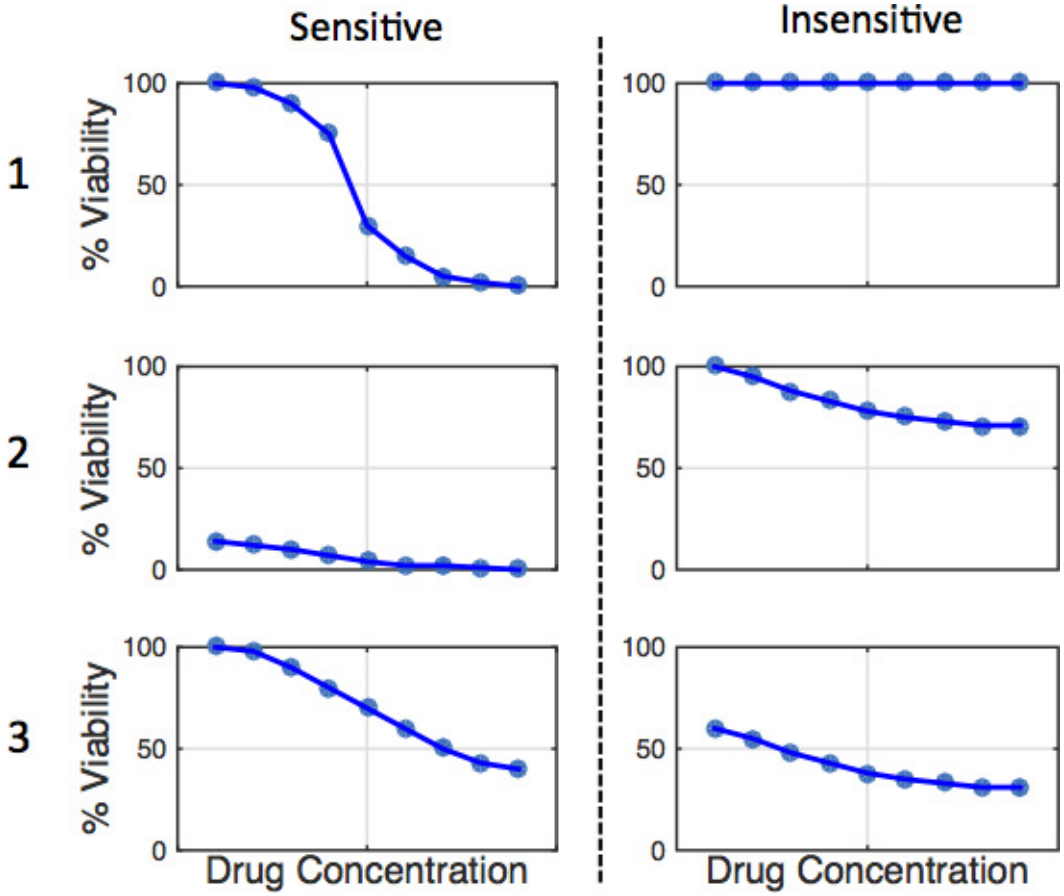
1. Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483,** 603–607 (2012).
2. Garnett, M. J. *et al.* Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature* **483,** 570–575 (2012).
3. Haibe-Kains, B. *et al.* Inconsistency in large pharmacogenomic studies. *Nature* **504,** 389–393 (2013).

Supplementary Information is available in the online version of the paper.

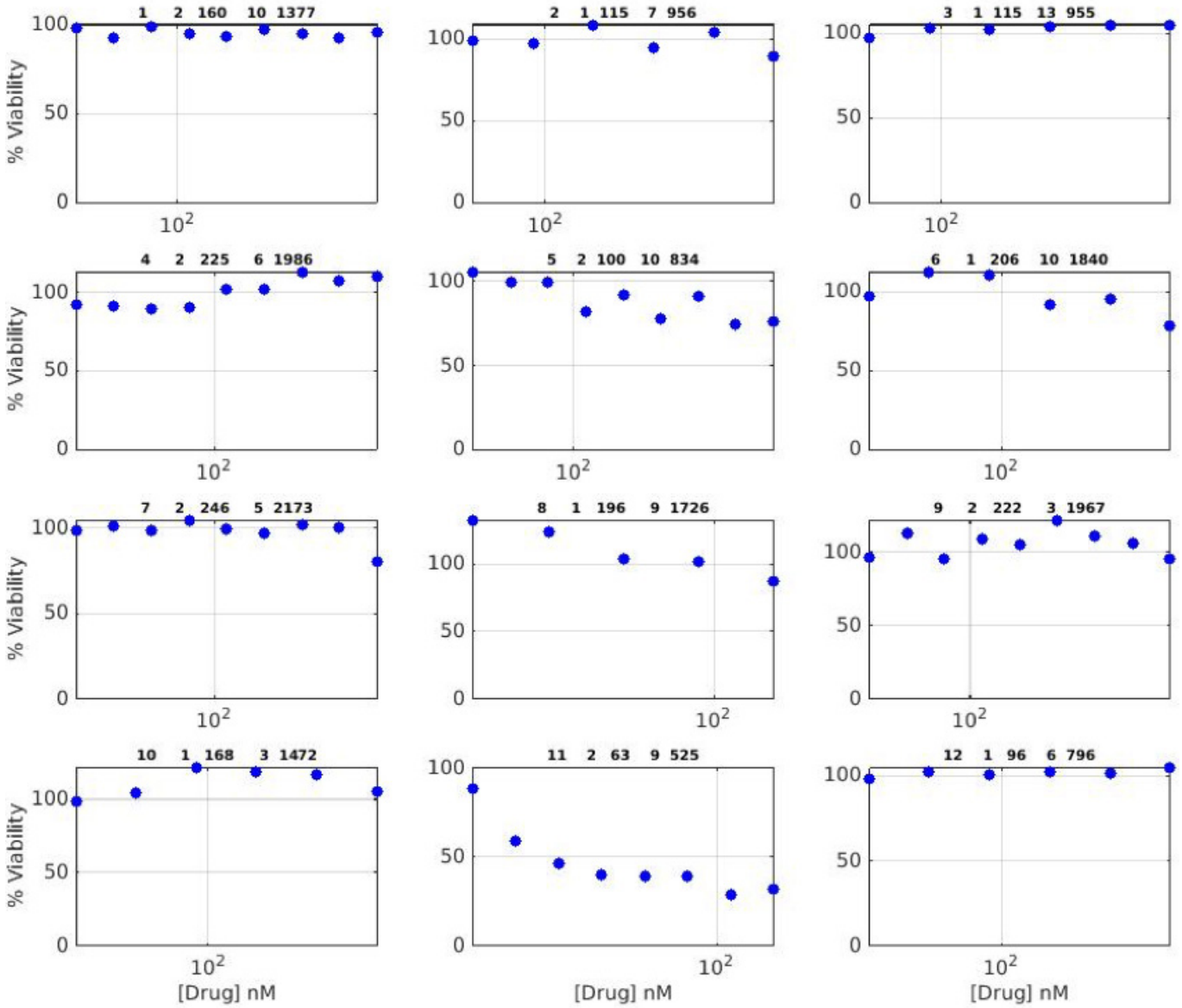
Author Contributions M.R.B. conceived of the study. M.B. and M.R.B. performed the analyses and wrote the paper. M.S.D. prepared CCLE and CGP data for analysis and performed preliminary analyses. E.A.R., E.C., H.Y.H., D.C.J., G.R.S., A.D.S., S.S.S. and T.V.T. served as manual curators.

Competing Financial Interests Declared none.

doi:10.1038/nature20580



Extended Data Figure 1 | Examples of typical sensitive versus insensitive dose–response curves. This document was given to manual curators as example dose–response curves. These idealized data represent various dose–response curves one might encounter in CCLE and/or CGP and indicate how they should be classified.



Extended Data Figure 2 | Examples of what the manual curators received. This is one page, as an example, from the data given to manual curators, which they were instructed to rate as either sensitive or insensitive.

Safikhani *et al.* reply

REPLYING TO M. Bouhaddou *et al. Nature* **540,** http://dx.doi.org/10.1038/nature20580 (2016)

In the accompanying Comment[1], the authors make two main claims: (1) that viability metrics computed over the drug concentration range shared between datasets yield higher consistency than the same metrics computed over the full, often only partially overlapping, drug concentration range; and (2) that binary drug sensitivity classification (insensitive versus sensitive) as determined by manual curators increases the consistency of pharmacological profiles between the Cancer Genome Project (CGP)[2] and Cancer Cell Line Encyclopedia (CCLE)[3]. We appreciate the innovative approach followed by the authors, and our reanalysis confirms the marginal, but statistically significant, increase in consistency achieved through the use of viability metrics computed across a reduced but common range. However, our results indicate that manual classification of drug dose–response curves does not significantly increase the agreement between drug sensitivity calls compared to computational approaches. Notably, it is unclear whether the authors' manual approach will improve reproducibility of the biomarker discovery process, as collapsing of complex curves into discrete categories may result in a substantial information loss. Here we provide specific responses to the main results reported by Bouhaddou *et al.*[1]

Similar to Pozdeyev *et al.*[4] and the Comment by Mpindi *et al.*[5], the authors investigated whether sensitivity metrics computed from the drug concentration range shared between CGP and CCLE yield higher consistency[1]. Using the authors' code[1], we were able to implement their slope (m_s) and area under the curve (AUC_s) metrics (in which subscript '*s*' denotes shared dose range) in our PharmacoGx platform[6] with a minor improvement of the m_s metric to prevent highly sensitive cell lines with flat drug dose–response curves to be classified as insensitive (see Supplementary Methods). We compared the AUC and *m* metrics computed from the full and shared drug concentration range for the pooled set of drug sensitivities. Our implementation of the drug dose–response curve fitting and sensitivity computation further improved the authors' results: initial correlation for m_s ($\rho = 0.52$) and AUC_s ($\rho = 0.61$) both increased to 0.67. We then tested whether the common viability metrics constituted significant improvement over computations on the full drug concentration range. We observed a small but statistically significant improvement for both the m_s and AUC_s metrics (test of difference in correlations, $P < 0.01$; Supplementary Fig. 1). Stratifying our analysis per drug, we observed that the improvement in consistency, although significant, was marginal, with the exception of nilotinib (Supplementary Fig. 2). However, most of the drugs still yielded poor consistency ($\rho < 0.5$), which is in line with both our initial report[7] and our more recent reanalysis[8].

The authors investigated discretization of their continuous metrics to test whether binary classification would yield higher consistency, as estimated by the overall percentage agreement in drug sensitivity calls. However, such a statistic does not take into account the agreement that would be expected purely by chance owing to the large proportion of cell lines being insensitive to the tested drugs. The Matthews correlation coefficient (MCC)[9] addresses this issue. It is a balanced measure that can be used when the classes are of different sizes, and its significance can be computed using the χ^2 statistic for binary classes (Supplementary Methods). We illustrate the case of four drugs with different patterns of consistency in Supplementary Fig. 3. Although all four drugs yield an overall agreement of greater than 92%, they exhibit a wide range of MCC values. Nilotinib is a good example of a consistent drug phenotype across cell lines (MCC = 0.86; Supplementary Fig. 3a). PLX4720 yields moderate consistency (MCC = 0.68; Supplementary Fig. 3b). AZD0530 and erlotinib show only poor consistency (MCC = 0.42 and −0.05, respectively; Supplementary Fig. 3c, d). These examples support MCC as an appropriate statistic to discriminate between highly consistent drug sensitivity calls and those with poor concordance. We therefore used the MCC to compare different classification schemes, including those proposed by the authors.

Recognizing the difficulty of summarizing drug dose–response curves computationally, Bouhaddou *et al.*[1] used an unconventional approach to increase the consistency of drug sensitivity calls: they gathered a team of eight curators and asked them to classify each drug dose–response curve as either sensitive or insensitive. The authors report a Cohen's kappa (κ) calue of 0.53, which is in line with our estimated MCC value of 0.53 (Supplementary Fig. 4). The authors qualified their manual classification as a high and statistically significant consistency. We disagree with the authors' claim[1] that their results provide evidence for high consistency. We refer them to the standards for strength of agreement for κ defined previously[10], which would only classify observed consistency as moderate. More importantly, when classifications are stratified by drug, we do not observe a significant improvement of manual curation over the computational classifications based on AUC_s and m_s values ($P > 0.12$, Wilcoxon signed rank test; Supplementary Fig. 5). Consistent with our previous report, 10 out of 15 drugs (66.7%) yielded poor consistency (MCC < 0.5).

By pooling drug sensitivity data across drugs, the authors noticed a good quantitative agreement between the two studies, with estimated Pearson correlation coefficients (ρ) of 0.52 and 0.61 for AUC_s and m_s values, respectively[1]; our improved implementation of their method increased the correlation to 0.67. Nevertheless, we disagree that this level of correlation constitutes evidence for good agreement, and define it as only moderate consistency based on the interpretation scale of our initial study[7]. More importantly, the common viability metrics only marginally improved consistency at the level of individual drugs (except for nilotinib), with most of the drug yielding inconsistent drug sensitivity values ($\rho < 0.5$; Supplementary Fig. 2). The authors made a similar observation[1] in their figure 1f, undermining their claim of drug response consistency in CGP and CCLE. Moreover, the main goal of CGP and CCLE consisted of finding new associations between molecular features and sensitivity to specific drugs[2,3]. Since biomarkers are to be found for each drug separately, it is vital that pharmacological profiles are highly consistent at the level of individual drugs and not merely when averaged across a larger dataset.

In conclusion, our re-analysis of the new AUC_s and m_s metrics described by Bouhaddou *et al.*[1] showed that they represent a statistically significant improvement over the published drug sensitivity values, but the increase in consistency is only marginal for the vast majority of the drugs tested both in CCLE and CGP. Furthermore, manual classification of the drug dose–response curves does not appear to substantially improve the consistency of binary sensitivity calls over computational approaches and is not a scalable method. However, the authors[1] showed that manually classified drug dose–response curves could be used as a benchmark to train nonlinear computational predictors that could take into account the peculiar features of each individual dataset. Although there is no evidence that the authors' approaches[1] improve reproducibility of biomarker discovery for individual drugs, their work may open a new avenue of research in pharmacogenomics. Manual curation and further exploration of new

large datasets such as CTRPv2 (ref. 11) and GDSC1000 (ref. 12)—containing approximately 395,000 and 225,000 individual curves, respectively—will present major challenges, but the investment in these large pharmacogenomic warrants such efforts.

Author A. C. Jin was a student in A.H.B.'s laboratory and left shortly after publication of the initial study, and did not participate in in the writing of this Reply. Authors Z.S., P.S. and M.F. developed the PharmacoGx software package, which enabled the analyses presented here; A.G. helped with the comparison of the different drug sensitivity metrics, and participated in the interpretation of the results and writing of this Reply.

Methods

The methods are described in detail in the Supplementary Information. The code and associated files required to reproduce this analysis are publicly available on the cdrug-rebuttals GitHub repository (https://github.com/bhklab/cdrug-rebuttals). The procedure to set up the software environment and run our analysis pipeline is provided in the Supplementary Information. This work complies with the guidelines proposed previously[13] in terms of code availability and replicability of results.

Zhaleh Safikhani[1,2], Nehme El-Hachem[3], Petr Smirnov[1], Mark Freeman[1], Anna Goldenberg[4,5], Nicolai J. Birkbak[6], Andrew H. Beck[7,8], Hugo J. W. L. Aerts[8,9,10], John Quackenbush[9,11] & Benjamin Haibe-Kains[1,2,5,12]

[1]Princess Margaret Cancer Centre, University Health Network, Toronto, Ontario M5G 2M9, Canada.
email: benjamin.haibe.kains@utoronto.ca
[2]Department of Medical Biophysics, University of Toronto, Toronto, Ontario M5G 1L7, Canada.
[3]Institut de recherches cliniques de Montréal, Montreal, Quebec H2W 1R7, Canada.
[4]Hospital for Sick Children, Toronto, Ontario M5G 1X8, Canada.
[5]Department of Computer Science, University of Toronto, Toronto, Ontario M5S 2E4, Canada.
[6]The Francis Crick Institute, University College London, London NW1 1AT, UK.
[7]Beth Israel Deaconess Medical Center, Boston, Massachusetts 02215, USA.
[8]Harvard Medical School, Boston, Massachusetts 02115, USA.
[9]Dana-Farber Cancer Institute, Boston, Massachusetts 02115, USA.
[10]Brigham and Women's Hospital, Boston, Massachusetts 02115, USA.
[11]Harvard School of Public Health, Boston, Massachusetts 02115, USA.
[12]Ontario Institute of Cancer Research, Toronto, Ontario M5G 1L7, Canada.

1. Bouhaddou, M. *et al.* Drug response consistency in CCLE and CGP. *Nature* **540,** http://dx.doi.org/10.1038/nature20580 (2016).
2. Garnett, M. J. *et al.* Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature* **483,** 570–575 (2012).
3. Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483,** 603–607 (2012).
4. Pozdeyev, N. *et al.* Integrating heterogeneous drug sensitivity data from cancer pharmacogenomic studies. *Oncotarget* http://dx.doi.org/10.18632/oncotarget.10010 (2016).
5. Mpindi, J. P. *et al.* Consistency in drug response profiling. *Nature* **540,** http://dx.doi.org/10.1038/nature20171 (2016).
6. Smirnov, P. *et al.* PharmacoGx: An R package for analysis of large pharmacogenomic datasets. *Bioinformatics* **32,** 1244–1246 (2015).
7. Haibe-Kains, B. *et al.* Inconsistency in large pharmacogenomic studies. *Nature* **504,** 389–393 (2013).
8. Safikhani, Z. *et al.* Revisiting inconsistency in large pharmacogenomic studies. *F1000Research* http://dx.doi.org/10.12688/f1000research.9611.1 (2016).
9. Matthews, B. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta* **405,** 442–451 (1975).
10. Landis, J. R. & Koch, G. G. The measurement of observer agreement for categorical data. *Biometrics* **33,** 159–174 (1977).
11. Seashore-Ludlow, B. *et al.* Harnessing connectivity in a large-scale small-molecule sensitivity dataset. *Cancer Discov.* **5,** 1210–1223 (2015).
12. Iorio, F. *et al.* A landscape of pharmacogenomic interactions in cancer. *Cell* **166,** 740–754 (2016).
13. Sandve, G. K., Nekrutenko, A., Taylor, J. & Hovig, E. Ten simple rules for reproducible computational research. *PLoS Comput. Biol.* **9,** e1003285 (2013).

doi:10.1038/nature20581