SUPPLEMENTARY FILE 1

Iterations of the Brief Communication Arising, reviews and responses

Consistency in drug response profiling

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

The recent comparative analysis by Haibe-Kains *et al.*[1] concluded that data from two large-scale studies of cancer cell lines[2,3] showed highly discordant results for drug sensitivity measurements, while gene expression data were reasonably concordant. Here, we cross-compared the two original data sets[2,3] against our own data of drug response profiles in overlapping panels cancer cell lines. Our results indicate that it is possible to achieve concordance between different laboratories for drug response measurements by paying attention to the harmonization of assays and experimental procedures.

Haibe-Kains *et al*.[1] reported on a comparative evaluation of two drug sensitivity and molecular profiling datasets, one from the Cancer Genome Project (CGP)[2] and the other one from the Cancer Cell Line Encyclopedia (CCLE)[3]. Gene expression profiles between hundreds of shared cancer cell lines across all genes showed high consistency (median rank correlation of 0.85) between the two studies, whereas drug response data for 15 compounds were highly discordant (median correlation of 0.28 for IC_{50} response metric). This report[1] and the accompanying News & Views commentary[4] suggested that differences in laboratory protocols, compounds and their tested concentration ranges, and computational methods may account for the differences, but the report did not elaborate which of these factors are important and whether these can be controlled for.

Here, we reanalyzed the raw dose-response data from both CGP and CCLE using a single standardized area under the curve (AUC) response metric, which we call the drug sensitivity score (DSS)[5]. We then compared the CGP and CCLE data with a new dataset of drug response profiles from our own laboratory,[5] covering 308 drugs across 106 cancer cell lines. This included 48 compounds in common with CGP and 14 with the CCLE. In the AUC calculation, we further unified the drug concentration ranges across the CGP, CCLE and FIMM data. We observed a significantly (p=2.1×10^{-5}) higher level of consistency, especially between the CCLE and FIMM drug response data (median rank correlation 0.74), as compared to the comparison of FIMM and CGP data (0.54) (**Figure 1**).

Similar experimental protocols were applied at FIMM and CCLE, including the same readout (CellTiter-Glo, Promega), similar controls (vehicle as negative control and positive controls consisting of toxic compounds 100 µM benzethonium chloride or 1 µM MG132). However, there were also differences, such as the plate format (1536 vs. 384 wells) and obviously no effort to standardize cell numbers used or any other parameters, such as the source, passage number and media used for cells, nor the origin and handling of drugs. Therefore, this observed level of drug response agreement could be substantially improved by further standardization of the laboratory protocols. The CGP experimental protocol differed from the two others in terms of readout (fluorescent nucleic acid stain Syto 60, Life Technologies) and in the use of controls (drug-free cells as negative and no cells as positive controls). CGP data were analyzed either in 96- or 384-well plates. Since the CGP data included a narrower dose-range of the drugs, the other datasets had to be filtered accordingly, leading to loss of valuable information.

We compared the comprehensive drug-response profiles between the same cell lines from different laboratories, in line with the analysis by Haibe-Kains *et al*., showing consistency in gene expression profiling between the same cell lines from CGP and CCLE (median rank correlation of 0.85)[1]. Analysis of correlations at the level of individual drugs shows more variability, due to the fact that some drugs are not effective in any of the cell lines tested. Analogously, gene expression correlations vary more widely when analyzed at the gene-level (median rank correlation of 0.58 between CGP and CCLE), as certain genes are not expressed at a level above technical noise. Preferably, the same types of evaluation setups should be used when comparing the consistency of gene expression measurements with that of drug responses. However, we note that the analysis of Haibe-Kains *et al*. is in line with the way these data were used in the original publications[2,3].

In summary, we show that standardization of assay methods and laboratory conditions will help to improve the inter-laboratory agreements in drug response profiling. Global standards, similar to the MIAME standard for the microarray data[6], should be developed.

Methods

We scaled the drug response readouts using the available positive and negative controls from CGP, CCLE and FIMM screens. Curve fitting for the scaled dose-response curves was based on the original dose ranges, using the 4-parameter logistic model and the

Levenberg-Marquardt algorithm[7,8]. DSS was calculated based on the fitted dose-response models (DSS R-package available at http://dss-calculation.googlecode.com/svn/trunk/).[5] The DSS integration was restricted to the common concentration window shared between CGP, CCLE and FIMM, which was limited by the narrow dose-range used in the CGP data. The compounds were identified and matched using IUPAC International Chemical Identifiers InChIKeys[9]. Cell line matching was based on the name mapping file provided by Haibe-Kains *et al*.[1]

John Patrick Mpindi[1], Bhagwan Yadav[1], Päivi Östling[1], Astrid Murumägi[1], Akira Hirasawa[1], Sara Kangaspeska[1], Prson Gautam[1], Krister Wennerberg[1], Olli Kallioniemi[1] & Tero Aittokallio[1]

[1]Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland

References

1. Haibe-Kains B, *et al*. Inconsistency in large pharmacogenomic studies. *Nature* **504**, 389-393 (2013).
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. Weinstein, J. N. & Lorenzi, P. L. Discrepancies in drug sensitivity. *Nature* **504**, 381-383 (2013).
5. Pemovska, T. *et al.* Individualized systems medicine strategy to tailor treatments for patients with chemorefractory acute myeloid leukemia. *Cancer Discovery* **3**, 1416-29 (2013).
6. Brazma, A. *et al*. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat. Genet.* **29**, 365-71 (2001).
7. Levenberg, K. A method for the solution of certain problems in least squares. *Quart. Appl. Math.* **2**, 164-8 (1944).
8. Marquardt, D. W. An algorithm for least-squares estimation of nonlinear parameters. *J. Soc. Ind. Appl. Math.* 11, 431-441 (1963).
9. Nicola, G., Liu, T. & Gilson, M.K. Public domain databases for medicinal chemistry. *J. Med. Chem.* **55**, 6987-7002 (2012).

Figure 1. Consistency between drug response profiles of cell lines across FIMM, CCLE and CGP. The Spearman's rank correlation coefficient was calculated for the pairwise overlapping cell lines and over shared sets of compounds between FIMM-CCLE (27 cell lines, 14 compounds), FIMM-CGP (42 cell lines, 48 compounds), and CCLE-CGP (284 cell lines, 13 compounds). The DSS drug response profiles between FIMM and CCLE showed improved correlation compared with the correlation between FIMM and CGP or CGP and CCLE (p=2.1×10^{-5} or p=0.0058, respectively, two-sided Wilcoxon rank-sum test). The box and horizontal bar represent the interquartile range and median of the correlation coefficients, respectively, and the whiskers the most extreme data points. The two horizontal dotted lines indicate the median rank correlations for the IC$_{50}$ or AUC response metrics, respectively, when the drug response profiles between CGP and CCLE were compared at the level of individual compounds, as reported in Haibe-Kains *et al.*[1]

Response to "Consistency in drug response profiling"

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

We read the report by Mpindi *et al.* with great interest and we welcome the opportunity to comment. We find this work to be an important contribution to the discussion of pharmacogenomic consistency for a variety of reasons. Mpindi *et al.* were able to reproduce our initial finding of a substantial inconsistency between CCLE and CGP drug sensitivity data and to explore potential reasons behind the problem and possible solutions. To do this, they compared the CCLE and CGP to a new, unpublished dataset referred to as FIMM, that includes 308 drugs tested across 106 cancer cell lines; these important pharmacological data, as well as the authors' analysis code, should be shared with the scientific community to ensure reproducibility of their findings. Overall, their comparative analysis of this new data set supports our published finding of improved consistency between studies when there is greater similarity in experimental methods. Mpindi *et al.* also claim that harmonizing the readout, drug concentration range, and statistical estimator makes it possible to achieve greater consistency across pharmacogenomic studies, but the presentation of their results hides some subtleties in the analysis that should be highlighted. We provide more detailed comments below.

Consistency *across* vs. *between* cell lines. In our study we computed Spearman's correlation coefficients for each individual drug *across* cell lines, which was relevant to the overall goal of the CCLE and CGP studies to "discover new genomic biomarker of drug response to speed the emergence of personalized therapeutic regimens". In contrast, Mpindi *et al.* compute the correlation of cell line' sensitivity data across drugs (referred to as *between* cell lines in their presentation). While it is not clear how the application of the *between* cell line correlation analysis is relevant for identifying drug response biomarkers we agree with the authors that these different types of correlation should be thoroughly analyzed. Figure 1 shows both the *across* and *between* cell correlations and, consistent with our original conclusions, gene expression and mutation data were significantly more concordant than IC_{50} and AUC values in all comparisons (Wilcoxon rank sum test $p<0.01$; Figure 1), except for agreement of presence/absence of mutations and AUC sensitivity calls across cell lines (Cohen's kappa: 0.28 vs. 0.23 for mutations and AUC sensitivity calls, $p=0.16$; Figure 1). To ensure proper comparison of correlation estimates, we suggest that Mpindi *et al.* divide their first Figure into two panels showing the boxplots separately for correlations computed from *across* and *between* cell lines analyses.

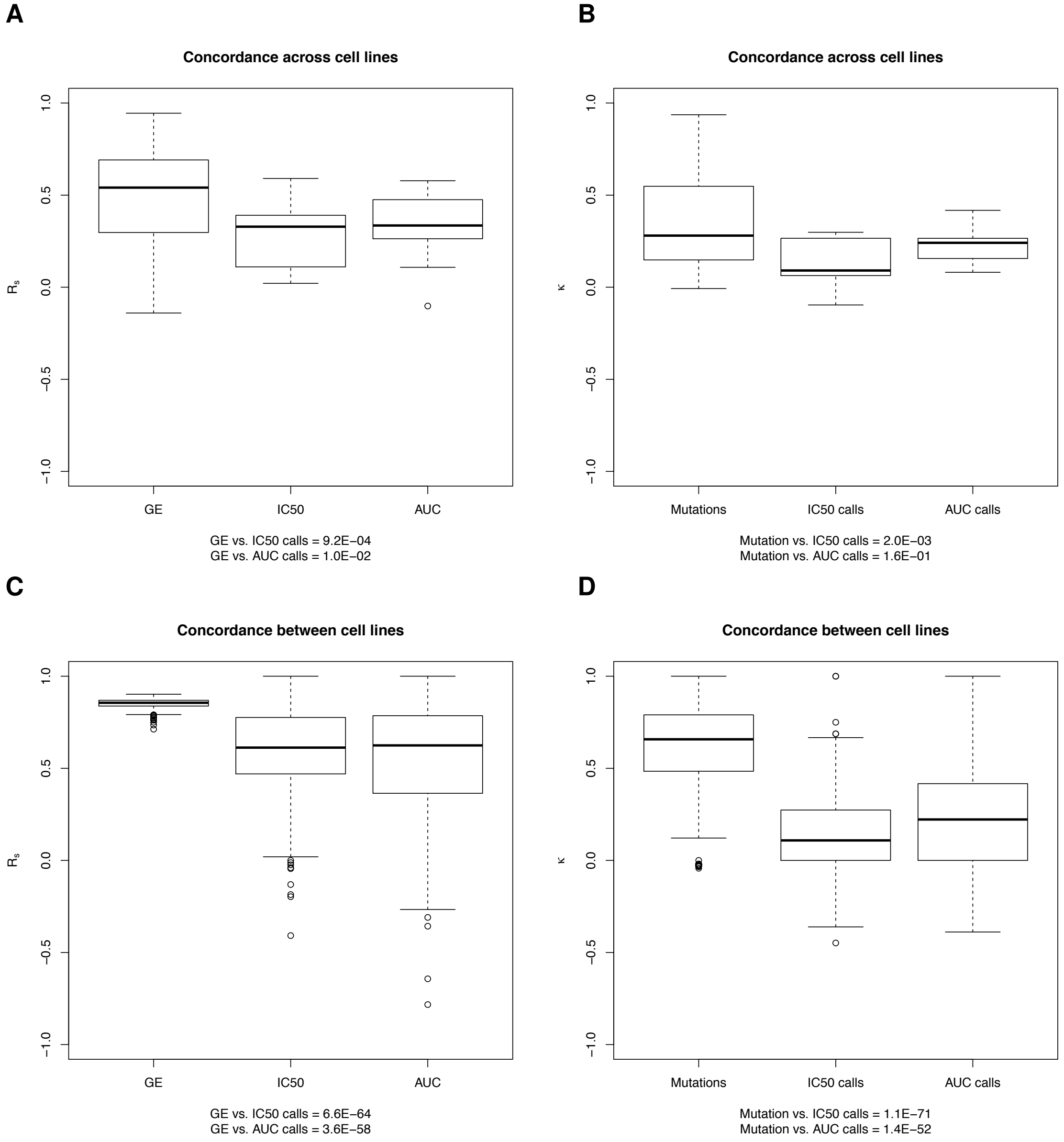
Improvement due to harmonization of experimental protocols. From Figure 1 in Mpindi *et al.*, it is not possible to properly quantify the improvement due to the authors' harmonization of drug concentration range and statistical estimator because the authors did not display the correlations computed from the original AUC values (before harmonization). Our analysis showed that the median of correlations computed from "unharmonized" AUC values to be 0.624, with similar interquartile range to that shown in the "CGP vs. CCLE" box in the Mpindi *et al's* Figure 1, therefore calling into question the benefit of the authors' harmonization. However, we concur with the Mpindi *et al* that studies with more similar experimental protocols should achieve significantly higher consistency. This is an important piece of evidence that is consistent with our published analysis of the GSK dataset, which, for a small number of drugs, suggested greater concordance with CCLE (which used the same cell viability assay) than with CGP. Overall we are very encouraged that the analysis by Mpindi *et al* supports our overall conclusion that large pharmacogenomic studies will require greater standardization and cross-validation to assure robustness and reproducibility of the associated data.

Minor comment

The authors claimed that they reanalyzed the raw dose-response curve from both CGP and CCLE. We requested similar access to data when performing our analysis and the CGP team provided their raw response data data. However, the CCLE team was unable to provide access to the actual raw data (only summary statistics, no data related to the actual technical replicates) as these data are stored in Novartis' platform software and not publicly available (Joseph Lehar; *personal communication*; Jan 21, 2014). We suspect that Mpindi *et al* did not have access to the actual assay data from the CCLE and that it is likely that higher consistency would have been found if the actual CCLE raw data were used.

Figures

Figure 1: (**A**) Box plot comparing Spearman's correlation coefficients estimating the concordance of gene expression, IC_{50} and AUC measures *across* cell lines in CGP and CCLE. (**B**) Box plot comparing Cohen's κ coefficients estimating the concordance of presence/absence of mutations, IC_{50} sensitivity calls and AUC sensitivity calls across cell lines in CGP and CCLE. (**C**) Box plot comparing Spearman's correlation coefficients estimating the concordance of gene expression, IC_{50} and AUC measures between cell lines in CGP and CCLE. (**D**) Box plot comparing Cohen's κ coefficients estimating the concordance of gene expression, IC50 and AUC measures between cell lines in CGP and CCLE. Significance of the difference between concordance observed for genomic and pharmacological data, as computed using the Wilcoxon rank sum test, is provided under each plot. GE: gene expression; R*s*: Spearman's rank-ordered correlation; κ: Cohen's κ coefficient.



Response to the comments of Referee #2:

Referee #2 (expertise in cancer genetics and bioinformatics):

Comment: In their original article, Haibe-Kains et al. (Nature 2013, 504, 389-393) analyzed two previously published large-scale genomic studies (The Cancer Genome Project (CGP) and the Cancer Cell line Encyclopedia (CCLE) to evaluate the correlation between genomic and drug response data. They concluded that although the genomic data between the studies remains concordant, drug response data was found to be highly discordant. The authors, although not conclusively, point out that poor correlation between drug response phenotypes is largely due to lack of standardization in experimental assays and data analysis methods.

In response, Mpindi et al. compared CCLE, CGP data with their own data (FIMM) of drug response in overlapping cell lines. A significant correlation of drug response was observed between the FIMM and CCLE dataset that shared 14 compounds and followed similar experimental procedures for drug response estimation. The authors conclude that correlation between drug response profiles can be substantially improved by using standardized assay methods.

Overall both groups of authors reach the same conclusion that methods to assay drug response are the major source of discordant results. Mapindi et al. also saw low correlation between CGP and CCLE drug response data, as seen by Haibe-Kains et al. In summary, this discourse is of low to modest interest to the general research community.

Our response: The way we and many others read the original Nature paper by Haibe-Kains et al. was that their main conclusion was the general inconsistency of drug testing, thereby questioning the findings made in cancer pharmacogenomics, and even the potential of personalized medicine.

Haibe-Kains et al. emphasized inconsistency, while our results indicated that it is possible to achieve a reasonably good concordance between different laboratories for drug testing measurements. The novel analysis was done using our unpublished in-house FIMM drug testing datasets as a common reference point. Of note, there was no attempt to specifically standardize the conditions between any of the three laboratories, and yet two sites were concordant with each other. Thus, with further standardization, we believe that it will be possible to achieve highly reliable and reproducible results. We feel our analysis therefore represents an important and novel angle to this topic, which is now being surrounded by considerable confusion, and therefore provides an important message both for the personalized medicine in general, as well as for the growing community of laboratories and scientists involved in drug sensitivity testing, as well as those analyzing integrated cancer omics and pharmacogenomics data.

In the revised version, we further demonstrate the relative contributions from the three levels of standardization: the effect of harmonization of (i) experimental assays, (ii) data analysis procedures, and (iii) comparison setup, on the inter-laboratory correlation, as shown in the new Figure 1.

Comment 1. Both Mpindi et al. and Haibe-Kains et al. explored the concordance of the different data sets generated by different research groups from two angles. Mpindi et al. compared each individual drug activity between the same cell lines from different laboratories, while the Haibe-Kains et al. calculated the correlation for each drug in all cell lines. While we acknowledge the significance of Haibe-Kains et al. study (Nature 2013), Mpindi et al. is more relevant to the overall goal of the personalized therapeutic regimens emphasized in both the CCLE and CGP studies; particularly, PARP-1 inhibitor sensitizes Ewing's sarcoma cell lines (Garnett et al., Nature 2012; Brenner et al., Cancer Res. 2012).

Our response: We fully agree that the way we made the inter-laboratory comparison of drug response profiles is relevant to the overall goal of personalized treatments (this is now mentioned at the end of page 2). We originally chose this approach because this was exactly the same approach that Haibe-Kains et al. used for comparing the gene expression and mutation profiles between the CCLE and CGP studies. Measuring drug response on a drug-by-drug basis can promote inconsistency, as the variability and dynamic range for some drugs can be minimal or non-existent compared to measurement noise. Since Haibe-Kains et al. analyzed drug response correlations in a different way than those of other parameters, this makes it inconsistent, if not even unfair, to make the conclusion that drug response data were substantially inferior in consistency. Our analyses demonstrate that results from drug sensitivity data show comparable consistency when evaluated the same way as they did for the gene expression (Fig. 1a). However, when using the Haibe-Kains et al. approach for the drug response comparisons, then the correlations are at a much lower level (Fig. 1b).

Comment 2. Standardization of the laboratory protocol will substantially improve the experimental agreement, particularly the drug response activity measured by CellTiter-Glo, which is very sensitive to cell numbers. While Mpindi et al. should provide the correlations of the original AUC before harmonization, it's expected that the improvement would be minor for "CGP vs CCLE" as both data sets used a completely different assay platform.

Our response: We have now provided the correlation results separately for analyses done before and after the harmonization (un-harmonized and harmonized plots, respectively, in Figure 1). While we cannot harmonize the laboratory protocols used by CCLE and CGP, we made our best effort to harmonize the data analysis protocols when processing the drug response data from CCLE, CGP and FIMM. The data analysis harmonization included careful matching of the compound IDs using IUPAC International Chemical Identifiers InChIKeys, as well as harmonization of the concentration ranges of the drugs between the three assays, and the way AUC was calculated based on the measured dose-response curves using our drug sensitivity score (DSS). These steps are now detailed in the Methods section.

As pointed out by the Referee, the improvements gained by such data analysis harmonization were rather modest in the “between cell lines” comparison, perhaps due to the rather limited number of common drugs between the three laboratories (Fig. 1a). However, the improvement was much larger in the “across cell line” comparison of Haibe-Kains et al. (Fig. 1b), especially for CCLE vs FIMM comparison that share more standardized assay protocols. This analysis demonstrates that while harmonization of the experimental assays is important (as shown by the differences in the inter-laboratory consistencies), harmonization of the data analysis procedures also contributes in part to the improved consistency of drug testing results. However, the biggest impact originated from the way by which the comparisons were calculated (i.e. whether one uses the "between cell lines" or "across cell lines" comparison).

Comment 3. Mpindi et al. should declare how they access the CCLE raw data as well as provide full access to their own FIMM data set in order to benefit the entire research community.

Our response: We originally downloaded the CCLE drug response dataset from the CCLE website (http://www.broadinstitute.org/ccle/data/browseData?conversationPropagation=begin, file name: CCLE_GNF_data_090613.xls). CCLE provides the percent inhibition values at each individual concentration point, enabling logistic function modeling using any dose-response curve fitting algorithm. The same logistic curve fitting and response modeling was used for CCLE, FIMM and CGP data using a modified area-under-the-curve metric as a response parameter. This was calculated using the drug sensitivity score (DSS) that we have recently made available (new Ref. 5, please see details in Methods). We have now made full access to our in-house FIMM DSS dataset (please see the next comment).

Comment 4. That fact that FIMM data set is unpublished, make it difficult to reproduce any of the results that are presented in communication form Mpindi et al.

Our response: We appreciate the importance of making the drug sensitivity data publicly available, and therefore we have put together the FIMM data set that was used in this comparison. We will make these still unpublished data available with this Brief Communication Arising, as Supplementary Dataset. The source-code of the DSS algorithm that was used to quantify drug response in the harmonized analyses is also available at https://bitbucket.org/BhagwanYadav/drug-sensitivity-score-dss-calculation.

We have now carefully cross-checked that the FIMM data release is comparable with the CCLE and CGP datasets. In the process of making sure that we are comparing exactly the same cell lines and drug compounds between the three datasets, we corrected a few annotation mismatches in our initial processing, which were due to combining together data from a number cancer cell line screens from FIMM. Importantly, these corrections did not affect our original conclusions.

Comment 5. Haibe-Kains et al. states that the CCLE raw data was not available for the analysis? This should be clarified.

Our response: This is true. We have now removed the word "raw" from text (last paragraph of page 1), and explained in the revised Methods section that we used the median of the triplicate measurements from CCLE (top of page 3).

Response to the comments from the original Nature authors:

Our response: We thank Haibe-Kains et al. for their comments. We agree with many of the points in their response, especially those we already raised in our original submission. In particular, we indeed reproduced their original results between CCLE and CGP data, but more importantly, we provided novel results using our unpublished in-house FIMM dataset as unique reference point. These results actually demonstrated that it is possible to achieve a good inter-laboratory concordance in drug-response measurements, after standardization of the experimental assay protocols and the comparison setup.

To ensure the reproducibility of these results, we have now made available both the FIMM drug testing dataset (Suppl. Dataset), as well as our implementation of the drug sensitivity score (DSS) (Methods).

Comment: Consistency across vs. between cell lines.

Our response: In their response, Haibe-Kains et al. showed the results both from the across and between cell line comparison. However, in their original paper, they chose to present the gene expression and mutation comparisons using the “between cell lines” correlation analyses, whereas the drug response comparisons were done using the “across cell lines” approach. We feel this led to inconsistent comparison results, which may have led to an unfair conclusion about the level of consistency of drug response measurements. Indeed, like we demonstrated in our submitted manuscript, and Haibe-Kains et al. later in their response, the design of this comparison has a major effect on the correlation results.

In new Figure 1, we now show the effect of the “between cell lines” and “across cell lines” correlation analyses in each of the comparisons separately (CCLE vs CGP, CCLE vs FIMM and CGP vs FIMM).

Comment: Improvement due to harmonization of experimental protocols.

Our response: Revised Figure 1 also demonstrates that while harmonization of the experimental assays is important (as shown by differences in the inter-laboratory consistencies between the three sites), harmonization of the data analysis procedures may also contribute in part to improved consistency of drug testing results, especially when using the “across cell line” comparison setup of Haibe-Kains et al. (Fig. 1b). However, by far the biggest impact on the consistency originates from the harmonization of the comparison setup (i.e. whether one uses “between cell lines” or “across cell lines” comparison).

Minor comment.

Our response: It is true that we did not have an access to the actual assay data from CCLE (i.e. raw data at the level of plates, wells and replicates), rather we used the median-summarized data accessible from their website (http://www.broadinstitute.org/ccle/data/browseData?conversationPropagation=begin, file name: CCLE_GNF_data_090613.xls). We fully agree with Haibe-Kains et al. that it is likely that even higher consistency could have been achieved if we had access to actual raw data, enabling also quality control steps like we do in our in-house drug screening.