SUPPLEMENTARY FILE 1

Iterations of the Brief Communication Arising, reviews and responses

Drug Response Consistency in CCLE and CGP

Mehdi Bouhaddou[1], Matthew S. DiStefano[1], Eric A. Riesel[1], T. Victoria Thompson[1], Emilce Carrasco[1], SreeHarish Muppirisetty[1], Marc R. Birtwistle[1,2,3,#]

[1]Department of Pharmacology and Systems Therapeutics, Icahn School of Medicine at Mount Sinai, New York, NY 10029

[2]DeTOX LINCS Center, Icahn School of Medicine at Mount Sinai, New York, NY 10029

[3]Systems Biology Center New York, Icahn School of Medicine at Mount Sinai, New York, NY 10029

[#r]To whom correspondence should be addressed

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 response across cell lines, providing a public resource relating biochemistry to drug sensitivity. A recent study[3] compared correlations between reported dose response metrics and found inconsistency between CCLE and CGP. This result questions the usefulness of not only these, but other current and future costly large-scale studies. We find that by asking a slightly different question—do the two studies agree for binary drug sensitivity classification—reasonable consistency is observed, in stark contrast to the previous conclusion[3]. Our results revive confidence that these important public resources may be useful for understanding molecular correlates of drug sensitivity.

Human vision is exquisitely good at pattern recognition, and may detect associations where algorithmic and quantitative assessments fail. We asked three independent people to manually evaluate consistency—in terms of drug sensitivity—for all matching cell line (285)/drug(24) pairs in the CCLE and CGP within their shared dose range (Appendix and Fig. 1a; ~75-80% consistency among curators). The data suggested that the two studies are reasonably consistent (~80%). This level of consistency is on par with that of the microarray data[3]. The data also show that cell lines are "insensitive" to most tested drugs (~65%).Characterizing such insensitive dose response data with a standard sigmoid response model meant for sensitive cell lines may lead to incorrect dataset consistency inferences. Moreover, the CCLE and CGP used different drug dose ranges, which on its own could unfairly influence consistency. We defined a simple linear dose response slope (% viability versus $\log_{10}$ dose) within the shared concentration range to compare sensitivity between the two studies (Fig. 1b); large negative slope implies sensitivity, and slope close to zero implies insensitivity. Even with a simple universal slope cutoff for sensitivity classification (slope<-16 gives approximately equal misclassification probability based on a two normal population mixture model fit), the

CCLE and CGP data exhibit far more agreement than disagreement, and are also quite consistent with the independent manual curations (Fig. 1a-b), even on the level of individual drugs (Fig. 1c). Of course, machine learning classification[1] may improve consistency inferences.

We next compared IC_{50} data only from sensitive drug/cell line combinations (in CCLE *or* CGP by slope cutoff—all but top right quadrant in Fig. 1b) with defined IC_{50} values (Methods—we re-estimated IC_{50} values to eliminate variability due to differing CCLE and CGP dose-response model assumptions). We found good correlation between the two studies (Pearson ρ=0.67, Fig. 1d). However, stratification by drug yields poor IC_{50} correlations (drugs with greater than 19 points—Fig. 1e). Why does pooling data across drugs improve correlation? Our interpretation is that these large scale studies cannot precisely determine IC_{50} for individual cell line/drug pairs, but can faithfully report on IC_{50} order of magnitudes—i.e. sensitivity vs. insensitivity. Supporting this line of reasoning, IC_{50} values averaged by drug have excellent correlation (Pearson ρ=0.92, Fig. 1d, inset). Haibe-Kains et al.[3] stratified IC_{50} by drug, which likely contributed to their conclusion of inconsistency.

We conclude that the CCLE and CGP are largely consistent for drug sensitivity classification based on dose response slope. Inferences of specific IC_{50} values would however require more detailed studies. 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 and genomic drift, different drug suppliers/batches, labs/equipment, and serum composition. This suggests that the measured genomic, mutation and gene expression parameters may provide a robust cellular context that dictates drug sensitivity.

References

1 Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603-607, doi:10.1038/nature11003

nature11003 [pii] (2012).

2 Garnett, M. J. *et al.* Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature* **483**, 570-575, doi:10.1038/nature11005

nature11005 [pii] (2012).

3 Haibe-Kains, B. *et al.* Inconsistency in large pharmacogenomic studies. *Nature* **504**, 389-393, doi:10.1038/nature12831

nature12831 [pii] (2013).

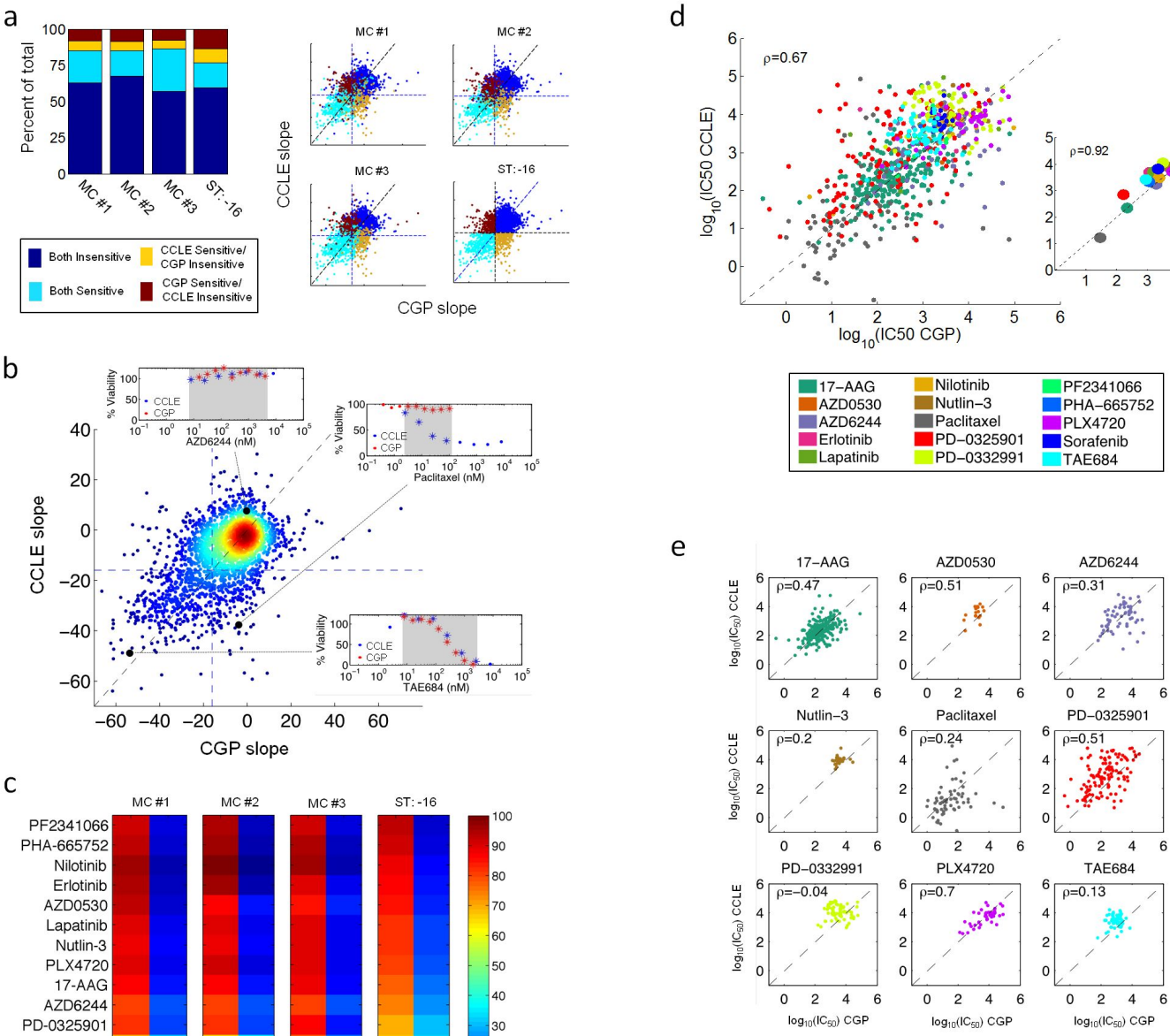
Methods

For each cell line/drug pair found in both CCLE and CGP, we calculated the slope of each dose response curve (cell viability data converted to a common 0-100% scale) only in the shared dose range using a linear regression and the $\log_{10}$ drug dose. One CCLE point and one CGP point defined boundaries of the shared dose range in a way that maximized 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 between 10^{-1} and 10^{5} nM, and were otherwise undefined.

Figure Legend

Figure 1. Consistency Between Pharmacological Data in CCLE and CGP

a, Three people independently evaluated consistency of drug sensitivity for all overlapping cell line/drug combinations (2520) within the shared drug dose range. Manual curators categorized each as either (*i*) both insensitive, (*ii*) both sensitive, (*iii*) CCLE sensitive/CGP insensitive, or (*iv*) CGP sensitive/CCLE insensitive. They were given minimal instruction, other than to look for a downward trend that they felt extended robustly past 50% for sensitivity. They were only given data points in the shared range. MC = Manual Curator; ST = Slope Threshold. All dose response curves and their manually curated classification results are provided in the Appendix. **b,** Linear slopes (based on $\log_{10}$ drug concentration) were calculated over the shared drug dose range for the 2520 overlapping cell line/drug pairs. Color indicates density of dots. The black dashed line is $x=y$. The blue dashed lines depict our slope threshold ($y=-16;x=-16$), which segregates the data into quadrants: upper right—consistent and insensitive; lower left—consistent and sensitive; top left—inconsistent, sensitive in CGP and insensitive in CCLE; bottom right—inconsistent, insensitive in CGP and sensitive in CCLE. In example dose response curves, stars and shading represent the shared dose range. Cell lines of examples in top right, bottom right, and bottom left quadrants are AU565 (breast cancer), U-87-MG (glioma), and SIG-M5 (leukemia), respectively. The slope cutoff was determined by analysis of a fit of the distribution of slopes to a mixture model of two normally distributed populations (one model for each study), one population with mean near zero as insensitive and one population with larger negative mean as sensitive. The point of intersection between the pdfs of these two underlying populations corresponds to the point at which there is equal probability of misclassifying a member of the sensitive population as insensitive, and vice versa. This point of intersection was slope~-16 in both studies. **c,** Data from Fig. 1a-b stratified by drug. Percent consistent is the sum of the upper right and lower left quadrants; percent inconsistent is the sum of the upper left and lower right quadrants. **d,** IC_{50} values from all sensitive cell line/drug combinations as determined by slope threshold, which includes those found in the lower left, upper left, and lower right quadrants, i.e. sensitive in either CCLE or CGP, and also were determined to have well-defined IC_{50} values (between 0.1 nM and 100 μM, see Methods). The black dashed line is $x=y$. Inset: IC_{50} values averaged by drug. **e,** IC_{50} values from all sensitive cell line/drug pairs (same as in Fig. 1d) stratified by drug, for drugs having greater than 19 points. All correlation coefficients are Pearson.


a
Percent of total
MC #1
MC #2
MC #3
ST: -16
Both Insensitive
CCLE Sensitive/ CGP Insensitive
Both Sensitive
CGP Sensitive/ CCLE Insensitive
CCLE slope
CGP slope
b
% Viability
CCLE
CGP
AZD6244 (nM)
Paclitaxel (nM)
TAE684 (nM)
c
PF2341066
PHA-665752
Nilotinib
Erlotinib
AZD0530
Lapatinib
Nutlin-3
PLX4720
17-AAG
AZD6244
PD-0325901
Paclitaxel
Sorafenib
PD-0332991
TAE684
% Consistent
% Inconsistent
d
ρ=0.67
ρ=0.92
log10(IC50 CCLE)
log10(IC50 CGP)
17–AAG
AZD0530
AZD6244
Erlotinib
Lapatinib
Nilotinib
Nutlin–3
Paclitaxel
PD–0325901
PD–0332991
PF2341066
PHA–665752
PLX4720
Sorafenib
TAE684
e
17–AAG
ρ=0.47
AZD0530
ρ=0.51
AZD6244
ρ=0.31
Nutlin–3
ρ=0.2
Paclitaxel
ρ=0.24
PD–0325901
ρ=0.51
PD–0332991
ρ=–0.04
PLX4720
ρ=0.7
TAE684
ρ=0.13
log10(IC50) CCLE
log10(IC50) CGP

Appendix

Drug Response Consistency in CCLE and CGP

Mehdi Bouhaddou[1], Matthew S. DiStefano[1], Eric A. Riesel[1], T. Victoria Thompson[1], Emilce Carrasco[1], SreeHarish Muppirisetty[1], Marc R. Birtwistle[1,2,3,#]

[1]Department of Pharmacology and Systems Therapeutics, Icahn School of Medicine at Mount Sinai, New York, NY 10029

[2]DeTOX LINCS Center, Icahn School of Medicine at Mount Sinai, New York, NY 10029

[3]Systems Biology Center New York, Icahn School of Medicine at Mount Sinai, New York, NY 10029

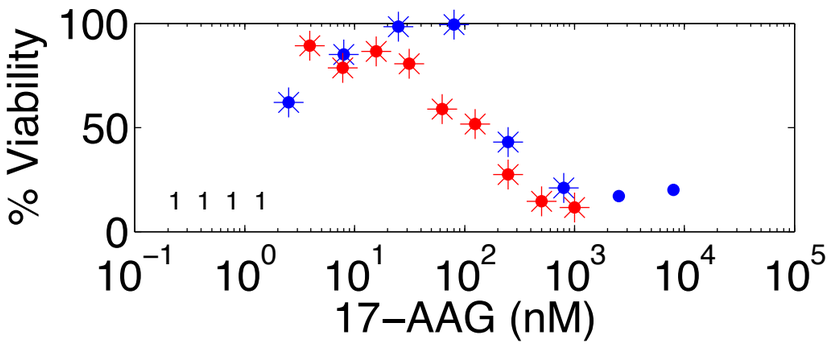
[#]To whom correspondence should be addressed

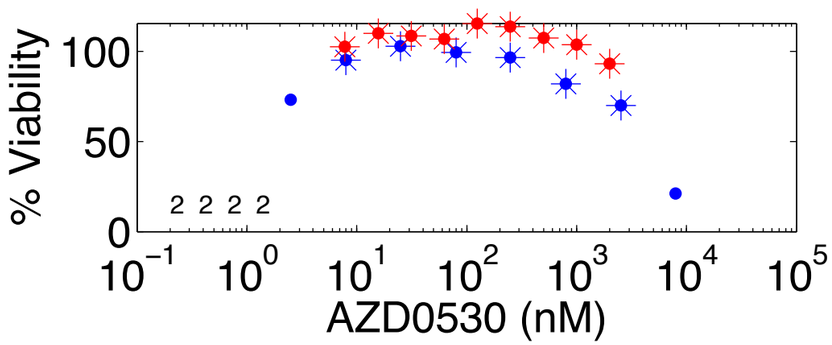
Appendix. Dose response curves for all shared cell line/drug pairs in CCLE and CGP.

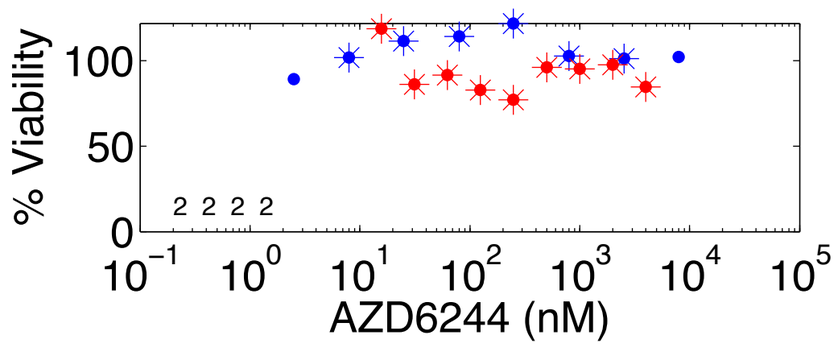
Each page depicts a different cancer cell line, the name of which is given at the top of each page. On each page, all of the drugs that were tested by both CCLE and CGP are displayed. Drug dose is on the x-axis and percent cell viability is on the y-axis. Blue corresponds to CCLE and red to CGP. Stars indicate the shared dose range between the two studies. The numbers in the bottom left corner of each subplot signify the sensitivity categorization given by Manual Curator #1, Manual Curator #2, Manual Curator #3, and by the Slope Threshold (-16), respectively. 1= Both Sensitive; 2= Both Insensitive; 3= CCLE Sensitive/ CGP Insensitive; 4= CGP Sensitive/ CCLE Insensitive.

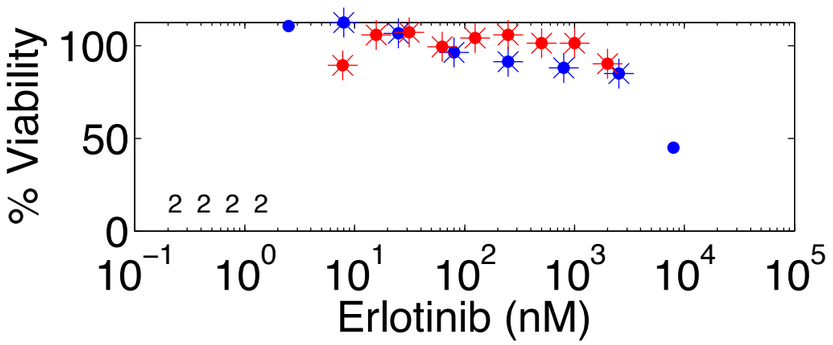
8–MG–BA

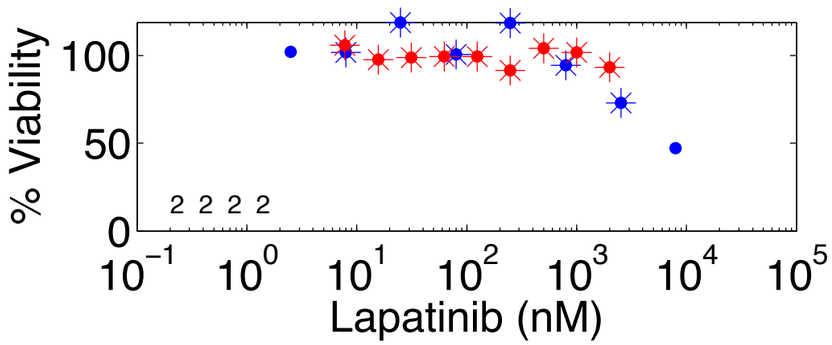
CCLE
CGP

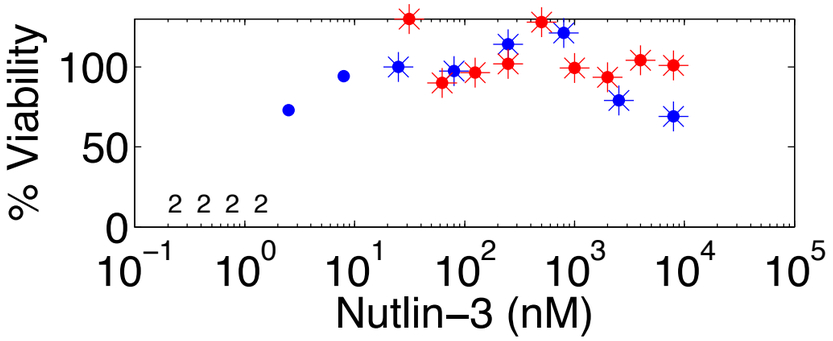


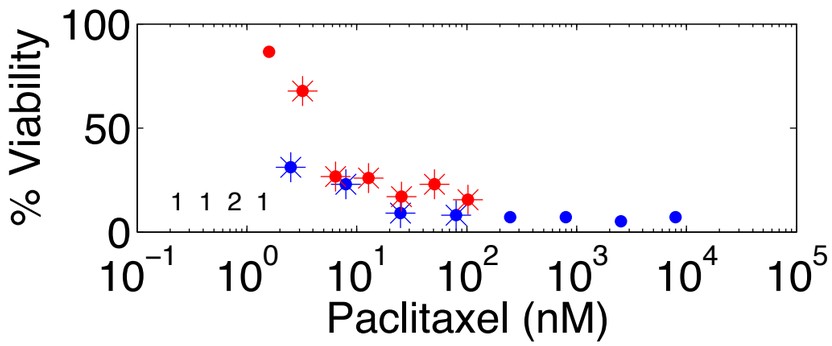


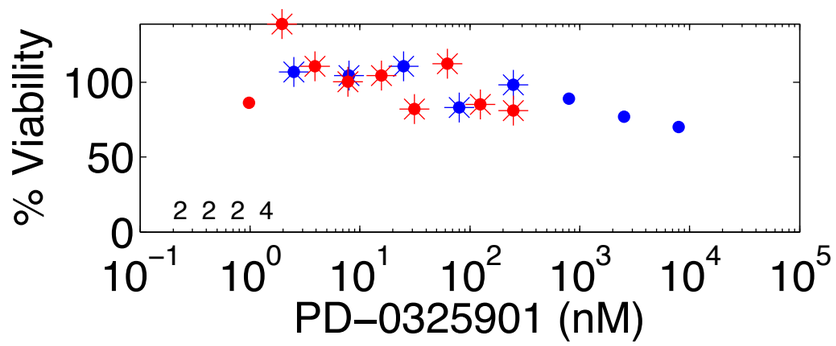


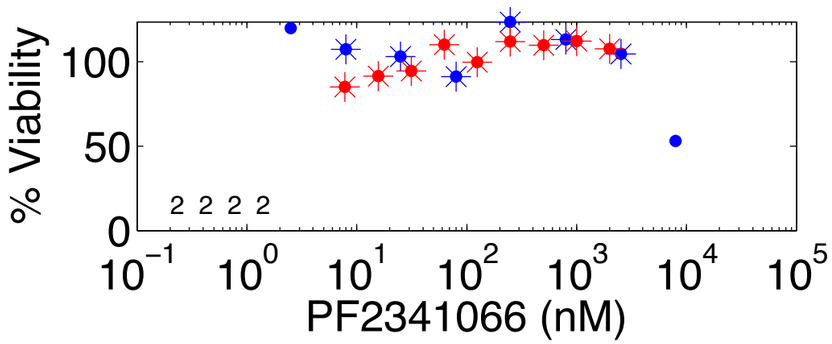


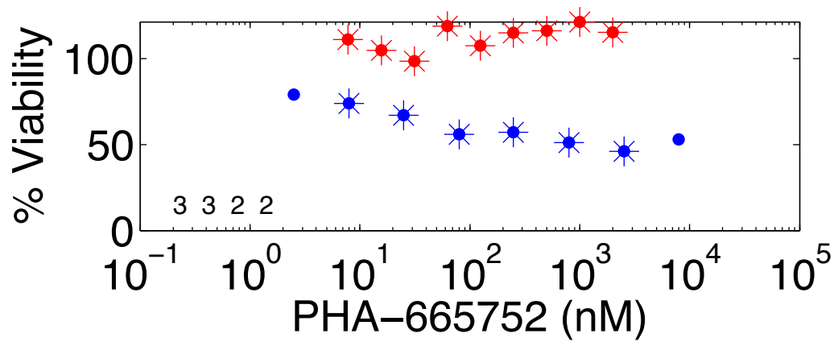


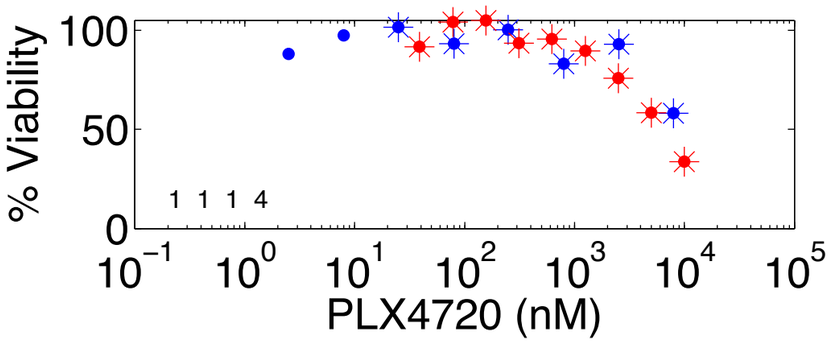


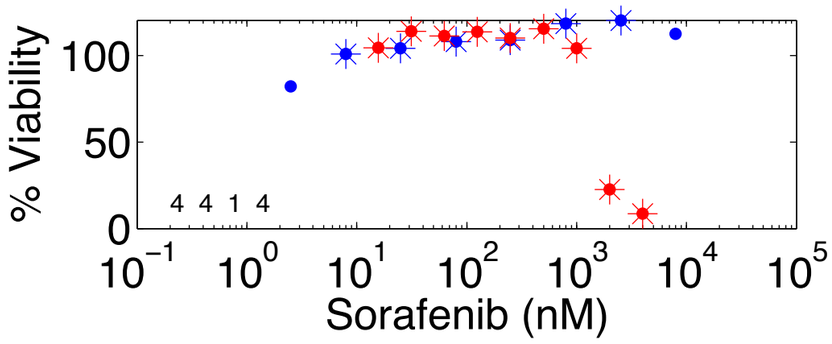


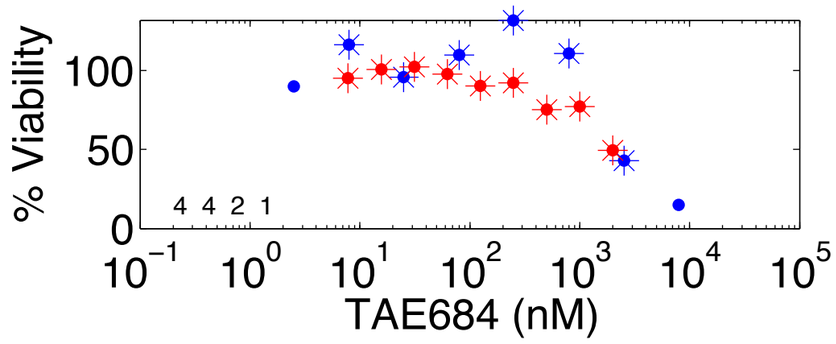












Response to "Drug Response Consistency in CCLE and CGP"

Zhaleh Safikhani, Petr Smirnov, Nehme El-Hachem, Nicolai Juul Birkbak, Andrew H. Beck, Hugo J.W.L. Aerts, John Quackenbush, Benjamin Haibe-Kains

We welcome the opportunity to respond to the report by Bouhaddou et al. The authors tested whether binary drug sensitivity calls (insensitive vs. sensitive), as determined by manual curators and by the linear dose response slope (ST), were consistent between CGP [1] and CCLE [2]. While we appreciate the innovative approach followed by the authors, our re-analysis and interpretation of their results do not support their claim that "CCLE and CGP are largely consistent for drug sensitivity based on dose response slope". Indeed we demonstrated that the dose response slope statistic exhibits the same level of (in)consistency than reported in our initial publication still stand [3]. However the manual classification of the drug dose response curves, if correctly carried out, holds the potential of improving consistency across studies, although it remains unknown whether this approach will improve reproducibility of the biomarker discovery process.. Below we provide our response to the main results reported by Bouhaddou et al.

Overall percentage agreement is not a good measure of consistency. The authors assessed consistency using overall percentage of agreement. However such a statistic does not take into account the agreement that would be expected purely by chance. Only agreement beyond that expected by chance can be considered "true" agreement. Cohen's Kappa (κ) is such a measure of "true" agreement [4], as it indicates the proportion of agreement beyond that expected by chance [5]. We used κ in our initial publication [3] to assess the agreement of drug sensitivity calls based on the Waterfall approach [2]. Based on the standards for strength of agreement for κ defined by Landis and Koch [6], which are ≤ 0□ poor, [0.01, 0.20□] slight, [0.21–0.40□] fair, [0.41–0.60□] moderate,

[0.61–0.80□] substantial, and [0.81–1□] almost perfect, we observed that only two drugs yielded moderate agreement (AZD0530 and AZD6244), while 6 and 7 drugs yielded fair and slight agreement, respectively (Supplementary Figure 5 of our initial publication [3]). Thanks to the authors' code, we were able to reproduce and analyse their new binary sensitivity classification schemes by using κ as a measure of consistency (Supplementary Figure 1). When data are pooled for all drugs, we observed that the drug sensitivity calling defined by the the curators yielded substantial agreement between CGP and CCLE, a major improvement over simple dichotomization based on AUC > 0.2 (Supplementary Figure 1A). However, the ST method does not outperform the simple AUC dichotomization scheme and yielded only fair agreement across studies (Supplementary Figure 1A).

Potential bias in manual evaluation of drug dose-response curves. Based on the rationale that human vision may outperform computer algorithms at pattern recognition tasks, the authors asked three manual curators to assess the drug sensitivity calls of all the cell line/drug pairs screened in CGP and CCLE. This simple approach has the potential to yield better phenotypic measurements if carried out in a rigorous way. The curators should assess drug sensitivity for each study separately in a blinded manner (for each cell line/drug pair, a single curve should be displayed, restricted to the shared concentration range, with the data source being hidden). Instead, the authors asked the curators to assess consistency between CGP and CCLE by evaluating whether the two curves, with their full concentration range available, are either both sensitive, both insensitive or inconsistent. Such an evaluation is likely to be biased as the curators are visualizing to the two curves simultaneously before taking their decision. This constitutes an information leak that may lead to overoptimistic estimates of consistency, which could partially explain the significant superiority of curators compared to computational schemes. It is therefore crucial to further confirm these promising results using an appropriate blinded experimental design.

Manual classification is significantly more consistent than computational schemes. To test whether the level of consistency observed from the data pooled across drugs translates to individual drugs, we computed the κ coefficient for all the binary sensitivity calling schemes for each drug separately (Supplementary Figure 2). The curators yielded almost perfect agreement for AZD0530, substantial agreement for 7 drugs (erlotinib, lapatinib, nilotinib, sorafenib, PLX4720, PD-0325901, nutlin-3, 17-AAG), moderate and fair agreement for 5 (PHA665752, crizotinib, PD-0332991, AZD6244) and 2 (paclitaxel, TAE684) drugs, respectively. Although the level of consistency varied considerably across curators, their classification of drug dose-response curves was significantly more reproducible than AUC dichotomization and ST (Wilcoxon rank sum test $p < 0.008$). This is in contrast with ST and AUC dichotomization schemes which yielded only slight or fair agreement for most drugs.

Limitations of the dose response slope for drug sensitivity calling. We further investigated the use of the dose response slope to classify cell line/drug pair into insensitive and sensitive categories. By carefully looking at the drug dose-response curves, we noticed an unexpected behaviour of the ST statistic in cases where the drug actually induced high level of growth inhibition at all drug concentrations, as for the EM-2 cell line treated with nilotinib (Supplementary Figure 3A) and the JVM-3 cell line treated with paclitaxel (Supplementary Figure 3B). In this situation, ST will be close to 0, classifying the cell line as insensitive, while it is actually extremely sensitive to the drug of interest. Other curves have unusual shapes (Supplementary Figure 4), probably resulting from normalization issues or experimental artifacts, which prevents proper computation of the dose response slope. Although these cases are relatively rare, they illustrate the limitations of the authors' approach and call into question the robustness of the ST statistic.

Consistency of drug sensitivity data for pooled versus individual drugs. By pooling drug sensitivity data across drugs, the authors noticed an improved consistency of IC_{50} values between CGP and CCLE, which was not confirmed at the level of individual drugs, which seems in stark contrast with the title of the authors' manuscript. The authors interpret their results by hypothesizing that the "[CGP and CCLE] studies cannot precisely determine IC_{50} for individual cell line/drug pairs, but can faithfully report on sensitivity vs. insensitivity categories". However, this interpretation seems contradictory with the low consistency observed for the ST sensitivity calling in multiple individual drugs (Supplementary Figure 2). The authors averaged IC_{50} values, which yielded a high correlation between CGP and CCLE. These results confirm that drugs with broad inhibitory effects yield much lower average IC_{50} than drugs more targeted effects in both studies. This observation is quite trivial and surely does not warrant the investment in these large pharmacogenomic studies. The main goal of CGP and CCLE consisted in finding new associations between molecular features and drug sensitivity to specific drugs. The authors did not demonstrate that their drug sensitivity calling approaches -- manual curation or dose response slope -- actually improved the reproducibility of the biomarker discovery process.

In conclusion, our re-analysis of the authors' ST method showed that it yields similar levels of (in)consistency than those reported in our initial publication [3]. However, manual classification of the drug dose-response curves seems to substantially improve the consistency of binary sensitivity calls but these results must be confirmed in a blinded setting. Importantly, the authors did not demonstrate that binary drug sensitivity classification improves reproducibility of biomarker discovery for individual drugs, which was the primary goal of the CGP and CCLE studies. Several challenges will still remain, as this approach is rather low-throughput while CGP and CCLE datasets contain approximately 80,000 and 12,000 individual curves, but the investment in these large pharmacogenomic certainly warrant such efforts.

Additional comments

The authors stated that “[drug sensitivity calling] level of consistency is on par with that of the microarray data”. The authors did not dichotomize the gene expression data to test whether drug sensitivity calls are as concordant as expression data between CGP and CCLE. Further analyses should be performed to support such a claim.

The authors stated that “CCLE and CGP used different drug dose ranges, which on its own could unfairly influence consistency”. The meaning of “unfairly” is unclear. The goal of our comparative study was to quantitatively assess the consistency of the data generated across CGP and CCLE. The fact that they used different concentration ranges and pharmacological assays is part of their experimental protocols and should be considered when assessing consistency. Moreover, the vast majority of the scientific community would rely on the data released by the two groups instead of re-estimating all the statistics from the raw sensitivity data by throwing away data points that might prove valuable for the estimation of the drug dose-response curve. However we agree that the use of different concentration ranges in an important factor of inconsistency, one of the few that can be accounted for by computational means. We evaluated the gain in consistency of the dose response slope estimated from the full and shared concentration range (Supplementary Figures 1 and 2) but the gain was not significant (one sided Wilcoxon rank sum $p = 0.72$).

The authors stated that “Of course, machine learning classification (ref 1) may improve consistency inferences.” It is not clear why the CCLE publication is provided as reference to support the potential of machine learning algorithms to improve estimation or pattern recognition of the drug dose-response curve.

References

1. Garnett, M. J. *et al.* Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature* **483,** 570–575 (2012).
2. Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483,** 603–607 (2012).
3. Haibe-Kains, B. *et al.* Inconsistency in large pharmacogenomic studies. *Nature* **504,** 389–393 (2013).
4. Cohen, J. A Coefficient of Agreement for Nominal Scales. *Educ. Psychol. Meas.* **20,** 37–46 (1960).
5. Sim, J. & Wright, C. C. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Phys. Ther.* **85,** 257–268 (2005).
6. Landis, J. R. & Koch, G. G. The measurement of observer agreement for categorical data. *Biometrics* **33,** 159–174 (1977).

Supplementary Information

Response to

”Drug Response Consistency in CCLE and CGP”

from Bouhaddou et al.

Zhaleh Safikhani, Petr Smirnov, Nehme El-Hachem, Nicolai Juul Birkbak,
Andrew H. Beck, Hugo J.W.L. Aerts, John Quackenbush, Benjamin Haibe-Kains

Contents

1 Supplementary Methods

1.1 Data retrieval and curation

1.1.1 CGP (release 4, March 2013)

Gene expression, mutation data and cell line annotations were downloaded from ArrayExpress. Drug sensitivity measurements and drug information were downloaded from the CGP website (link) and the Nature website (link), respectively.

Minimum and maximum screening concentrations for each drug/cell line were extracted from `gdsc_compounds_conc_w2.csv` (μM).

The natural logarithm of IC_{50} measurements were retrieved from `gdsc_manova_input_w2.csv` in column `"*_IC_50"` (referred to as x) and subsequently transformed using $-\log_{10}(\exp(x))$; high values are representative of cell line sensitivity to drugs.

The AUC measurements were retrieved from `gdsc_manova_input_w2.csv` in column `"*_AUC"` (referred to as x); high values are representative of cell line sensitivity to drugs.

1.1.2 CCLE (release March 2013)

Gene expression, mutation data cell line annotations and drug information were downloaded from the CCLE website (link) Drug sensitivity data were downloaded from the Nature website (link);

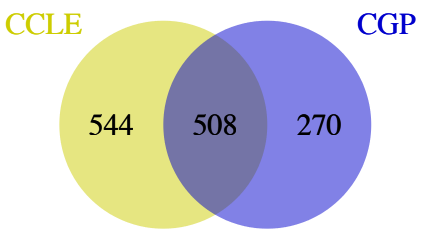
Screening concentrations for each drug/cell line were extracted from `CCLE_NP24.2009_Drug_data_2012.02.20.csv` (μM).

IC_{50} measurements were retrieved from `CCLE_NP24.2009_Drug_data_2012.02.20.csv` ("IC_{50} μM `(norm)`") (referred to as x) and subsequently transformed into logarithmic scale, $-\log_{10}(x)$; high values are representative of cell line sensitivity to drugs.

AUC measurements were retrieved from `CCLE_NP24.2009_Drug_data_2012.02.20.csv` ("`ActArea (norm)`") and subsequently divided by the number of drug concentrations tested (8); high values are representative of cell line sensitivity to drugs.

1.2 Cell line annotations

Cell line names were harmonized in both CGP and CCLE to match identical cell lines; this was done through manual search over alternative names of cell lines, as reported in CGP and CCLE cell line annotation files and online databases such as hyperCLDB and BioInformationWeb. In our comparative analysis published in Nature [3], we focused on the set of 471 cell lines for which both gene expression and drug sensitivity were available. In the present work we extended our curation to all the cell lines for which at least one data type (gene expression, mutation or drug sensitivity) is available, increasing the shared set of cell lines to 508:



Tissue type nomenclature from CGP [2] was chosen throughout this study, CCLE tissue type information [1] was therefore updated to follow this nomenclature, which resulted in 24 tissue types:

Tissue type	Number of cell lines
lung	109
haematopoietic_and_lymphoid_tissue	77
breast	39
central_nervous_system	36
large_intestine	33
skin	30
oesophagus	21
urinary_tract	17
ovary	16
pancreas	16
stomach	16
autonomic_ganglia	12
soft_tissue	12
upper_aerodigestive_tract	12
liver	11
kidney	10
bone	8
endometrium	8
thyroid	8
pleura	6
prostate	4
biliary_tract	1
salivary_gland	1
small_intestine	1

All the curation steps have been documented in the scripts `cdrug2_normalization_cgp.R`, `cdrug2_normalization_ccle.R`, and `cdrug2_format.R`.

2 Full Reproducibility of the Analysis Results

We will describe how to fully reproduce the figures and tables reported in the main manuscript. We automated the analysis pipeline so that minimal manual interaction is required to reproduce our results. To do this, one must simply:

1. Set up the software environment
2. Run the R scripts
3. Generate the Supplementary Information

The code and associated files are publicly available on GitHub: https://github.com/bhaibeka/cdrug2.

2.1 Set up the software environment

We developed and tested our analysis pipeline using R running on linux and Mac OSX platforms.

To mimic our software environment the following R packages should be installed:

- R version 3.1.1 (2014-07-10), `x86_64-unknown-linux-gnu`
- Base packages: base, datasets, graphics, grDevices, grid, methods, parallel, splines, stats, utils
- Other packages: affxparser 1.36.0, affy 1.42.3, affyio 1.32.0, amap 0.8-12, AnnotationDbi 1.26.0, Biobase 2.24.0, BiocGenerics 0.10.0, BiocInstaller 1.14.3, biomaRt 2.20.0, bitops 1.0-6, corpcor 1.6.6, DBI 0.3.0, epibasix 1.3, fingerprint 3.5.2, Formula 1.1-2, frma 1.16.0, gdata 2.13.3, genefu 1.14.0, GenomeInfoDb 1.0.2, gplots 2.14.1, Hmisc 3.14-5, hthgu133acdf 2.14.0, hthgu133afrmavecs 1.1.0, igraph 0.7.1, inSilicoDb 2.1.1, jetset 1.6.0, lattice 0.20-29, lsa 0.73, magicaxis 1.9.3, MASS 7.3-34, mclust 4.3, MetaGx 0.0.2, mgcv 1.8-3, mRMRe 2.0.5, nlme 3.1-117, OptimalCutpoints 1.1-3, org.Hs.eg.db 2.14.0, PharmacoGx 0.0.3, plotrix 3.5-7, prodlim 1.4.5, rcdk 3.3.0, RColorBrewer 1.0-5, RCurl 1.95-4.3, rjson 0.2.14, R.methodsS3 1.6.1, R.oo 1.18.0, RSQLite 0.11.4, R.utils 1.33.0, sm 2.2-5.4, SnowballC 0.5.1, stringr 0.6.2, survcomp 1.15.1, survival 2.37-7, sva 3.10.0, vcd 1.3-2, VennDiagram 1.6.8, WriteXLS 3.5.0, XML 3.98-1.1, xtable 1.7-4
- Loaded via a namespace (and not attached): acepack 1.3-3.3, Biostrings 2.32.1, bit 1.1-12, bootstrap 2014.4, caTools 1.17.1, cluster 1.15.3, codetools 0.2-9, colorspace 1.2-4, ff 2.2-13, foreach 1.4.2, foreign 0.8-61, GenomicRanges 1.16.4, gtools 3.4.1, IRanges 1.22.10, iterators 1.0.7, KernSmooth 2.23-13, latticeExtra 0.6-26, lava 1.2.6, Matrix 1.1-4, nnet 7.3-8, oligo 1.28.2, oligoClasses 1.26.0, png 0.1-7, preprocessCore 1.26.1, rcdklibs 1.5.8.4, rJava 0.9-6, rmeta 2.16, rpart 4.1-8, stats4 3.1.1, SuppDists 1.1-9.1, survivalROC 1.0.3, tcltk 3.1.1, tools 3.1.1, XVector 0.4.0, zlibbioc 1.10.0

All these packages are available on CRAN[1] or Bioconductor[2], except for jetset which is available on the CBS website[3].

Run the following commands in a R session to install all the required packages:

[1] http://cran.r-project.org
[2] http://www.bioconductor.org
[3] http://www.cbs.dtu.dk/biotools/jetset/

```
source("http://bioconductor.org/biocLite.R")
biocLite(c("AnnotationDbi", "affy", "affyio", "hthgu133acdf",
 "hthgu133afrmavecs", "hgu133plus2cdf", "hgu133plus2frmavecs",
 "org.Hs.eg.db", "genefu", "biomaRt", "frma", "Hmisc", "vcd",
 "epibasix", "amap", "gdata", "WriteXLS", "xtable", "plotrix",
 "R.utils", "DBI", "GSA", "gplots", "magicaxis"))
```

Note that you may need to install Perl[4] and its module Text::CSV_XS for the WriteXLS package to write xls file; once Perl is installed in your system, use the following command to install the Text::CSV_XS module through CPAN[5]:

```
cpan Text/CSV_XS.pm
```

Lastly, follow the instructions on the CBS website to properly install the jetset package or use the following commands in R:

```
download.file(url="http://www.cbs.dtu.dk/biotools/jetset/current/jetset_1.4.0.tar.gz",
 destfile="jetset_1.4.0.tar.gz")
install.packages("jetset_1.4.0.tar.gz", repos=NULL, type="source")
```

Once the packages are installed, clone the `cdrug2` GitHub repository (https://github.com/bhaibeka/cdrug2) This should create a directory on the file system containing the following files that are relevant to reproduce our results presented here:

`cdrug2_foo.R` Script containing the definitions of all functions required for the analysis pipeline.

`cdrug2_birtwistle.R` The R code used to perform all the analyses performed in our response.

All the files required to run the automated analysis pipeline are now in place. It is worth noting that raw gene expression and drug sensitivity data are voluminous, please ensure that at least 25GB of storage are available.

2.2 Run the R scripts

Open a terminal window and go to the `cdrug2` directory. You can easily run the analysis pipeline either in batch mode or in a R session. Before running the pipeline you can specify the number of CPU cores you want to allocate to the analysis (by default only 1 CPU core will be used). To do so, open the script `cdrug2_pipeline.R` and update line #41:

```
nbcore <- 4
```

to allocate four CPU cores for instance.

To run the full pipeline in batch mode, simply type the following command:

```
R CMD BATCH code/cdrug2_pipeline.R Rout &
```

The progress of the pipeline could be monitored using the following command:

```
tail -f Rout
```

To run the full analysis pipeline in an R session, simply type the following command:

```
source("code/cdrug2_pipeline.R")
```

Key messages will be displayed to monitor the progress of the analysis.

The analysis pipeline was developed so that all intermediate analysis results are saved in the directories `data` and `saveres`. Therefore, in case of interruption, the pipeline will restart where it stopped.

[4] http://www.perl.org/get.html
[5] http://www.cpan.org/modules/INSTALL.html

2.3 Additional parameters

download.method Method to download all the data from the CCLE and CGP websites; options are 'wget' (default), 'curl', "lynx', or 'auto'.

cosmic.version Version of the COSMIC database ('v68' as for 2014-04-07; see Sanger's FTP)

saveres Path to the directory where all teh results should be stored. Default is "./saveres"

max.cellfiles Maximum number of CEL files taht can be processed at once by `frma`.

minsample Minimum number of samples to compute the Spearman correlation. Default is 10.

genedrugm Method to estimate the association gene-drug, controlled for tissue type. Method "lm" = linear regression or logistic regression, depending on the output variable

concordance.method Estimator for concordance between gene-drug associations; possible options are "spearman" (default), "cosine", and "pearson" for Spearman's rank-ordered correlation coefficient, cosine similarity and Pearson's correlation coefficient, respectively.

2.4 Generate the Supplementary Information

After completion of the analysis pipeline a directory `saveres` will be created to contain all the intermediate results, tables and figures reported in the main manuscript and this Supplementary Information.

3 List of Abbreviations

AUC	Area under the drug sensitivity curve.
CGP	Cancer Genome Project initiated by the Wellcome Sanger Institute.
CCLE	The Cancer Cell Liners Encyclopedia initiated by Novartis and the Broad Institute.
IC_{50}	Concentration in micro molar [μM] at which the drug inhibited 50% of the cellular growth.
FDR	False Discovery Rate
R_s	Spearman correlation coefficient
ST	Slope of the drug dose-response curve for the concentration range shared between CGP and CCLE.
ST*	Slope of the drug dose-response curve for the full concentration range in CGP and CCLE.

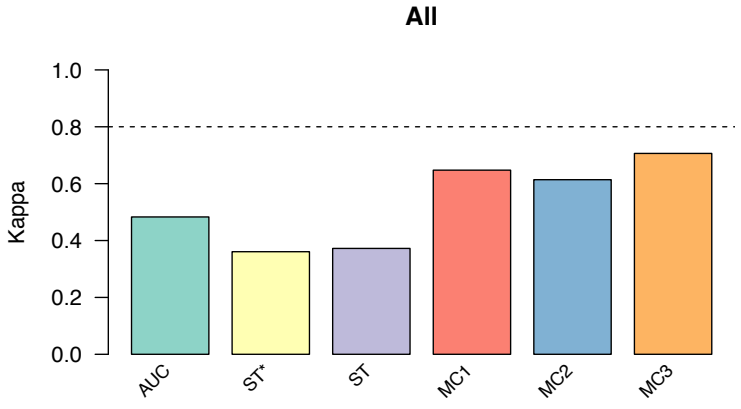
4 Supplementary Tables

Supplementary Table 1: Description of the 15 anticancer drugs screened both in CGP and CCLE studies.

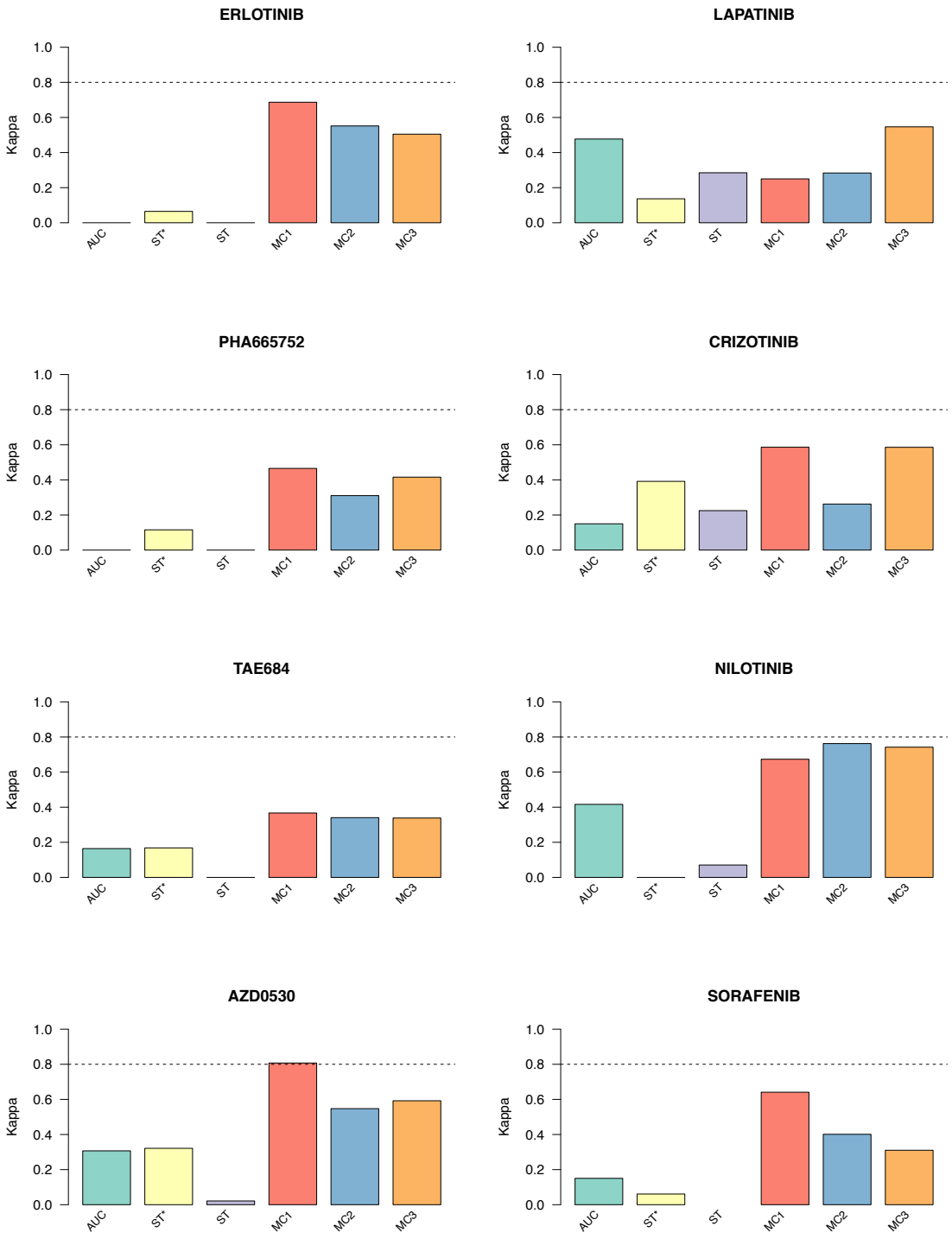
Compound	Class	Target(s)	Class	Organization
Erlotinib	Targeted	EGFR	Kinase inhibitor	Genentech
Lapatinib	Targeted	EGFR, HER2	Kinase inhibitor	GlaxoSmithKline
PHA-665752	Targeted	c-MET	Kinase inhibitor	Pfizer
Crizotinib	Targeted	c-MET, ALK	Kinase inhibitor	Pfizer
TAE684	Targeted	ALK	Kinase inhibitor	Novartis
Nilotinib	Targeted	Abl/Bcr-Abl	Kinase inhibitor	Novartis
AZD0530	Targeted	Src, Abl/Bcr-Abl, EGFR	Kinase inhibitor	AstraZeneca
Sorafenib	Targeted	Flt3, C-KIT, PDGFRbeta, RET, Raf kinase B, Raf kinase C, VEGFR-1, KDR, FLT4	Kinase inhibitor	Bayer
PD-0332991	Targeted	CDK4/6	Kinase inhibitor	Pfizer
PLX4720	Targeted	RAF	Kinase inhibitor	Plexxikon
PD-0325901	Targeted	MEK	Kinase inhibitor	Pfizer
AZD6244	Targeted	MEK	Kinase inhibitor	AstraZeneca
Nutlin-3	Targeted	MDM2	Other	Roche
17-AAG	Targeted	HSP90	Other	Bristol-Myers Squibb
Paclitaxel	Cytotoxic	beta-tubulin	Cytotoxic	Bristol-Myers Squibb

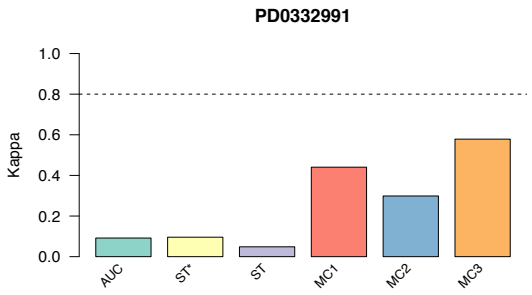
5 Supplementary Figures

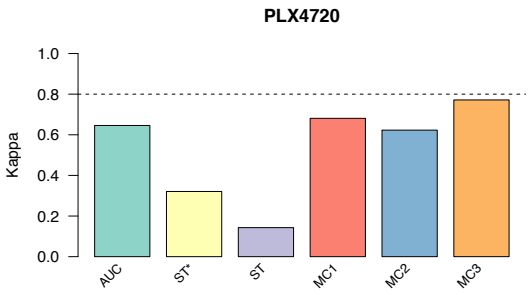
Supplementary Figure 1: Consistency of drug sensitivity calling using 6 different methods for pooled data across drugs. The drug sensitivity calling methods are the following: MC1, MC2, and MC3 represent the manual curators; ST and ST* represent the dose response slope estimation from the shared and full concentration range, respectively; and AUC represents the simple dichotomization based on AUC $>$ 0.2.

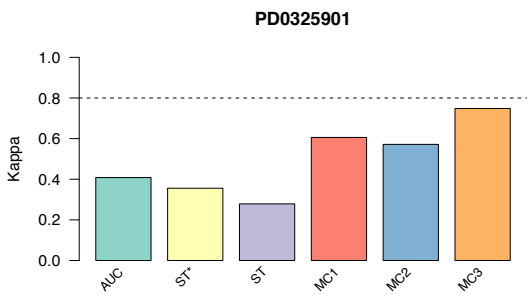


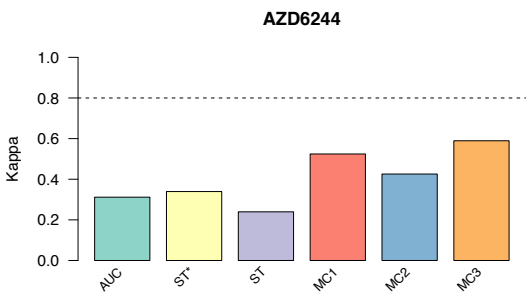
Supplementary Figure 2: Consistency of drug sensitivity calling using 6 different methods for each drug separately. The drug sensitivity calling methods are the following: MC1, MC2, and MC3 represent the manual curators; ST and ST* represent the dose response slope estimation from the shared and full concentration range, respectively; and AUC represents the simple dichotomization based on AUC $>$ 0.2.

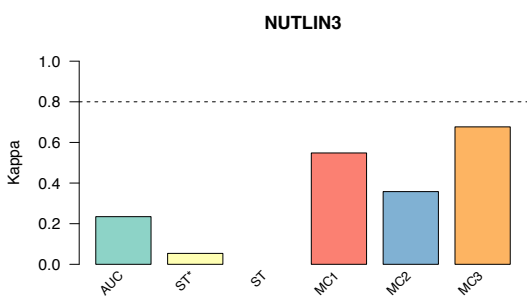


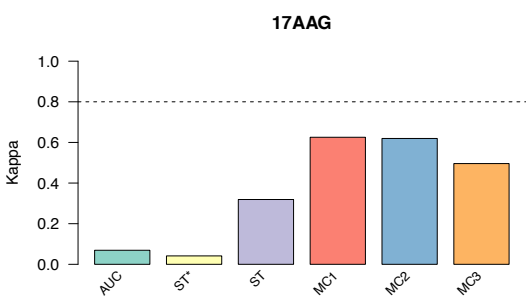

PD0332991
Kappa
1.0
0.8
0.6
0.4
0.2
0.0
AUC
ST*
ST
MC1
MC2
MC3

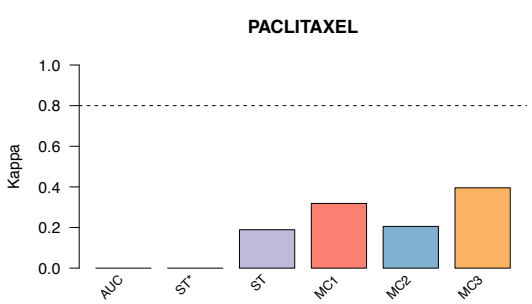

PLX4720
Kappa
1.0
0.8
0.6
0.4
0.2
0.0
AUC
ST*
ST
MC1
MC2
MC3


PD0325901
Kappa
1.0
0.8
0.6
0.4
0.2
0.0
AUC
ST*
ST
MC1
MC2
MC3

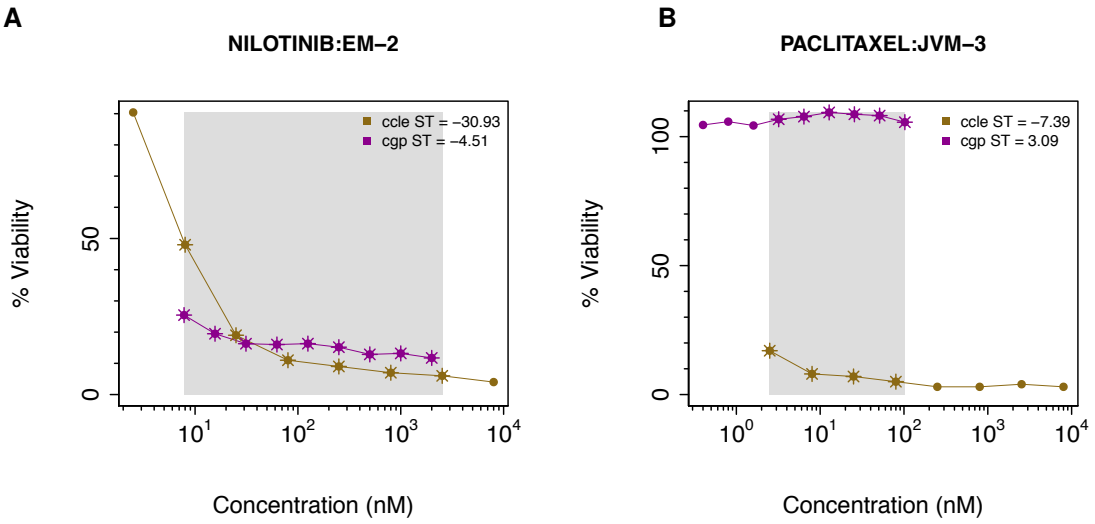

AZD6244
Kappa
1.0
0.8
0.6
0.4
0.2
0.0
AUC
ST*
ST
MC1
MC2
MC3


NUTLIN3
Kappa
1.0
0.8
0.6
0.4
0.2
0.0
AUC
ST*
ST
MC1
MC2
MC3

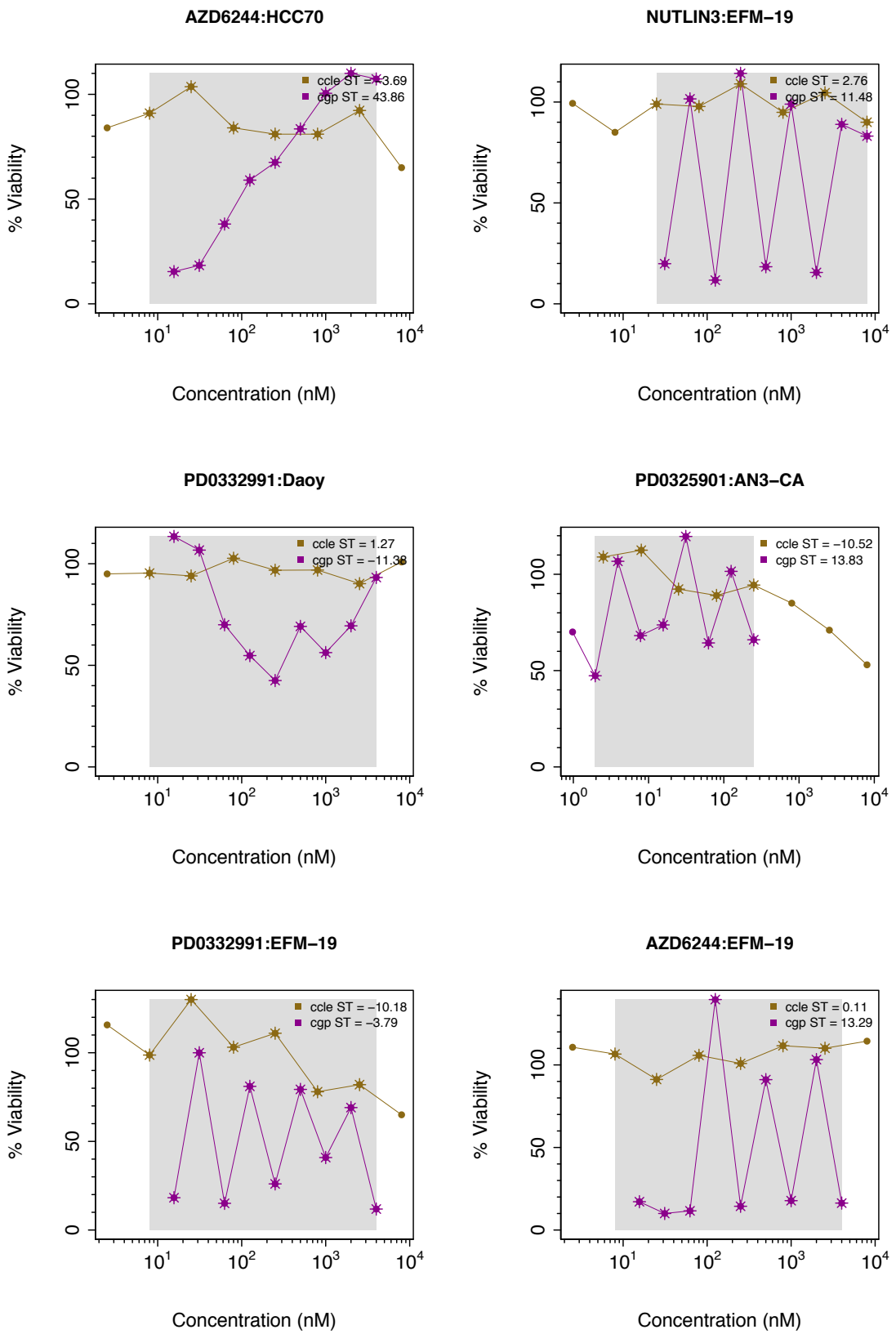

17AAG
Kappa
1.0
0.8
0.6
0.4
0.2
0.0
AUC
ST*
ST
MC1
MC2
MC3


PACLITAXEL
Kappa
1.0
0.8
0.6
0.4
0.2
0.0
AUC
ST*
ST
MC1
MC2
MC3

Supplementary Figure 3: Problematic cases where ST is close to 0 while the drug dose response curves indicate strong sensitivity. (**A**) EM-2 cell line treated with nilotinib where the CGP curve should have been classified as sensitive. (**B**) JVM-3 cell line treated with paclitaxel where the CCLE curve should have been classified as sensitive.



Supplementary Figure 4: Examples of drug dose-response curves with unusual shapes.



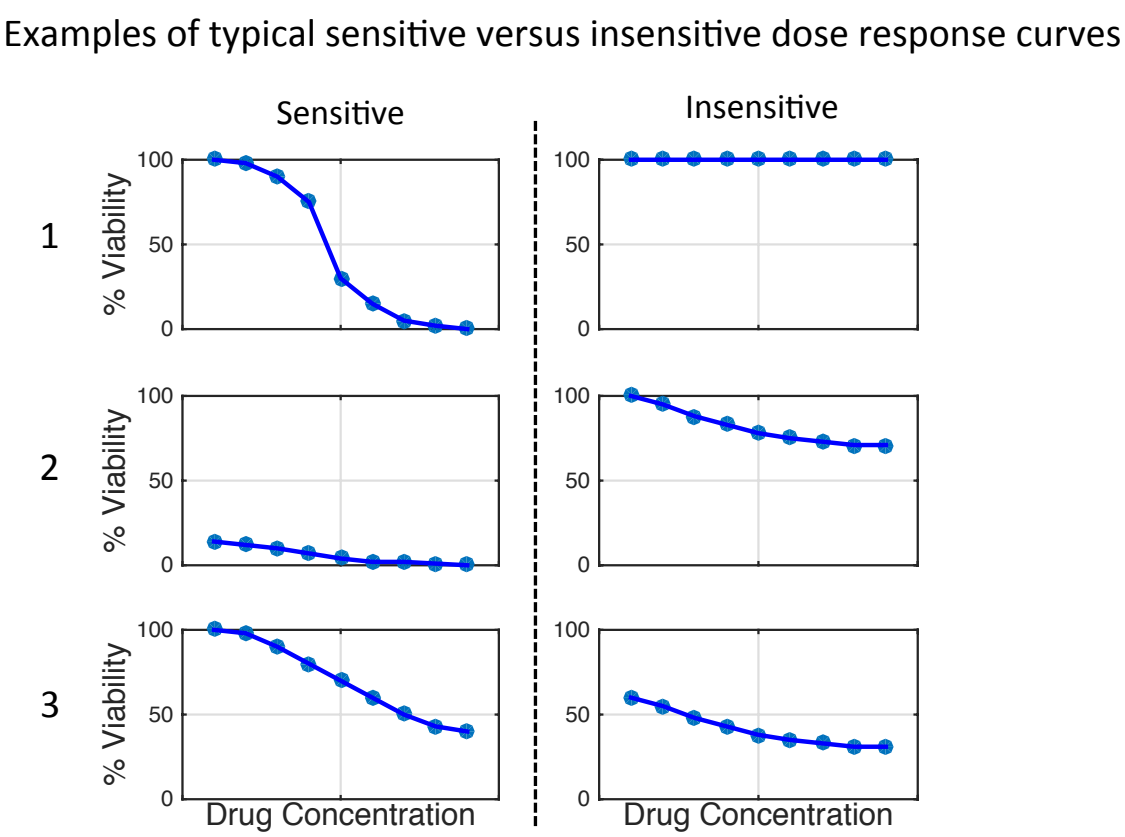
Response to Referee

The Manuscript "Drug Response Consistency in CCLE and CGP" by Bouhaddou et. al questions the validity of the main claim from Haibe-Kains et al. "Inconsistency in large pharmacogenomic studies. Nature 504, 389-393". The Haibe-Kains et al. study has one critical flaw; it fails to take into account that only viabilities of sensitive cell lines are expected to correlate between replicate studies and thus global measures of correlation will always underestimate the degree of concordance. At the most crude level cell lines can be divided into sensitive and insensitive, and the concordance of replicates can be assessed by comparing these binary calls. This is the approach taken by Bouhaddou et al. They assess the concordance of the large screens (GDSC and CCLE) by comparing the concordance of the binary calls (sensitive/insensitive) for each cell line/ drug combination. The study correctly identifies the major shortcoming of Haibe-Kains et al. but has some issues on its own. Safikhani et al. in their response to this critique identify several of the shortcomings, but also fail to convincingly argue about the validity of their original study (Haibe-Kains et al.) Overall, neither of the manuscripts is a valuable addition to the discussion.

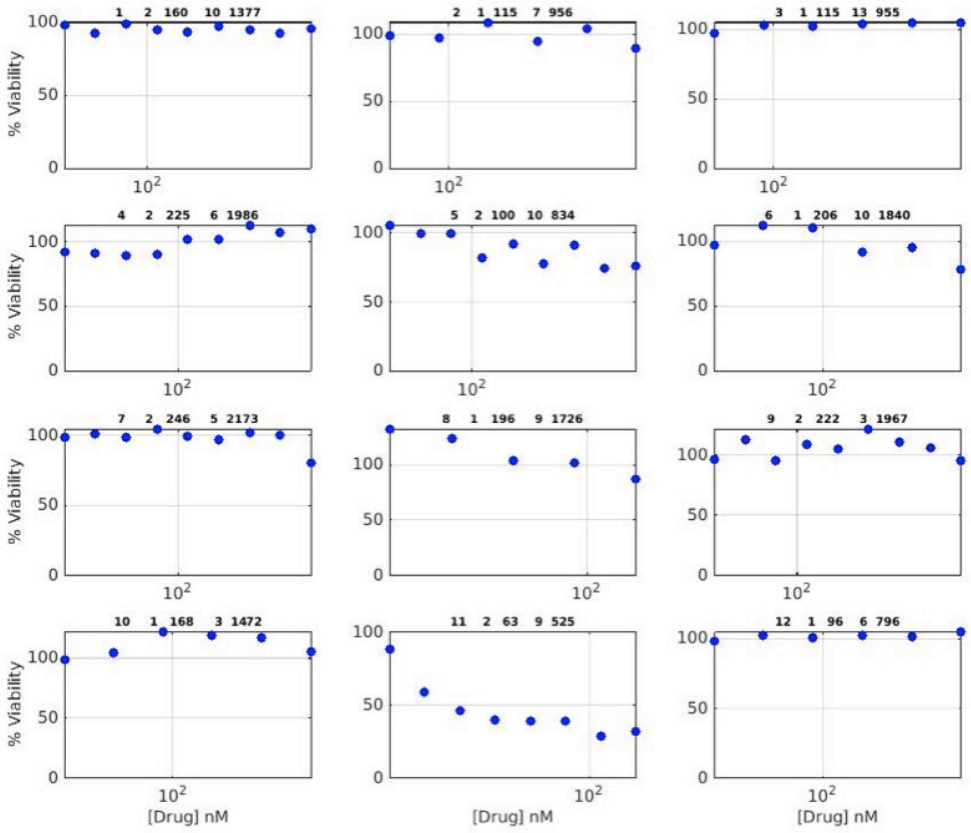
Major Issues:

1. Bouhaddou et al. follow an unusual approach of asking three independent reviewers to assess cell line viability. As pointed out by Haibe-Kains et al. the design of the study should be blinded and the studies should be evaluated independently.

We thank the Reviewer and Safikhani et al. for these suggestions to improve our manual curation process. In our revision, we have completely redone the manual curation in a blinded manner that provides independent evaluation of each study (i.e. curves from only one study at a time are presented to curators). We also randomized the curation with respect to the study, the drug, and the cell line. Lastly, we more than doubled the number of the manual curators (now 8 as opposed to 3 initially), and each curator was given a minimal, yet clear and consistent set of instructions for calling sensitivity. Extended Data Figures 1 and 2 contain an example of what manual curators were given, which we also paste below.



Extended Data Figure 1 | Example dose response curves given to manual curators. These idealized data represent various dose response curves one might find in CCLE and/or CGP and indicate how they should be classified.



Extended Data Figure 2 | One example page from the data given to manual curators. Each plot was rated as either sensitive (1) or insensitive (2). The numbers on the plot titles are the result of randomization—plot index, study, cell line, drug and cell line/drug index. Curators did not have the key. Each curator was given a different randomization.

2. The "% agreement" used by Bouhaddou et al. is a flawed metric of agreement and Safikhani et al. are correct in re-analyzing the data using Cohen's Kappa. Their data suggest that GDSC and CCLE are consistent. So, either the manual effort by Bouhaddou et al. is flawed or the original Haibe-Kains et al. publication was wrong (i.e. it was a false-negative due to the failure to do a more advanced or deeper analysis). This is an important controversy and should be settled by:

○ trying multiple alternative (perhaps non-linear) statistical approaches that capture the intents of manual classification;

○ repeating the manual effort using a larger number of curators using a proper blinded design;

○ evaluating consistency using multiple statistical tests - Kappa underestimates the concordance if one of the classes is small (e.g. a rare sensitive cell line among many insensitive ones);

○ expanding beyond binary classification (e.g. 3-class) to see at which point CCLE GSC become incongruent.

We thank the Reviewer for these suggestions which have strengthened our revised manuscript. There are multiple new results in our manuscript that address these points:

- As described in our above response to Point 1, we have redone the manual curation with a blinded (and randomized) design.
- We now compare AUC data between the two studies in addition to the slope data we previously compared. AUC data is shown to also have good concordance between CCLE and CGP (see below and Fig. 1a). We train a support vector machine (SVM) classifier based on both the AUC and slope data which captures the intents of the manual curation well (see below and Fig. 1b).

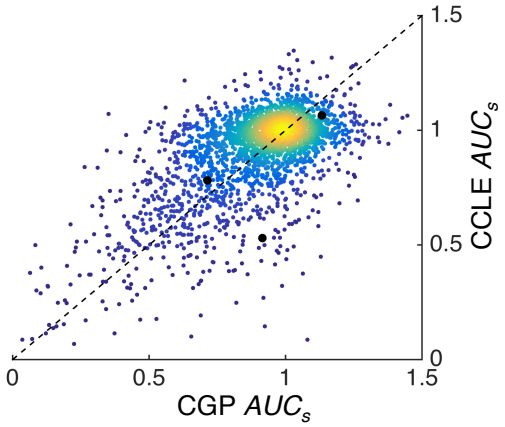


Figure 1a | Area under the curve (AUC) correlation between two studies. All areas are calculated considering a shared dose range. Color indicates density of dots. Dashed line is x=y line.

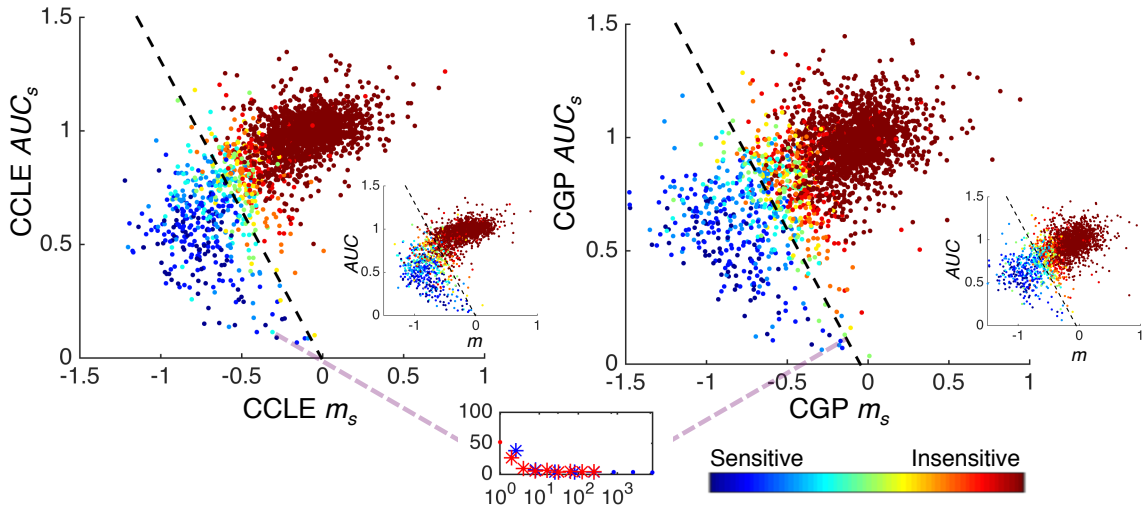


Figure 1b | Relationship between m_s and AUC_s for each database (inset *m* and *AUC* defined with the entire dose range as opposed to the shared dose range). The SVM classifier decision boundary divides the plot into sensitive and insensitive cell line/drug combinations, as indicated by the black dashed 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. Color 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.

- We report Cohen's Kappa and % agreement for consistency between the studies as a whole (i.e. no stratification by drug), both of which demonstrate significant consistency. However, application of Cohen's Kappa is inappropriate when one category is small, as the Reviewer pointed out, and most drugs have a very small number of sensitive cell lines. For example, data from PHA-665722 are in almost perfect agreement between the studies (see Table below), but Cohen's Kappa is identically 0. We therefore only calculate % agreement for individual drugs.

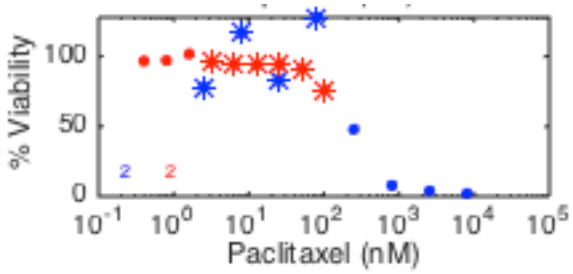
		CCLE	
		Sensitive	Insensitive
CGP	Sensitive	0	0
	Insensitive	1	89

- We agree it is important to understand the point at which CCLE and CGP become unacceptably incongruent. However, in our new manual curation test runs (prior to the entire exercise), we found that including more than two

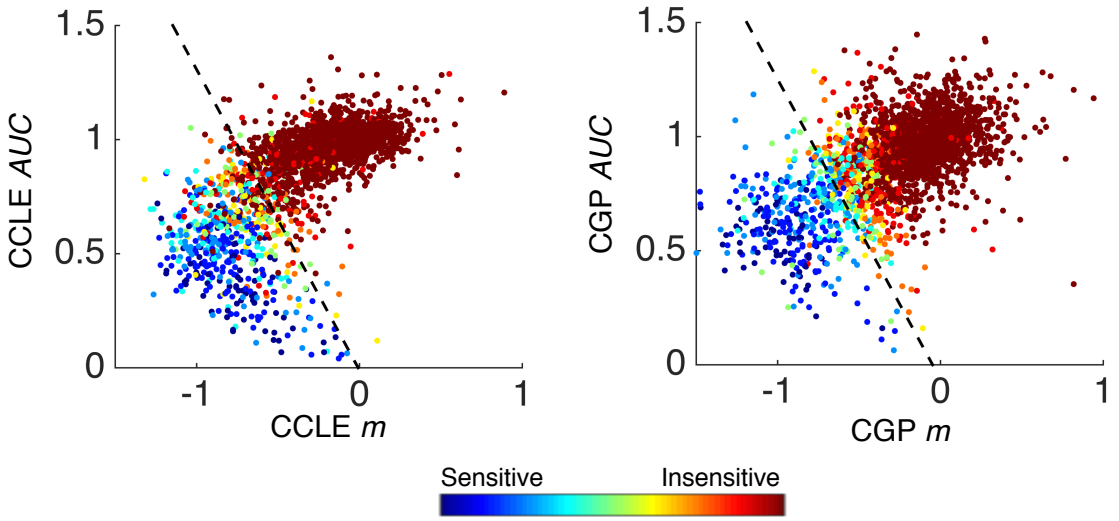
categories was confusing to curators, and thus problematic. Therefore, for the current manuscript, we retained the simplest possible binary description of sensitivity for manual curation, and leave more complex analyses (e.g. 3-class) for future work.

3. Bouhaddou et al. use only the shared range of concentrations to evaluate the slope of viability (% viability versus log10 dose); this means that some of the valuable data points have been thrown away for the sake of a simpler linear model. It should be shown that the conclusions also hold when all data points are used to estimate some form of "slope" or sensitivity.

The intent of our manuscript is to compare the CGP and CCLE, and in this context, we use the shared dose range because comparing outside of that range can lead to incorrect conclusions of inconsistency. For instance, in the example below with Paclitaxel (cell line SHP-77), if we were to consider the entire range of drug doses assayed by CCLE (blue dots) and CGP (red dots), CCLE would be rated as sensitive and CGP as insensitive only because the dose ranges differed. Within the dose range shared by both studies (starred), both databases are consistent.



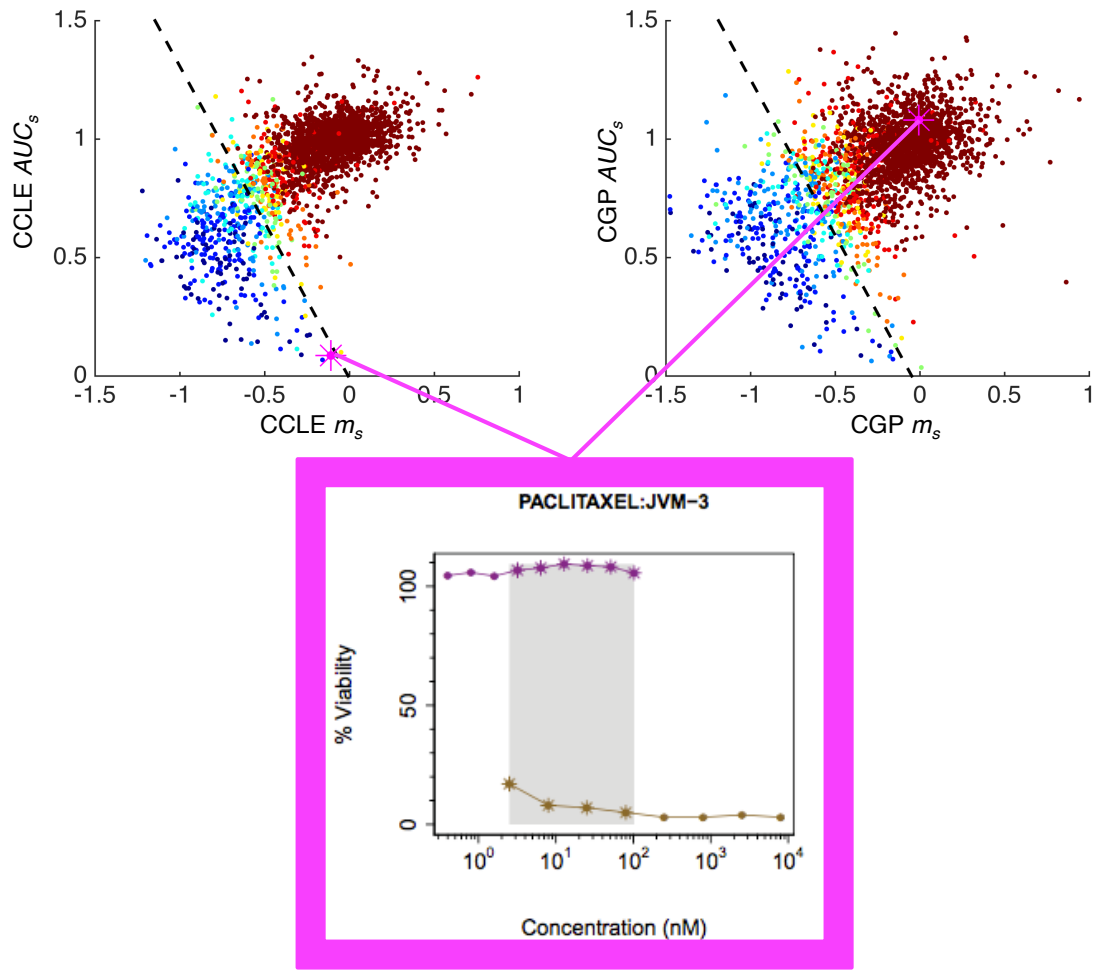
Nevertheless, the majority of CCLE and CGP data use scenarios will not focus on their comparison, and could thus benefit from using the full range of data. Therefore, in the revised manuscript we analyze the full range of available data for each study independently (see insets of Fig. 1b pasted below).



These results demonstrate our suggested sensitivity classification approaches also work well on the full dose range available in either study.

4. Safikhani et al. question the robustness of the ST statistic. Their examples are valid, but also suggest that ST under-estimates concordance rather than over-estimates it. Since Bouhaddou et al. only argue that GDSC and CCLE are concordant ST can be seen as a conservative approach.

In our revision we now consider area under the curve (AUC) in addition to slope (see above response to Point 2). This bivariate treatment is more robust for the cases where % viability is low starting from the lowest dose. As an example, we present below the Paclitaxel case highlighted by Safikhani et al., which our previous slope metric was insufficient to characterize. Our bivariate consideration of AUC in addition to slope now places this dose response properly as inconsistent between the studies (CCLE sensitive/CGP insensitive).



5. Bouhaddou et al. say that large studies can only "faithfully report on sensitivity vs. insensitivity categories" and not individual IC50 values, whereas Safikhani et al. say that this is a "surely" (sic) trivial finding. Both are very strong claims that are not backed up by data. Bouhaddou et al. do not show any attempts to estimate more accurate IC50 values (e.g. by utilizing cell-line drug combinations not shared by both of the studies or some additional experimental data points), while Safikhani et al. fail to realize (again) that IC50 values are not expected to be correlated across insensitive cell lines.

We have removed these claims from the manuscript. They were an aside to the main point of our paper, which remains: the CCLE and CGP can largely be considered consistent for drug sensitivity.

Editorial questions:

Many of the points below in this "*Editorial questions*" section do not seem to require a direct response from us, or have already been addressed in the above responses. We provide limited responses as needed below.

Is the criticism valid?

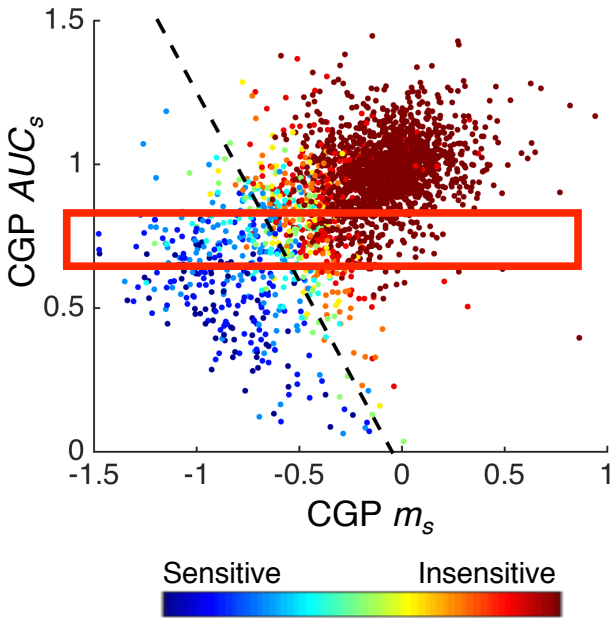
The criticism is valid, but the analysis to support the criticism is sub-par.

• *"We asked three independent people to manually evaluate consistency" This is not a technically rigorous strategy.*

We have significantly strengthened the manual curation; please see the above response to Point 1.

• *"Moreover, the CCLE and CGP used different drug dose ranges, which on its own could unfairly influence consistency." This could be a reasonable critique; however, the original article acknowledges - "however, the two studies used different experimental protocols...the pharmacological assay used, the range of drug concentrations tested, and choice of an estimator for summarizing the drug dose-response curve." The critique should specifically address why AUC is still not sufficient to compare the two studies.*

In the revised manuscript we now consider AUC and we find that AUC is also quite consistent between the two studies. To compare AUC between the studies, we recalculated AUC within the shared dose range, which the original article did not do. In comparing AUC values to the results of our new manual curation, we found that univariate classification based on AUC is worse than bivariate classification based on AUC and slope. Consider for instance the data points that fall within the red rectangle delineated on the plot from Fig. 1b pasted below. Although those dots all have a very similar AUC value (y-axis), approximately half are classified as insensitive and half as sensitive by manual curation. By evaluating the data along two dimensions, AUC and slope, we are able to better discriminate sensitivity.



Is there non-specialist interest?

Not much. Damage has been done by publishing the Haibe-Kains manuscript. The readers should be made aware that the work is flawed, but the Bouhaddou et al. manuscript is not written for a non-specialist readership and uses non-standard methodology. Overall it appears to be too much of a technical comment rather than a convincing illustration.

• Understanding the validity/quality of data produced from large-scale studies is important to a broader audience. Some technical points may be hard to understand, but I think the most important aspects of this response are made relatively clearly.

• The appendix is a useful resource for readers interested in a particular drug. It would be a bit easier to parse in a web interface, but I this is not essential to provide.

We have strived throughout the revision to improve clarity for non-specialists while retaining and enhancing technical rigor.

Is the argument persuasive?

The reply is thorough and identifies the shortcomings of Bouhaddou et al., but it fails to correct or soften the bold and largely sensationalist claims from Haibe-Kains et al. The additional remarks can be omitted.

• Fig. 1b does a good job of making the authors' point.

• Fig. 1d,e, as well as the corresponding discussion in the main text, do a good job of showing how correlations can be very different depending upon the assumed precision of measurement.

• This response, as well as the independent previous response, both make the point that most cell lines are not responsive to most drugs. Here, they say "The data also show that cell lines are "insensitive" to most tested drugs (~65%)". In general, this is important to take into consideration when defining a measure for consistency. If revised, this response could do an acceptable job of making this point.

Can the points be made more concisely?

Yes - but the authors do a good job of keeping the main text brief.

• The linear regression plots, and overlaid plots showing response across different drug doses, are more important than the "binary drug sensitivity calls".

• Remove Fig. 1a and c, basically anything involving "manual curation".

• Remove claims that are not followed with supporting evidence, such as "of course, machine learning classification1 may improve consistency inferences."

Given the large number of critiques of the Haibe-Kains et al. paper, which reflect the growing concern with the study in the community, some form of form of critique needs to be published. In the current form neither Bouhaddou et al. nor Safikhani et al. are sufficiently clear and persuasive.

Response to Safikhani et al.:

Overall percentage agreement is not a good measure of consistency.

We now report Cohen's Kappa where appropriate in addition to the standard measure of % consistency. Stratification by drug results in very few sensitive calls, for which Cohen's Kappa is known to be an inadequate statistic. Unfortunately, there seems to be little statistical literature that supports a useful metric for such cases. Please also see above response to Point 2, Bullet 3.

Potential bias in manual evaluation of drug dose-response curves

We have redone the manual curation to alleviate these concerns (please see above response to Point 1)

Limitations of the dose response slope for drug sensitivity calling

We have developed a bivariate SVM classifier based on slope and AUC, that addresses these limitations. Please see the above response to Point 2, Bullet 2, and Point 4.