bioRxiv preprint doi: https://doi.org/10.1101/213553; this version posted November 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

A multi-center study on factors influencing the reproducibility of *in vitro* drug-response studies

Mario Niepel*[1], Marc Hafner*[1], Elizabeth H. Williams[1], Mirra Chung[1], Anne Marie Barrette[2], Alan D. Stern[2], Bin Hu[2], LINCS Consortium, Joe W. Gray[3], Marc R. Birtwistle[2,§,††], Laura M. Heiser[3,††], and Peter K. Sorger[1,††]

* These authors contributed equally

[††]Co-corresponding authors

[1]HMS LINCS Center
Laboratory of Systems Pharmacology
Department of Systems Biology
Harvard Medical School
Boston, MA 02115, USA

[2]Drug Toxicity Signature Generation (DToxS) LINCS Center
Department of Pharmacological Sciences
Mount Sinai Institute for Systems Biomedicine
Icahn School of Medicine at Mount Sinai
One Gustave L. Levy Place Box 1215
New York NY 10029

[3]Microenvironment Perturbagen (MEP) LINCS Center
Oregon Health Sciences University
Collaborative Life Sciences Building (CLSB) - CL3G
2730 SW Moody Avenue
Portland, OR 97201

[§]Current address:
Clemson University
Dept. of Chemical and Biomolecular Engineering
Earle Hall
206 S. Palmetto Blvd.
Clemson, SC 29634

†Contact author:
Peter K. Sorger,
200 Longwood Avenue
Boston, MA 02115,

bioRxiv preprint doi: https://doi.org/10.1101/213553; this version posted November 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

peter_sorger@hms.harvard.edu copying christopher_bird@hms.harvard.edu and heiserl@ohsu.edu
Phone (617) 432-6902
Fax (617) 432-6990
ORCID ID: 0000-0002-3364-1838

RUNNING TITLE: Reproducibility in drug-response assays

KEYWORDS: Reproducibility, cancer, drug screening, pharmacology

bioRxiv preprint doi: https://doi.org/10.1101/213553; this version posted November 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

ABSTRACT

Evidence that some influential biomedical results cannot be recapitulated has increased calls for data that is findable, accessible, interoperable, and reproducible (FAIR). Here, we study factors influencing the reproducibility of a prototypical cell-based assay: responsiveness of cultured cell lines to anti-cancer drugs. Such assays are important for drug development, mechanism of action studies, and patient stratification. This study involved seven research centers comprising the NIH LINCS Program Consortium, which aims to systematically characterize the responses of human cells to perturbation by gene disruption, small molecule drugs, and components of the microenvironment. We found that factors influencing the measurement of drug response vary substantially with the compound being analyzed and thus, the underlying biology. For example, substitution of a surrogate assay such as CellTiter-Glo® for direct microscopy-based cell counting is acceptable in the case of neratinib or alpelisib, but not palbociclib or etoposide. Uncovering and controlling for such context sensitivity requires systematic measurement of assay robustness in the face of biological variation, which is distinct from assay precision and sensitivity. Conversely, validating assays only over a narrow range of conditions has the potential to introduce serious systematic error in a large dataset spanning many compounds and cell lines.

INTRODUCTION

The goal of making biomedical data more findable, accessible, interoperable, and reusable (the FAIR principles (Wilkinson, Dumontier et al. 2016)) as well as reports from industry that call into the question the reproducibility of published data (Arrowsmith 2011, Prinz, Schlange et al. 2011, Begley and Ellis 2012, Baker 2016) have increased interest in data reliability (Errington, Iorns et al. 2014, Morrison 2014). However, only a limited number of studies (https://f1000research.com/channels/PRR) such as the Science Exchange Reproducibility Initiative

bioRxiv preprint doi: https://doi.org/10.1101/213553; this version posted November 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

(http://validation.scienceexchange.com/#/reproducibility-initiative) have addressed the issue through new experiments and the results of such reproducibility experiments have themselves been controversial (eLIFE-Editorial 2017, Ioannidis 2017, Nature-Editorial 2017, Nosek and Errington 2017). In this paper, we investigate a prototypical class of cell-based assays—rather than a specific scientific result—through collaboration among multiple geographically dispersed laboratories. This work was carried out as part of our participation in the NIH Library of Network-Based Cellular Signatures Program (LINCS) consortium (http://www.lincsproject.org/). The overall goal of LINCS is to generate broadly useful datasets characterizing the responses of cells to perturbation by small molecule drugs, components of the microenvironment, and gene depletion or overexpression. For such a resource to be useful, it must be reproducible.

The assay studied by LINCS Centers in this paper involves measuring the responses of tissue culture cells to small molecule anti-cancer drugs across a dose range. Drug-response assays in cultured cells are widely used in preclinical pharmacologyc(Cravatt and Gottesfeld 2010, Schenone, Dancik et al. 2013) and the study of cellular pathwaysc(Barretina, Caponigro et al. 2012, Garnett, Edelman et al. 2012, Heiser, Sadanandam et al. 2012). In the case of anti-cancer drugs, cells are exposed to drugs or drug-like compounds for several days (commonly three days) and the number of viable cells then determined, either by direct counting using a microscope or using a surrogate assay such as CellTiter-Glo® (Promega), which measures ATP levels in a cell lysate. With some important caveats, the amount of ATP in a lysate from a single microtiter well is proportional to the number of viable cells in that well(Tolliday 2010). Several large-scale datasets describing the responses of hundreds of cell lines to libraries of anti-cancer drugs have recently been publishedc(Barretina, Caponigro et al. 2012, Garnett, Edelman et al. 2012, Seashore-Ludlow, Rees et al. 2015, Haverty, Lin et al. 2016), but their reproducibility and utility is being debatedc(Haibe-Kains, El-Hachem et al. 2013, Consortium and Consortium 2015, Safikhani, Freeman et al. 2015, Bouhaddou, DiStefano et al. 2016).

bioRxiv preprint doi: https://doi.org/10.1101/213553; this version posted November 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Our approach was simple: five experimentally-focused LINCS Data and Signature Generation Centers measured the sensitivity of the widely-used, non-transformed MCF 10A mammary epithelial cell line to eight small molecule drugs with different targets and mechanisms of action. The LINCS Data Coordination and Integration Center (DCIC) then processed and visualized the data working with LINCS Center 1. LINCS investigators have established that the way in which drug response is conventionally calculated from cell viability assays is confounded by variability in rates of cell proliferation arising from changes in plating density, fluctuation in media composition, and intrinsic differences in cell division times(Hafner, Niepel et al. 2016, Hafner, Niepel et al. 2017). We corrected for these and other known confounders using the growth rate inhibition (GR) method(Hafner, Niepel et al. 2016, Hafner, Niepel et al. 2017, Niepel, Hafner et al. 2017), thereby focusing the current study on other sources of irreproducibility that remain poorly understood. Individual Centers were provided with identical aliquots of MCF 10A cells, drugs, and media supplements, as well as a common experimental protocol and data analysis procedure. Some variation in methods was inevitable, because not all laboratories had access to the same instrumentation or the same level of technical expertise; in our view, this is a positive feature of the study because it more fully replicates “real-world” conditions. In initial experiments, we observed substantial center-to-center variation. We then performed systematic studies to identify those factors with the largest impact on the measurement of drug response, and distributed this information to other centers to improve experimental and analytical procedures. We found that irreproducibility arose from a subtle interplay between experimental and computational methods and poorly understood sources of biological variation. Thus, a sustained commitment to characterizing and controlling for this variability was necessary to obtain reproducible data from different sites.

RESULTS

Measuring drug responses in collaboration

To establish the single-Center precision of dose-response assays, Center 1 performed multiple replicates of a 13-point dose-response assay with MCF 10A cells and the MEK1/2 kinase inhibitor

bioRxiv preprint doi: https://doi.org/10.1101/213553; this version posted November 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Trametinib at concentrations between 0.33 nM and 1 µM. Both technical and biological replicates were performed (Figure 1): for technical replicates, multiple drug dilution series were assayed on a single microtiter plate at a single time. For biological replicates, three sets of assays were performed on successive days in different plates; each biological replicate involved three technical replicates. In all cases, viable cell number was determined by differentially staining live and dead cells, collecting fluorescence images from each well, segmenting images using machine vision software, and then counting all viable cells in a well(Hafner, Niepel et al. 2016, Niepel, Hafner et al. 2017). Sigmoidal curves were fitted to the data and four response metrics derived: these measured potency (GR_{50}), maximal efficacy (GR_{max}), slope (Hill Coefficient or h_{GR}) of the dose response curve, and the integrated area over this curve (GR_{AOC})(Hafner, Niepel et al. 2016). Fitting procedures and response metrics have been described in detail previously(Hafner, Niepel et al. 2016, Hafner, Niepel et al. 2017) (Supplemental Figure 1), and all routines and data can be accessed on-line or via download at http://www.grcalculator.org/.

We found that response curves for technical replicates were nearly indistinguishable (Figure 1), showing that purely technical error resulting from inaccurate pipetting or errors in counting cells is small. For biological replicates, standard deviations in measurement of drug potency and efficacy ($\log_{10}(GR_{50})$ values and GR_{max} values, respectively) were ~ 0.07, representing the repeatability of the assay at a single research site across multiple days.

To measure reproducibility across laboratories while controlling for anticipated sources of variability, a single LINCS Center distributed identical MCF 10A aliquots, drug stocks, and media additives, as well as a detailed experimental protocol optimized for cell line-drug pairs under study. This protocol included optimal dose-ranges and separation between doses for reliable curve fitting. When individual LINCS centers first performed these assays, up to 200-fold variability in GR_{50} values was observed between centers (Supplementary Figure 2). We therefore performed directed experiments in Center 1 to systematically investigate the origins of assay variability.

bioRxiv preprint doi: https://doi.org/10.1101/213553; this version posted November 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Technical drivers of variability

First, we examined inter-Center variability in estimation of GR_{max} focusing on the topoisomerase inhibitor Etoposide and CDK4/6 inhibitor Palbociclib. One LINCS Center used the CellTiter-Glo® ATP-based assay and a luminescence plate reader as a proxy for the number of viable cells; CellTiter-Glo is a common substitute for direct cell counting when a suitable microscope is not available. When we performed side-by-side experiments we found that dose-response curves and GR metrics computed from (image-based) direct cell counts and CellTiter-Glo® values were the same for the EGFR inhibitor Neratinib, differed marginally for the PI3K inhibitor Alpelisib, and exhibited significant differences for the topoisomerase inhibitor Etoposide or CDK4/6 inhibitor Palbociclib (Figure 2A,B). This finding explains, in part, inter-Center differences in drug response observed in preliminary experiments (Supplementary Figure 2). It is known that CellTiter-Glo® and direct cell counts are poorly correlated when drugs cause dramatic changes in cell size or alter ATP metabolism, thereby changing the relationship between ATP level in cell extracts and cell number (Figure 2C for Palbociclib)(Salani, Marini et al. 2013, Harris, Koh et al. 2016, Soliman, Steenson et al. 2016). Our data show that the degree of agreement between cell counting and CellTiter-Glo® depended on the drug being assayed (previous work has shown that cell line also impacts the correlation between the assays(Niepel, Hafner et al. 2017)). Thus, direct cell counting and CellTiter-Glo® measurements can be substituted for each other in some cases but not in others. More generally, a substitution of assays that appears to be justified by pilot studies on a limited number of cell lines and drugs can be problematic when the number and chemical diversity of drugs is increased.

Edge effects and non-uniform cell growth are a second and substantial source of variation in cell based studies performed in microtiter plates(Coyle, Green et al. 1989, Bushway, Azimi et al. 2010) thought to arise in part from temperature gradients and uneven evaporation of media at the edges of plates. We have observed a variety of irregularities in plating and cell growth at least some of which are dependent on the batch of microtiter plate; we now test all batches of plates for uniformity of cell

bioRxiv preprint doi: https://doi.org/10.1101/213553; this version posted November 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

growth(Niepel, Hafner et al. 2017). Because variation in growth is often confined to specific regions of a plate (Figure 2D), it can generate systematic error in dose-response curves. A variety of approaches can be taken to minimize such effects (e.g. placing plates in humidified chambers to reduce evaporation from edge wells), but we found that automated and randomized compound dispensing is particularly helpful. Using the HP D300e Digital Dispenser it is possible to "print" compounds directly into microtiter plates in an arbitrary pattern, randomizing the locations of control and technical replicate samples in the plate. In this way, systematic error arising from edge effects is converted into random error, which doesn't result in a directional bias of response metrics and is more easily modeled(Niepel, Hafner et al. 2017). We have found that the use of simple washing and dispensing robots also reduces errors that humans make during repetitive pipetting operations. Most of these robots are small, robust, and relatively inexpensive, and our experience suggests that they greatly improve the reproducibility of medium- and high-throughput cell-based and biochemical studies.

A third source of error we explored involved the concentration range over which a drug was assayed and the impact of the range on curve fitting and parameter estimation. For example, if we follow general practice and assay Trametinib over a thousand-fold concentration range, growth is fully arrested at ~30 nM (Figure 3A, left plot, red arrow): phenotypic response does not change even when the dose is increased 100-fold to 1 µM and increasing the dose-range has no effect on curve fitting and parameter estimation (Figure 3A, left plot). However, when Dasatinib (a poly-selective SRC-family kinase inhibitor) was assayed over a thousand-fold range, curve fitting identified a plateau in GR value between 0.3 to 1 µM, but when the dose-range was extended GR values become negative, demonstrating a shift from cytostasis to cell killing (Figure 3A, right plot, and 3B). The subtlety here is that a dose-range that is adequate for analysis of Trametinib is not adequate for Dasatinib. This sort of variation is difficult to spot in a high-throughput experiment and suggests that efficient procedures are needed to optimize dose ranges for specific compounds careful quality control of curve fitting.

bioRxiv preprint doi: https://doi.org/10.1101/213553; this version posted November 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

A fourth source of inter-Center variation involved over-estimation of cell number in the presence of high doses of Dasatinib and Neratinib when we compared Centers that used imaging-based assays (Figure 3C). Above 1 µM, GR values were negative for both drugs denoting cell death, but cell counts at one LINCS Center implied purely cytostatic effects. Follow-up studies showed that the discrepancy arose from the use of image processing algorithms that included dead cells in the "viable cell count" and from over-counting the number of cells when the drugs induced frequent multi-nucleation(Roytta, Laine et al. 1987, Orth, Kohler et al. 2011). Observed differences in drug response across Centers could be recapitulated in a single laboratory using two different image processing routines and were also evident by visual inspection of the segmented images (Figure 3C,D). In retrospect, all Centers should have processed images in the same way, but such routines are often built into manufacturer's proprietary software making identical image analysis dependent on transfer of primary data. Furthermore, this level of harmonization is impossible to achieve when replicating published results. This example demonstrates the importance of locking down all steps in the data processing pipeline from raw measurements to final parameter estimation, as well as a relatively subtle interplay between biological and technical sources of variability.

Biological factors impacting repeatability

Variables that can change the biology of drug response, such as media composition, incubation conditions, microenvironment, media volume, and cell density, have been discussed elsewhere(Hafner, Niepel et al. 2016, Haverty, Lin et al. 2016) and were controlled to the greatest extent possible in this study through standardization of reagents and the use of GR metrics. In a truly independent repeat of the current study, experimental variables such as these would need to be considered as additional confounders, because it is difficult to fully standardize a reagent as complex as tissue culture media. However, one Center performed a preliminary comparison of batches of horse serum, hydrocortisone, cholera toxin, and insulin and found that the effects on drug response were smaller than the sources of variation discussed above (data not shown). At the outset of the study, we had anticipated that the origin

bioRxiv preprint doi: https://doi.org/10.1101/213553; this version posted November 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

of the MCF 10A isolate would be an important determinant of drug response. MCF 10A cells have been grown for many years, and karyotyping reveals differences among isolates(Soule, Maloney et al. 1990, Caruso, Reiners et al. 2001, Cowell, LaDuca et al. 2005, Kim, Yang et al. 2008, Zientek-Targosz, Kunnev et al. 2008, Marella, Malyavantham et al. 2009), which is why we distributed aliquots of a single isolate to all Centers. However, we detected very little variability in drug sensitivity when three different MCF 10A isolates were compared directly to each other and also to a histone H2B-mCherry-tagged subclone of one of the isolates (Supplemental Figure 3). Variability among MCF 10A isolates assayed at Center 1 (including different sub-clones from the same master stock) was also smaller than what was observed when a single isolate was assayed at different Centers. However, in other settings, clonal variation is likely to have a much larger impact(Ramirez, Rajaram et al. 2016).

The duration of drug exposure is not generally explored in *in vitro* studies, and instead researchers typically assay cells after a fixed time point after treatment. Here, we monitored the responses of MCF 10A cells to drugs in a live-cell experiment in which cell number was measured every two hours using an automated high-throughput microscope. We then quantified the response by calculating GR values over a 12 hour moving window (instantaneous GR values) and found that the effect of time and dose were substantial in some cases and varied with drug. For example, instantaneous GR values for cells exposed to Etoposide were nearly constant at all drug doses throughout a 50-hour assay period (Figure 4, top left plot), whereas instantaneous GR values for 0.1 μM of Neratinib varied from 0 to 1 over the same period (Figure 4, bottom left plot), with the highest variability at intermediate drug doses. As a consequence, GR dose-response curves and metrics derived from these curves, such as GR_{50} and GR_{max}, varied with time (Figure 4 and Supplemental Figure 4). The temporal dependence of drug response is likely to reflect biological adaptation, drug export, and other factors important in understanding drug mechanism of action(Fletcher, Haber et al. 2010, Muranen, Selfors et al. 2012, Hafner, Niepel et al. 2016, Harris, Frick et al. 2016, Fallahi-Sichani, Becker et al. 2017), factors which,

bioRxiv preprint doi: https://doi.org/10.1101/213553; this version posted November 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

in a high-throughput assay, remain unexplored and can add to the difficulties in reproducing experimental results.

Final results

When the factors described above were taken into account to the greatest extent possible, we found that reproducibility of drug-response measurements across five LINCS Centers could be reduced to a standard deviation of about 2.4-fold (standard deviation value of $\log_{10}(GR_{50})$ is ~0.38; see Figure 5). In the final dataset, one center still relied on manual compound dispensing because a robot was not available and a second center used CellTiter-Glo® assays rather than direct cell counting to estimate viable cell number. Despite such differences in procedure, which are typical of "real-world" conditions, inter-Center variability at the end of the study was dramatically lower than at the outset, while remaining about 5-fold higher than for biological repeats at a single center (Figure 1, fourth panel; 5A-B). Variability remained drug-dependent. For example, for Paclitaxel estimates of GR_{max} and Hill Coefficient (h_{GR}) were variable, and GR_{50} and h_{GR} values were variable for Palbociclib (Figure 5A-B). In contrast, for Neratinib, response was reproducibly measured (standard deviation for all metrics below 0.17 with a single outlier).

DISCUSSION

In this study, we show that five geographically dispersed laboratories can, with effort, generate reasonably reproducible drug-response data in tissue culture cells. Inter-lab precision was three- to five-fold lower than that achievable within a single laboratory, likely reflecting persistent differences in experimental procedure. This level of precision required some effort to achieve and, in our judgement, exceeds the norm for this class of experiments in the current literature (although this is not easy to prove). Further improvements in inter-center precision and reproducibility would necessitate the implementation of identical and automated compound handling, pipetting, and cell counting procedures in all laboratories. This was not achieved, because of the expense of acquiring the necessary instrument

bioRxiv preprint doi: https://doi.org/10.1101/213553; this version posted November 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

and a belief—belied by subsequent experiments—that counting cells is such a simple procedure that different assays can be substituted for each other without consequence.

At the outset of the study we had hoped to identify the specific biological, experimental, and computational factors that had the largest impact on irreproducibility across individual Centers and thereby generate a ranked list of the most important factors in ensuring reliable data collection. However, we discovered that in many cases irreproducibility was *itself* irreproducible and that the technical factors responsible for specific outlier datapoints were difficult to diagnose. We therefore undertook a systematic study of the assay itself, with an eye to identifying those variables with the greatest impact on data quality. We found that irreproducibility most commonly arises from unexpected interplay between experimental protocol and true biological variability. For example, estimating cell number from ATP levels using the CellTiter-Glo® assay produces very similar results as direct counting of cells in the case of Neratinib, but this is not true when cells are exposed to Etoposide or Palbociclib (Figure 2). The discrepancy most likely arises because ATP levels in lysates of drug-treated cells vary for reasons other than changes in cell number, including changes in cell size and metabolism. This observation has important implications for the design of experiments in which diverse compounds are screened. We have previously shown that the density at which cells are assayed also has a dramatic effect on drug response(Hafner, Niepel et al. 2016), but this too is context dependent. For some cell line-drug pairs, density has little or no effect, whereas for other pairs it increases drug sensitivity, and for yet others it decreases sensitivity.

Preliminary studies suggest that such context dependence reflects real changes in the underlying biology and not flaws in assay methodology itself. For example, cell density directly impacts media conditioning and the strength of autocrine signaling, which affects response to some drugs but not others(Yonesaka, Zejnullahu et al. 2008, Wilson, Fridlyand et al. 2012). As a consequence, changes in protocol that might seem unimportant based on a control study in one cell type or biological setting might, nonetheless, substantially affect results obtained with other cells or growth conditions. From this

bioRxiv preprint doi: https://doi.org/10.1101/213553; this version posted November 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

we conclude that there is no substitute for empirical analysis of even a seemingly simple assay. In general, the variables with the greatest effect on measurement of drug response differed from what we expected *a priori*. For example, isolate-to-isolate differences (evident in karyotypic changes) in MCF 10A cultures had less of an effect on drug response assays (Supplemental Figure 3) than the ways in which drugs and cells were plated into multi-well plates and counted (Figure 2D, 3). The only way to identify and control for such variation is to conduct comprehensive experiments aimed at empirically establishing the range of conditions over which data remain precise and exact for a significant number of cell lines and drugs used in a specific profiling effort.

Data processing routines are as important for reproducibility as well-controlled experiments(Sandve, Nekrutenko et al. 2013). Data and data analysis routines can interact in multiple ways, some of which are clear in retrospect but not necessarily anticipated. For example, collecting 9-point dose response curves generally represents good practice, but it is essential that the dose range effectively span the GR_{50} (the mid-point of the response) and account for the possibility of compound dose response curves (which is often ascribed to inhibition of different targets over different dose ranges). When this is not the case (as illustrated by Figure 3A), curve fitting is underdetermined and response metrics become unreliable. In many cases problems with dose range are not evident until an initial assay has been performed. Iterative design is straightforward in small scale studies, but substantially harder for large-scale screens; for a large dataset, data processing routines must be designed to automatically identify and flag problems with dose range. Another example involves image processing routines for automated cell counting: such routines must be optimized for cells that grow and respond to drugs in different ways (Figure 3C,D) and must be tested for performance at high and low cell densities. Processing pipelines for the type of data collected in this study are much less developed than the pipelines commonly used for genomics data(Lam, Clark et al. 2011, Bao, Huang et al. 2014, Ashley 2016), but much can be learned from this comparison. For example, computational platforms with provenance such as Galaxy(Goecks, Nekrutenko et al. 2010), Sage Bionetworks' Synapse(Omberg,

bioRxiv preprint doi: https://doi.org/10.1101/213553; this version posted November 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Ellrott et al. 2013), or Cancer Genomics Clouds, have been developed to support data sharing, reproducible analyses, and transparent pipelines, with a primary focus on genomics data. Galaxy also provides a shared platform on which to execute workflows, which serves to eliminate compute environment differences. With sufficient investment in pipeline development it should be possible to adapt such solutions to a wider range of assay types. Image processing algorithms present a unique challenge in that they are frequently proprietary software linked to a specific data acquisition microscope, which complicates common analysis across laboratories; open-source image analysis platforms are therefore preferable in principle(Carpenter, Jones et al. 2006). In the specific case of drug dose-response assays described here, we have developed an online tool at http://GRcalculator.org along with open source scripts for download(Clark, Hafner et al. 2017).

The elements of a workflow for reproducible collection of dose-response data are fairly simple: (i) *Standardization of reagents* including obtaining cell lines directly from repositories such as the ATCC, performing mass spectrometry-based quality control of small molecule drugs, and tracking lot numbers for all media additives; (ii) *Standardized data processing* starting with raw data and metadata through to reporting of final results; and (iii) *Use of automation* to improve reliability and enable experimental designs too complex or labor intensive for humans to execute reliably. The first two points are obvious, but not trivial to implement, because laboratories are not all equipped the same way and some data processing routines are embedded in a non-obvious way in instrument software. A major benefit of automation is that it makes random plate layouts feasible, thereby changing systematic edge effects into random error that has a reduced impact on the dose-response measures. Our data argue that small and relatively inexpensive bench-top dispensing and washing robots have an important role in the performance of reproducible cell-based assays.

The primary contribution of the current study is to show that future execution of reproducible drug dose-response assays in different cell types will require systematic experimentation aimed at establishing the robustness of assays over a full range biological settings and cell types. Such robustness

bioRxiv preprint doi: https://doi.org/10.1101/213553; this version posted November 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

is distinct from conventional measures of assay performance such as precision or repeatability in a single biological setting (Figure 5C). Testing of this type is not routinely performed for the simple reason that establishing and maintaining robust and reproducible assays is time consuming and expensive: we estimate that reproducibility adds ~20% to the total cost of a large-scale study such as drug-response experiments in panels of cell lines(AlQuraishi and Sorger 2016). Iterative experimental design is also essential, even though several leading biomedical scientists have argued that this is not feasible for large-scale studies.(Harris 2017) More generally, despite a push for adherence to the FAIR principles(Wilkinson, Dumontier et al. 2016) there is currently no consensus that the necessary investment is worthwhile, nor do incentives exist in the publication or funding processes for individual research scientists to meet FAIR standards(AlQuraishi and Sorger 2016, Goodspeed, Heiser et al. 2016). In developing these incentives, we must recognize that reproducible research is a public good whose costs are borne by individual investigators and whose benefits are conferred to the community as a whole.

A question raised by our analysis is whether, given their variability and context-dependence, drug response assays performed *in vitro* are useful for understanding drug response in other settings, especially in human patients(Wilding and Bodmer 2014). Concern about the translatability of *in vitro* experiments is long-standing, but we think the current work provides grounds for optimism rather than additional worry. Simply put, if *in vitro* data cannot be reproduced from one laboratory to the next, then it is no wonder that they cannot easily be reproduced in humans; conversely, paying greater attention to accurate and reproducible *in vitro* data is likely to improve translation. Moreover, many of the factors that appear to represent irreproducibility in fact arise from biologically meaningful variation. This includes the time-dependence of drug response, the impact of non-genetic heterogeneity at a single-cell level, and the influence of growth conditions and environmental factors(Cohen, Geva-Zatorsky et al. 2008, Yonesaka, Zejnullahu et al. 2008, Loewer and Lahav 2011, Muranen, Selfors et al. 2012, Wilson, Fridlyand et al. 2012). The simple assays of drug response in current use are unable to correct for such

bioRxiv preprint doi: https://doi.org/10.1101/213553; this version posted November 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

variability, and the problem is made worse by "kit-based science" in which technical validation of assays is left to vendors. However, if the challenge of understanding biological variability at a mechanistic level is embraced, it seems likely that we will improve our ability to conduct *in vitro* assays reproducibly and to apply data obtained in cell lines to human patients.

MATERIAL AND METHODS

Cell lines and drugs

Three isolates of MCF 10A, here referred to as MCF 10A-GM, MCF 10A-OHSU, and MCF 10A-HMS, were sourced independently at three different times from the ATCC. MCF 10A-H2B-mCherry cells were created by inserting an H2B-mCherry expression cassette into the AAVS1 safe harbor genomic locus of MCF 10A-HMS using CRISPR/Cas9(Hafner, Niepel et al. 2016). All lines were cultured in DMEM/F12 base media (Invitrogen #11330-032) supplemented with 5% horse serum (Sigma-Aldrich #H1138), 0.5 μg/mL hydrocortisone (Sigma # H-4001), 20 ng/mL rhEGF (R&D Systems #236-EG), 10 μg/mL insulin (Sigma #I9278), 100 ng/mL cholera toxin (Sigma-Aldrich #C8052), and 100 units/mL penicillin and 100 μg/mL streptomycin (Invitrogen #15140148 or #15140122 or other sources) as described previously(Debnath, Muthuswamy et al. 2003). Base media, horse serum, hydrocortisone, rhEGF, insulin, and cholera toxin where purchased by the MEP-LINCS Center and distributed to the remaining experimental sites. MCF 10A-GM was expanded by Gordon Mills at MD Anderson Cancer Center and distributed to all experimental sites. Cell identity was confirmed at individual experimental sites by short tandem repeat (STR) profiling, and the cells were found to be free of mycoplasma prior to performing experiments.

Drugs were obtained from commercial vendors by HMS LINCS, tested for identity and purity in house as described in detail in the drug collection section of the HMS LINCS Database (http://lincs.hms.harvard.edu/db/sm/), and distributed as 10 mM stock solutions dissolved in DMSO to all experimental sites.

bioRxiv preprint doi: https://doi.org/10.1101/213553; this version posted November 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Drug response experiments and data analysis

The experimental and computational protocols to measure drug response are described in detail in two prior publications(Hafner, Niepel et al. 2017, Niepel, Hafner et al. 2017). The following protocol was suggested for this study: cells were plated at 750 cells per well in 60 µL of media in 384-well plates using automated plate fillers and incubated for 24 h prior to drug addition. Drugs were added at the indicated doses with a D300 Digital Dispenser (Hewlett-Packard), and cells were further incubated for 72 h. At the time of drug addition and at the endpoint of the experiment cells were staining with Hoechst and LIVE/DEAD™ Fixable Red Dead Cell Stain (ThermoFisher Scientific) and cell numbers were determined by imaging as described(Hafner, Niepel et al. 2016, Niepel, Hafner et al. 2017) or by the CellTiter-Glo® assay (Promega). Some details of the experimental protocol differed across Centers and over time, e.g. manually dispensing of drugs or use of 96-well plates.

For live-cell experiments with MCF 10A-H2B-mCherry, cell counts were performed by imaging plates in an 2 hr interval over the course of 96 hours (only first 50 hours shown)(Hafner, Niepel et al. 2016, Niepel, Hafner et al. 2017). Data analysis was performed as described previously (Hafner, Niepel et al. 2016, Hafner, Niepel et al. 2017).

Irregularities in growth across microtiter plates was performed by plating MCF 10A cells at 750 cells per well in 60 µL of media in 384-well plates using automated plate fillers and determining cell numbers after 96 h through imaging as described(Hafner, Niepel et al. 2016, Niepel, Hafner et al. 2017).

.

DECLARATIONS

Acknowledgements

We thank K. Ward for arranging the HP D300 instrument trial at Mount Sinai.

Competing interests

The authors declare that they have no competing interests.

bioRxiv preprint doi: https://doi.org/10.1101/213553; this version posted November 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Funding

This work was funded by grants U54-HL127365 to PKS; U54-HG008098 to RI, MRB, and EAS; R01-GM104184 to MRB; U54HL127624 to MM and AM; U54-HG008100 to JWG, LMH, and JEK; U54-HG008097 to JDJ; U54-NS091046 to CNS; U54-HL127366 to TRG and AD. ADS and AMB were supported by a NIGMS-funded Integrated Pharmacological Sciences Training Program grant T32GM062754.

Authors' contributions

M.N., L.M.H., M.R.B., and E.H.W. designed the study and led its execution; P.K.S supervised the systematic analysis of assay variability and oversaw manuscript preparation. M.H. contributed to experimental design and performed all data analysis. A.M.B. and A.D.S. participated in data collection and initial processing. All authors participated in editing the manuscript.

Materials & Correspondence

Data presented in this paper are included in the additional material and available on the GR browser at http://www.grcalculator.org/grbrowser/.

LINCS Consortium

HMS-LINCS: Caroline E. Shamu[1]. **DTox**: Gomathi Jayaraman[2], Maria Alimova[2], Evren U. Azeloglu[2], Ravi Iyengar[2], Eric A. Sobie[2]. **MEP-LINCS**: Gordon B. Mills[4], James E. Korkola[3], Tiera Liby[3]. **LINCS-PCCSE**: Jacob D. Jaffe[5], Desiree Davison[5], Xiaodong Lu[5]. **LINCS-Transcriptomics**: Todd R. Golub[5], Aravind Subramanian[5]. **NeuroLINCS**: Brandon Shelley[6], Clive N. Svendsen[6]. **DCIC**: Avi Ma'ayan[7], Mario Medvedovic[8].

[4]M.D. Anderson Cancer Center
Systems Biology Department
Unit 1058
P.O. Box 301429
Houston, TX 77030

bioRxiv preprint doi: https://doi.org/10.1101/213553; this version posted November 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

[5]LINCS Proteomic Characterization Center for Signaling and Epigenetics; LINCS Center for Transcriptomics
The Broad Institute
415 Main St.
Cambridge, MA 02142

[6]NeuroLINCS Center
Board of Governors Regenerative Medicine Institute
Cedars-Sinai Medical Center
8700 Beverly Blvd, AHSP 8th Floor
Los Angeles, CA 90048

[7]BD2K-LINCS Data Coordination and Integration Center (DCIC)
Department of Pharmacological Sciences
Mount Sinai Center for Bioinformatics
Icahn School of Medicine at Mount Sinai
One Gustave L. Levy Place Box 1603
New York, NY 10029

[8]Cincinnati Children's Hospital Medical Center
Cincinnati, OH 45229

FIGURE LEGENDS

Figure 1: Results from harmonized drug responses.

Dose-response curves of MCF 10A treated with the MEK1/2 inhibitor Trametinib showing technical replicates (duplicate data collected on the same multi-well plate), biological replicates (averaged data for repeats collected on different days in the same laboratory), and independent experiments (averages of one or more biological replicates performed by different LINCS Centers, where circles represent data from image-based assays and triangles from CellTiter-Glo® assays). Dotted black lines show the averages across all technical replicates, biological replicates, and independent replicates, respectively, in the three plots.

Figure 2: Experimental causes of variability.

(A) Dose-response curves of MCF 10A cells treated with four different drugs measured by image-based cell count or ATP content (CellTiter-Glo®) on the same day by the HMS LINCS Center, which is

bioRxiv preprint doi: https://doi.org/10.1101/213553; this version posted November 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

equivalent to technical replicates. (B) Computed GR_{50} and GR_{max} values corresponding to the plots in A. (C) Representative images of MCF 10A cells treated with vehicle control (DMSO) or 1 µM Palbociclib. (D) Uneven growth of MCF 10A cells in a 384-well plate over the course of three days that demonstrates the presence of edge effects. In the heatmap, color represents the number of cells per well, as assessed by imaging. Plots show deviation from mean number (for the full plate based on the distance from the edge, by column, or by row. Error bars represent the standard deviation.

Figure 3: Technical causes of variability.

(A) Dose-response curves of MCF 10A cells treated with Trametinib or Dasatinib fitted to either the extended dose range (up to 1 µM and 10 µM, respectively) or omitting the last order of magnitude. (B) Computed GR_{50}, GR_{max}, and GR_{inf} values corresponding to the plots in A. (C) Results of cell counting for MCF 10A cells treated with Dasatinib or Neratinib using two different image processing algorithms (denoted as A (red) and B (blue)) included in the Columbus image analysis software package. (D) Number of dead cells (LIVE/DEAD™ Fixable Red Dead Cell Stain positive) and nuclei (Hoechst positive) counted for MCF 10A cells treated with 3.16 µM Dasatinib or 1 µM Neratinib based on the two different algorithms (corresponding to the plots in C).

Figure 4: Changes in drug response related to the underlying biology.

Left panels: Inhibition of MCF 10A growth (12-hour instantaneous GR values) measured in a time-lapse, live-cell experiment involving treatment with multiple doses of Etoposide (top) or Neratinib (bottom). Different colors indicated different drug concentrations ranging from 1 nM (yellow) to 10 µM (blue). Middle panels: Dose-response curves derived from 12-hour GR values computed at 24 (red) and 48 hours (blue) across three biological repeats. Right Panels: Computed GR_{50} and GR_{max} values at 24 and 48 hr.

bioRxiv preprint doi: https://doi.org/10.1101/213553; this version posted November 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure 5: Variability of the response measures across Centers.

(A) Dose-response curves of MCF 10A cells treated with eight drugs measured independently by the five LINCS Centers (circles represent data from image-based assays and triangles from CellTiter-Glo® assays). Dotted black lines show the dose-response curve when all independent replicates were averaged. (B) GR metrics describing the sensitivity of MCF 10A cells to eight drugs measured independently by five LINCS Centers (circles represent data from image-based assays and triangles from CellTiter-Glo® assays). The black line shows the mean sensitivity across all Centers, and the gray area shows the standard error of the mean. (C) Differences in repeatability/precision vs. biological sensitivity/stability/robustness.

Supplemental Figure 1: GR dose-response curve and metrics

Schematic of a dose-response curve under the GR model and the source of the derived metrics.

Supplemental Figure 2: **GR metrics describing the initial experiments to assess sensitivity of MCF10A cells to eight drugs measured at 3 Centers.** Center 3 and Center 4 represent the final results provided by each Center (as in Fig. 5A). Preliminary 1 represents initial experiments run by Center 3, which showed poor agreement with data from the other two Centers (see table to pipeline details). The disparate results reflect in part differences in readout (CTG vs. Imaging) for Etoposide, Vorinostat, Alpelisib, and Palbociclib. Preliminary 2 represents a coordinated effort by Centers 3 and 4: daughter drug plates used by Center 3 in Preliminary 1 experiments were shipped to Center 4 for cell culture. For many drugs, there is good agreement between Preliminary 1 and Preliminary 2, indicating consistency between cell culturing. The consistent discrepancy in responses to Trametinib between Preliminary 1-Preliminary 2 studies and Center 3-Center 4 studies indicate errors in construction of the drug dilution series.

bioRxiv preprint doi: https://doi.org/10.1101/213553; this version posted November 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Supplemental Figure 3: GR metrics across MCF 10A isolates

GR metrics describing the sensitivity of four different MCF 10A cell isolates to eight drugs measured independently at the HMS LINCS center. The black line shows the mean sensitivity measured across all isolates, and the gray box shows the standard error of the mean.

Supplemental Figure 4: Instantaneous GR metrics

Instantaneous GR_{max} (A) and GR_{50} (B) values in MCF 10A cells treated with Etoposide (left) and Neratinib (right) over the course of 24 hr for three biological repeats.

ADDITIONAL MATERIAL

Supplemental Data 1:

Final drug-response data generated by all LINCS Centers (Figure 5).

Supplemental Data 2:

Data generated by the HMS LINCS Center during follow-up experiments (Figures 2-4 and Supplemental Figure 3-4).

bioRxiv preprint doi: https://doi.org/10.1101/213553; this version posted November 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

REFERENCES

AlQuraishi, M. and P. K. Sorger (2016). "Reproducibility will only come with data liberation." Sci Transl Med **8**(339): 339ed337.

Arrowsmith, J. (2011). "Trial watch: Phase II failures: 2008-2010." Nat Rev Drug Discov **10**(5): 328-329.

Ashley, E. A. (2016). "Towards precision medicine." Nat Rev Genet **17**(9): 507-522.

Baker, M. (2016). "Biotech giant publishes failures to confirm high-profile science." Nature **530**(7589): 141.

Bao, R., L. Huang, J. Andrade, W. Tan, W. A. Kibbe, H. Jiang and G. Feng (2014). "Review of current methods, applications, and data management for the bioinformatics analysis of whole exome sequencing." Cancer Inform **13**(Suppl 2): 67-82.

Barretina, J., G. Caponigro, N. Stransky, K. Venkatesan, A. A. Margolin, S. Kim, C. J. Wilson, J. Lehar, G. V. Kryukov, D. Sonkin, A. Reddy, M. Liu, L. Murray, M. F. Berger, J. E. Monahan, P. Morais, J. Meltzer, A. Korejwa, J. Jane-Valbuena, F. A. Mapa, J. Thibault, E. Bric-Furlong, P. Raman, A. Shipway, I. H. Engels, J. Cheng, G. K. Yu, J. Yu, P. Aspesi, Jr., M. de Silva, K. Jagtap, M. D. Jones, L. Wang, C. Hatton, E. Palescandolo, S. Gupta, S. Mahan, C. Sougnez, R. C. Onofrio, T. Liefeld, L. MacConaill, W. Winckler, M. Reich, N. Li, J. P. Mesirov, S. B. Gabriel, G. Getz, K. Ardlie, V. Chan, V. E. Myer, B. L. Weber, J. Porter, M. Warmuth, P. Finan, J. L. Harris, M. Meyerson, T. R. Golub, M. P. Morrissey, W. R. Sellers, R. Schlegel and L. A. Garraway (2012). "The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity." Nature **483**(7391): 603-607.

Begley, C. G. and L. M. Ellis (2012). "Drug development: Raise standards for preclinical cancer research." Nature **483**(7391): 531-533.

Bouhaddou, M., M. S. DiStefano, E. A. Riesel, E. Carrasco, H. Y. Holzapfel, D. C. Jones, G. R. Smith, A. D. Stern, S. S. Somani, T. V. Thompson and M. R. Birtwistle (2016). "Drug response consistency in CCLE and CGP." Nature **540**(7631): E9-E10.

bioRxiv preprint doi: https://doi.org/10.1101/213553; this version posted November 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Bushway, P. J., B. Azimi, S. Heynen-Genel, J. H. Price and M. Mercola (2010). "Hybrid median filter background estimator for correcting distortions in microtiter plate data." Assay Drug Dev Technol **8**(2): 238-250.

Carpenter, A. E., T. R. Jones, M. R. Lamprecht, C. Clarke, I. H. Kang, O. Friman, D. A. Guertin, J. H. Chang, R. A. Lindquist, J. Moffat, P. Golland and D. M. Sabatini (2006). "CellProfiler: image analysis software for identifying and quantifying cell phenotypes." Genome Biol **7**(10): R100.

Caruso, J. A., J. J. Reiners, Jr., J. Emond, T. Shultz, M. A. Tainsky, M. Alaoui-Jamali and G. Batist (2001). "Genetic alteration of chromosome 8 is a common feature of human mammary epithelial cell lines transformed in vitro with benzo[a]pyrene." Mutat Res **473**(1): 85-99.

Clark, N. A., M. Hafner, M. Kouril, E. H. Williams, J. L. Muhlich, M. Pilarczyk, M. Niepel, P. K. Sorger and M. Medvedovic (2017). "GRcalculator: an online tool for calculating and mining dose-response data." BMC Cancer **in press**.

Cohen, A. A., N. Geva-Zatorsky, E. Eden, M. Frenkel-Morgenstern, I. Issaeva, A. Sigal, R. Milo, C. Cohen-Saidon, Y. Liron, Z. Kam, L. Cohen, T. Danon, N. Perzov and U. Alon (2008). "Dynamic proteomics of individual cancer cells in response to a drug." Science **322**(5907): 1511-1516.

Consortium, C. C. L. E. and G. o. D. S. i. C. Consortium (2015). "Pharmacogenomic agreement between two cancer cell line data sets." Nature **528**(7580): 84-87.

Cowell, J. K., J. LaDuca, M. R. Rossi, T. Burkhardt, N. J. Nowak and S. Matsui (2005). "Molecular characterization of the t(3;9) associated with immortalization in the MCF10A cell line." Cancer Genet Cytogenet **163**(1): 23-29.

Coyle, M. P., Jr., D. P. Green and E. H. Monsanto (1989). "Advances in carpal bone injury and disease." Hand Clin **5**(3): 471-486.

Cravatt, B. F. and J. M. Gottesfeld (2010). "Chemical biology meets biological chemistry minireview series." J Biol Chem **285**(15): 11031-11032.

Debnath, J., S. K. Muthuswamy and J. S. Brugge (2003). "Morphogenesis and oncogenesis of MCF-10A mammary epithelial acini grown in three-dimensional basement membrane cultures." Methods **30**(3): 256-268.

bioRxiv preprint doi: https://doi.org/10.1101/213553; this version posted November 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

eLIFE-Editorial (2017). "The challenges of replication." Elife **6**.

Errington, T. M., E. Iorns, W. Gunn, F. E. Tan, J. Lomax and B. A. Nosek (2014). "An open investigation of the reproducibility of cancer biology research." Elife **3**.

Fallahi-Sichani, M., V. Becker, B. Izar, G. J. Baker, J. R. Lin, S. A. Boswell, P. Shah, A. Rotem, L. A. Garraway and P. K. Sorger (2017). "Adaptive resistance of melanoma cells to RAF inhibition via reversible induction of a slowly dividing de-differentiated state." Mol Syst Biol **13**(1): 905.

Fletcher, J. I., M. Haber, M. J. Henderson and M. D. Norris (2010). "ABC transporters in cancer: more than just drug efflux pumps." Nat Rev Cancer **10**(2): 147-156.

Garnett, M. J., E. J. Edelman, S. J. Heidorn, C. D. Greenman, A. Dastur, K. W. Lau, P. Greninger, I. R. Thompson, X. Luo, J. Soares, Q. Liu, F. Iorio, D. Surdez, L. Chen, R. J. Milano, G. R. Bignell, A. T. Tam, H. Davies, J. A. Stevenson, S. Barthorpe, S. R. Lutz, F. Kogera, K. Lawrence, A. McLaren-Douglas, X. Mitropoulos, T. Mironenko, H. Thi, L. Richardson, W. Zhou, F. Jewitt, T. Zhang, P. O'Brien, J. L. Boisvert, S. Price, W. Hur, W. Yang, X. Deng, A. Butler, H. G. Choi, J. W. Chang, J. Baselga, I. Stamenkovic, J. A. Engelman, S. V. Sharma, O. Delattre, J. Saez-Rodriguez, N. S. Gray, J. Settleman, P. A. Futreal, D. A. Haber, M. R. Stratton, S. Ramaswamy, U. McDermott and C. H. Benes (2012). "Systematic identification of genomic markers of drug sensitivity in cancer cells." Nature **483**(7391): 570-575.

Goecks, J., A. Nekrutenko, J. Taylor and T. Galaxy (2010). "Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences." Genome Biol **11**(8): R86.

Goodspeed, A., L. M. Heiser, J. W. Gray and J. C. Costello (2016). "Tumor-Derived Cell Lines as Molecular Models of Cancer Pharmacogenomics." Mol Cancer Res **14**(1): 3-13.

Hafner, M., M. Niepel, M. Chung and P. K. Sorger (2016). "Growth rate inhibition metrics correct for confounders in measuring sensitivity to cancer drugs." Nat Methods **13**(6): 521-527.

Hafner, M., M. Niepel and P. K. Sorger (2017). "Alternative drug sensitivity metrics improve preclinical cancer pharmacogenomics." Nat Biotechnol **35**(6): 500-502.

bioRxiv preprint doi: https://doi.org/10.1101/213553; this version posted November 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Hafner, M., M. Niepel, K. Subramanian and P. K. Sorger (2017). "Designing Drug-Response Experiments and Quantifying their Results." Curr Protoc Chem Biol **9**(2): 96-116.

Haibe-Kains, B., N. El-Hachem, N. J. Birkbak, A. C. Jin, A. H. Beck, H. J. Aerts and J. Quackenbush (2013). "Inconsistency in large pharmacogenomic studies." Nature **504**(7480): 389-393.

Harris, E. A., E. J. Koh, J. Moffat and D. R. McMillen (2016). "Automated inference procedure for the determination of cell growth parameters." Phys Rev E **93**(1): 012402.

Harris, L. A., P. L. Frick, S. P. Garbett, K. N. Hardeman, B. B. Paudel, C. F. Lopez, V. Quaranta and D. R. Tyson (2016). "An unbiased metric of antiproliferative drug effect in vitro." Nat Methods **13**(6): 497-500.

Harris, R. F. (2017). Rigor mortis : how sloppy science creates worthless cures, crushes hope, and wastes billions. New York, Basic Books.

Haverty, P. M., E. Lin, J. Tan, Y. Yu, B. Lam, S. Lianoglou, R. M. Neve, S. Martin, J. Settleman, R. L. Yauch and R. Bourgon (2016). "Reproducible pharmacogenomic profiling of cancer cell line panels." Nature **533**(7603): 333-337.

Heiser, L. M., A. Sadanandam, W. L. Kuo, S. C. Benz, T. C. Goldstein, S. Ng, W. J. Gibb, N. J. Wang, S. Ziyad, F. Tong, N. Bayani, Z. Hu, J. I. Billig, A. Dueregger, S. Lewis, L. Jakkula, J. E. Korkola, S. Durinck, F. Pepin, Y. Guan, E. Purdom, P. Neuvial, H. Bengtsson, K. W. Wood, P. G. Smith, L. T. Vassilev, B. T. Hennessy, J. Greshock, K. E. Bachman, M. A. Hardwicke, J. W. Park, L. J. Marton, D. M. Wolf, E. A. Collisson, R. M. Neve, G. B. Mills, T. P. Speed, H. S. Feiler, R. F. Wooster, D. Haussler, J. M. Stuart, J. W. Gray and P. T. Spellman (2012). "Subtype and pathway specific responses to anticancer compounds in breast cancer." Proc Natl Acad Sci U S A **109**(8): 2724-2729.

Ioannidis, J. P. (2017). "Acknowledging and Overcoming Nonreproducibility in Basic and Preclinical Research." JAMA.

Kim, Y. M., S. Yang, W. Xu, S. Li and X. Yang (2008). "Continuous in vitro exposure to low-dose genistein induces genomic instability in breast epithelial cells." Cancer Genet Cytogenet **186**(2): 78-84.

bioRxiv preprint doi: https://doi.org/10.1101/213553; this version posted November 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Lam, H. Y., M. J. Clark, R. Chen, R. Chen, G. Natsoulis, M. O'Huallachain, F. E. Dewey, L. Habegger, E. A. Ashley, M. B. Gerstein, A. J. Butte, H. P. Ji and M. Snyder (2011). "Performance comparison of whole-genome sequencing platforms." Nat Biotechnol **30**(1): 78-82.

Loewer, A. and G. Lahav (2011). "We are all individuals: causes and consequences of non-genetic heterogeneity in mammalian cells." Curr Opin Genet Dev **21**(6): 753-758.

Marella, N. V., K. S. Malyavantham, J. Wang, S. Matsui, P. Liang and R. Berezney (2009). "Cytogenetic and cDNA microarray expression analysis of MCF10 human breast cancer progression cell lines." Cancer Res **69**(14): 5946-5953.

Morrison, S. J. (2014). "Time to do something about reproducibility." Elife **3**.

Muranen, T., L. M. Selfors, D. T. Worster, M. P. Iwanicki, L. Song, F. C. Morales, S. Gao, G. B. Mills and J. S. Brugge (2012). "Inhibition of PI3K/mTOR leads to adaptive resistance in matrix-attached cancer cells." Cancer Cell **21**(2): 227-239.

Nature-Editorial (2017). "Replication studies offer much more than technical details." Nature **541**(7637): 259-260.

Niepel, M., M. Hafner, M. Chung and P. K. Sorger (2017). "Measuring Cancer Drug Sensitivity and Resistance in Cultured Cells." Curr Protoc Chem Biol **9**(2): 55-74.

Nosek, B. A. and T. M. Errington (2017). "Making sense of replications." Elife **6**.

Omberg, L., K. Ellrott, Y. Yuan, C. Kandoth, C. Wong, M. R. Kellen, S. H. Friend, J. Stuart, H. Liang and A. A. Margolin (2013). "Enabling transparent and collaborative computational analysis of 12 tumor types within The Cancer Genome Atlas." Nat Genet **45**(10): 1121-1126.

Orth, J. D., R. H. Kohler, F. Foijer, P. K. Sorger, R. Weissleder and T. J. Mitchison (2011). "Analysis of mitosis and antimitotic drug responses in tumors by in vivo microscopy and single-cell pharmacodynamics." Cancer Res **71**(13): 4608-4616.

Prinz, F., T. Schlange and K. Asadullah (2011). "Believe it or not: how much can we rely on published data on potential drug targets?" Nat Rev Drug Discov **10**(9): 712.

bioRxiv preprint doi: https://doi.org/10.1101/213553; this version posted November 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Ramirez, M., S. Rajaram, R. J. Steininger, D. Osipchuk, M. A. Roth, L. S. Morinishi, L. Evans, W. Ji, C. H. Hsu, K. Thurley, S. Wei, A. Zhou, P. R. Koduru, B. A. Posner, L. F. Wu and S. J. Altschuler (2016). "Diverse drug-resistance mechanisms can emerge from drug-tolerant cancer persister cells." Nat Commun **7**: 10690.

Roytta, M., K. M. Laine and P. Harkonen (1987). "Morphological studies on the effect of taxol on cultured human prostatic cancer cells." Prostate **11**(1): 95-106.

Safikhani, Z., M. Freeman, P. Smirnov, N. El-Hachem, A. She, R. Quevedo, A. Goldenberg, N. Juul Birkbak, C. Hatzis, L. Shi, A. H. Beck, H. J. Aerts, J. Quackenbush and B. Haibe-Kains (2015). "Revisiting inconsistency in large pharmacogenomic studies." bioRxiv.

Salani, B., C. Marini, A. D. Rio, S. Ravera, M. Massollo, A. M. Orengo, A. Amaro, M. Passalacqua, S. Maffioli, U. Pfeffer, R. Cordera, D. Maggi and G. Sambuceti (2013). "Metformin impairs glucose consumption and survival in Calu-1 cells by direct inhibition of hexokinase-II." Sci Rep **3**: 2070.

Sandve, G. K., A. Nekrutenko, J. Taylor and E. Hovig (2013). "Ten simple rules for reproducible computational research." PLoS Comput Biol **9**(10): e1003285.

Schenone, M., V. Dancik, B. K. Wagner and P. A. Clemons (2013). "Target identification and mechanism of action in chemical biology and drug discovery." Nat Chem Biol **9**(4): 232-240.

Seashore-Ludlow, B., M. G. Rees, J. H. Cheah, M. Cokol, E. V. Price, M. E. Coletti, V. Jones, N. E. Bodycombe, C. K. Soule, J. Gould, B. Alexander, A. Li, P. Montgomery, M. J. Wawer, N. Kuru, J. D. Kotz, C. S. Hon, B. Munoz, T. Liefeld, V. Dancik, J. A. Bittker, M. Palmer, J. E. Bradner, A. F. Shamji, P. A. Clemons and S. L. Schreiber (2015). "Harnessing Connectivity in a Large-Scale Small-Molecule Sensitivity Dataset." Cancer Discov **5**(11): 1210-1223.

Soliman, G., S. Steenson and A. Etekpo (2016). "Effects of Metformin and a Mammalian Target of Rapamycin (mTOR) ATPCompetitive Inhibitor on Targeted Metabolomics in Pancreatic Cancer Cell Line." Metabolomics **6**(183).

Soule, H. D., T. M. Maloney, S. R. Wolman, W. D. Peterson, Jr., R. Brenz, C. M. McGrath, J. Russo, R. J. Pauley, R. F. Jones and S. C. Brooks (1990). "Isolation and characterization of a spontaneously immortalized human breast epithelial cell line, MCF-10." Cancer Res **50**(18): 6075-6086.

bioRxiv preprint doi: https://doi.org/10.1101/213553; this version posted November 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Tolliday, N. (2010). "High-throughput assessment of Mammalian cell viability by determination of adenosine triphosphate levels." Curr Protoc Chem Biol **2**(3): 153-161.

Wilding, J. L. and W. F. Bodmer (2014). "Cancer cell lines for drug discovery and development." Cancer Res **74**(9): 2377-2384.

Wilkinson, M. D., M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J. W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. t Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S. A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao and B. Mons (2016). "The FAIR Guiding Principles for scientific data management and stewardship." Sci Data **3**: 160018.

Wilson, T. R., J. Fridlyand, Y. Yan, E. Penuel, L. Burton, E. Chan, J. Peng, E. Lin, Y. Wang, J. Sosman, A. Ribas, J. Li, J. Moffat, D. P. Sutherlin, H. Koeppen, M. Merchant, R. Neve and J. Settleman (2012). "Widespread potential for growth-factor-driven resistance to anticancer kinase inhibitors." Nature **487**(7408): 505-509.

Yonesaka, K., K. Zejnullahu, N. Lindeman, A. J. Homes, D. M. Jackman, F. Zhao, A. M. Rogers, B. E. Johnson and P. A. Janne (2008). "Autocrine production of amphiregulin predicts sensitivity to both gefitinib and cetuximab in EGFR wild-type cancers." Clin Cancer Res **14**(21): 6963-6973.

Zientek-Targosz, H., D. Kunnev, L. Hawthorn, M. Venkov, S. Matsui, R. T. Cheney and Y. Ionov (2008). "Transformation of MCF-10A cells by random mutagenesis with frameshift mutagen ICR191: a model for identifying candidate breast-tumor suppressors." Mol Cancer **7**: 51.

bioRxiv preprint doi: https://doi.org/10.1101/213553; this version posted November 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

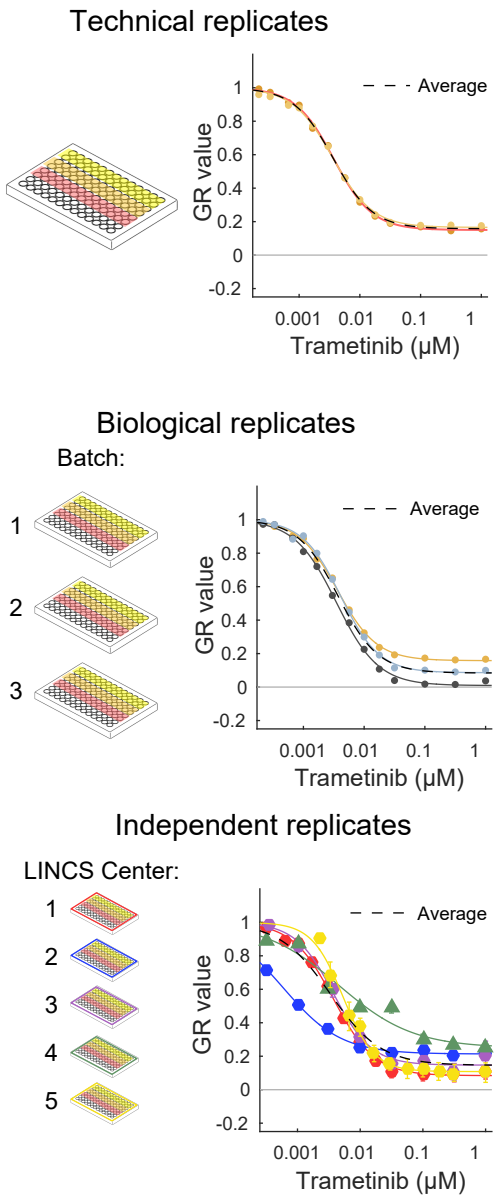
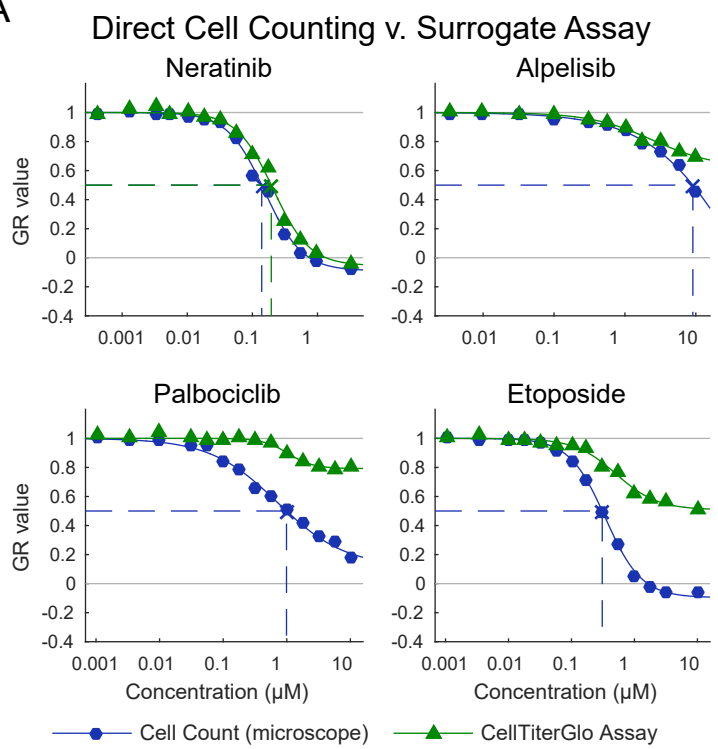


Figure 1

bioRxiv preprint doi: https://doi.org/10.1101/213553; this version posted November 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

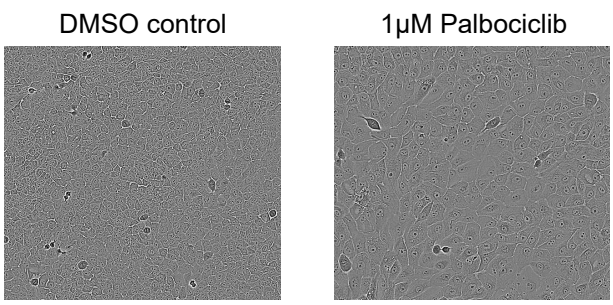
A



B

	GR_{50} (μM)		GR_{max}	
Neratinib	0.14	0.20	-0.07	-0.04
Alpelisib	9.08	Inf	0.46	0.70
Palbociclib	0.99	Inf	0.17	0.78
Etoposide	0.31	Inf	-0.06	0.51
	cell count		CTG Assay	

C



D

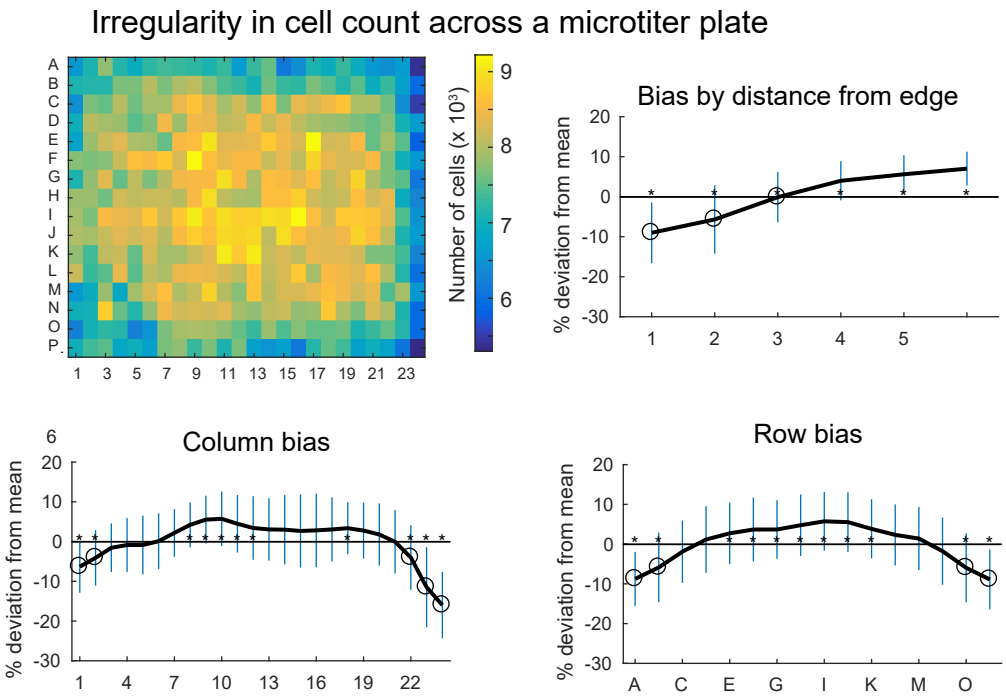
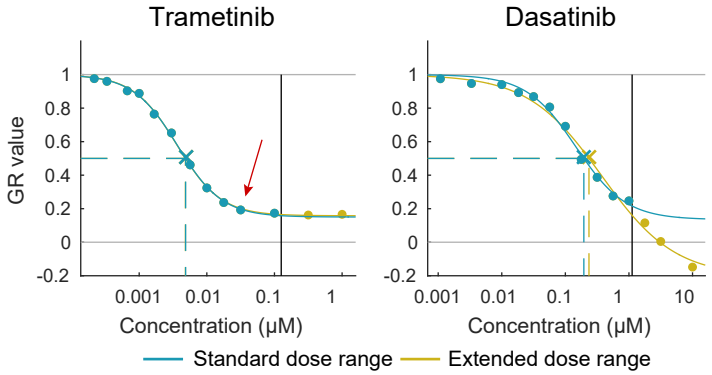


Figure 2

bioRxiv preprint doi: https://doi.org/10.1101/213553; this version posted November 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

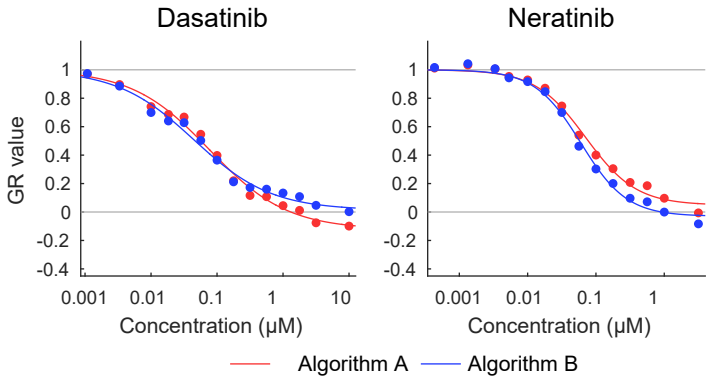
A Impact of dose range on dose-response parameters



B

	Trametinib		Dasatinib	
GR_{50} (μM)	0.00	0.00	0.20	0.24
GR_{max}	0.17	0.16	0.25	-0.15
GR_{inf}	0.15	0.16	0.13	-0.20
	Standard range		Extended range	

C Impact of image processing algorithm on cell count



D

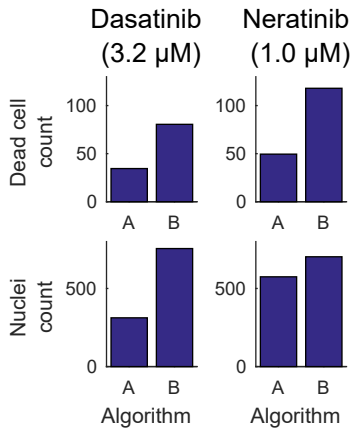
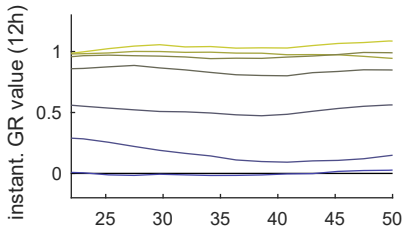


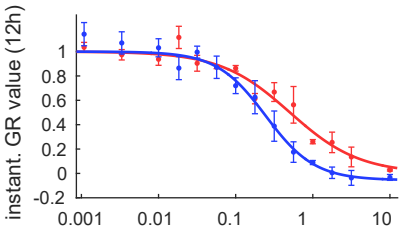
Figure 3

bioRxiv preprint doi: https://doi.org/10.1101/213553; this version posted November 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Impact of time on drug sensitivity

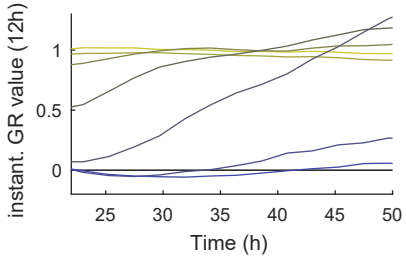
Etoposide

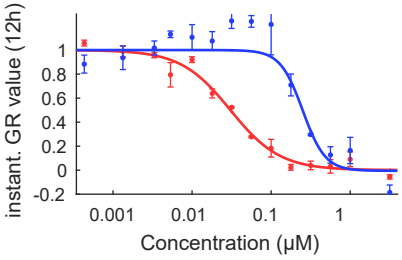




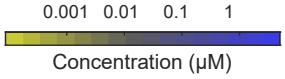
GR_{50} (μM)	0.51	0.23
GR_{max}	0.03	-0.04

Neratinib





GR_{50} (μM)	0.03	0.25
GR_{max}	-0.06	-0.19

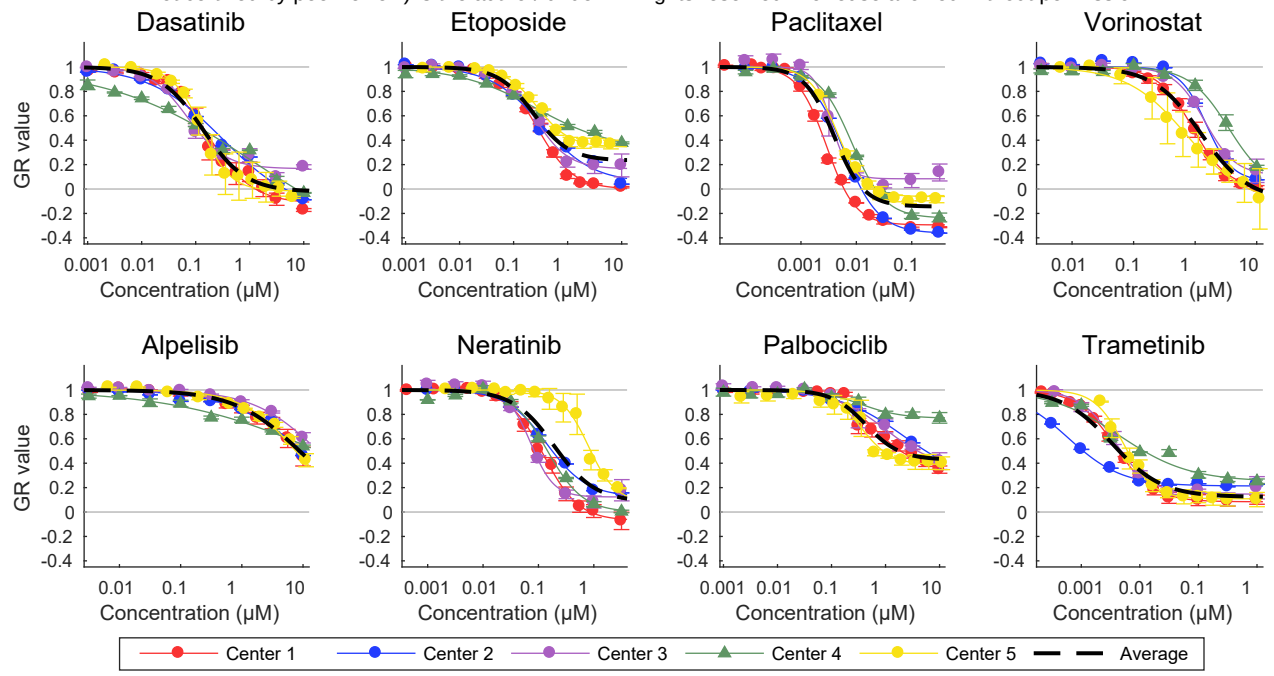


Time=24h Time=48h

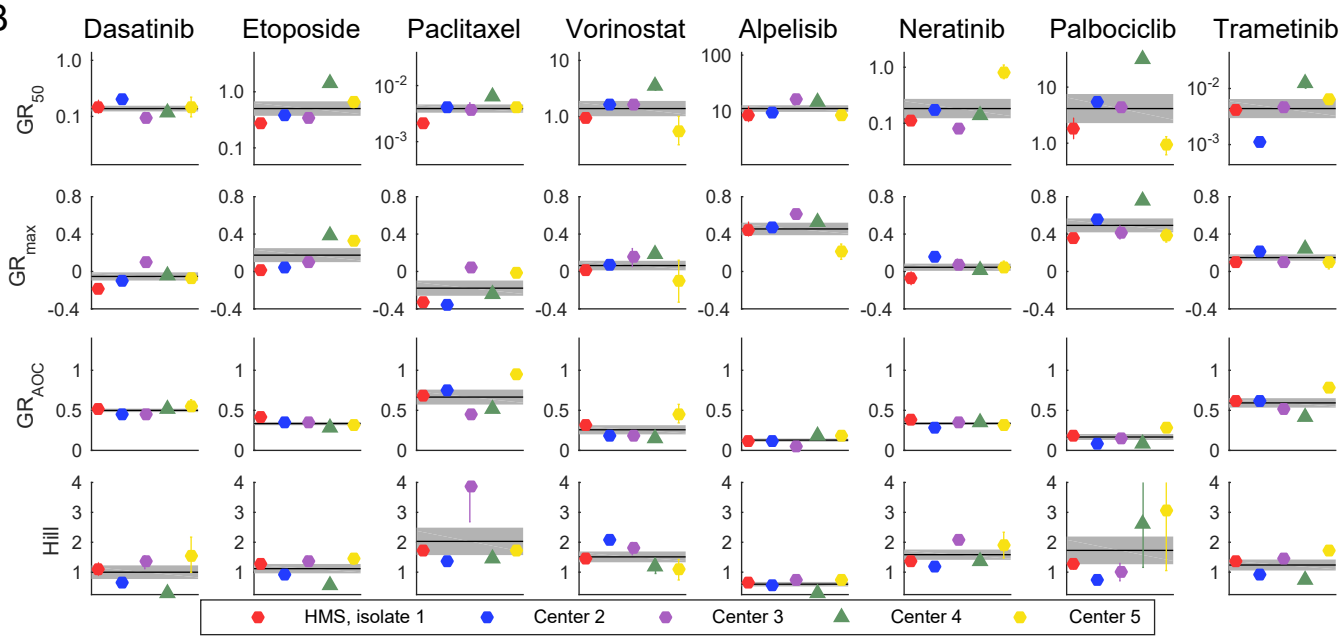
Figure 4

bioRxiv preprint doi: https://doi.org/10.1101/213553; this version posted November 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

A



B



C

Assay precision, reproducibility and biological robustness

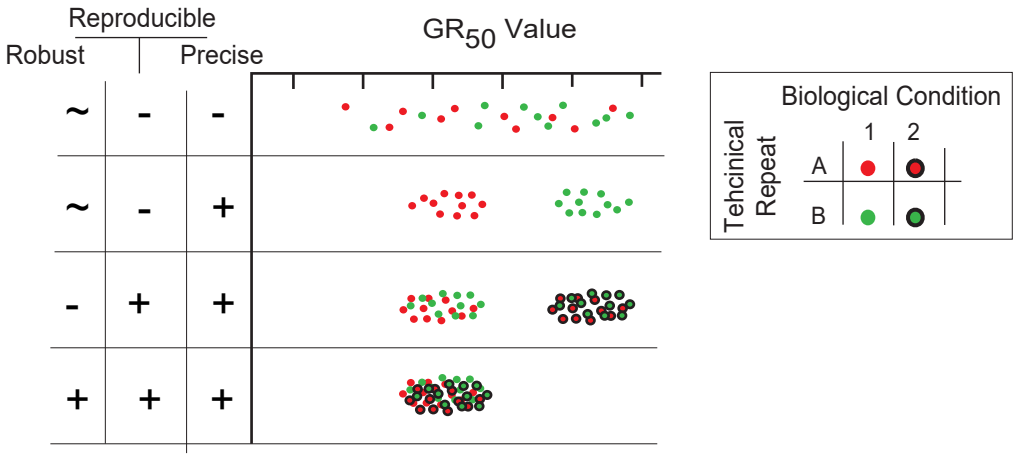
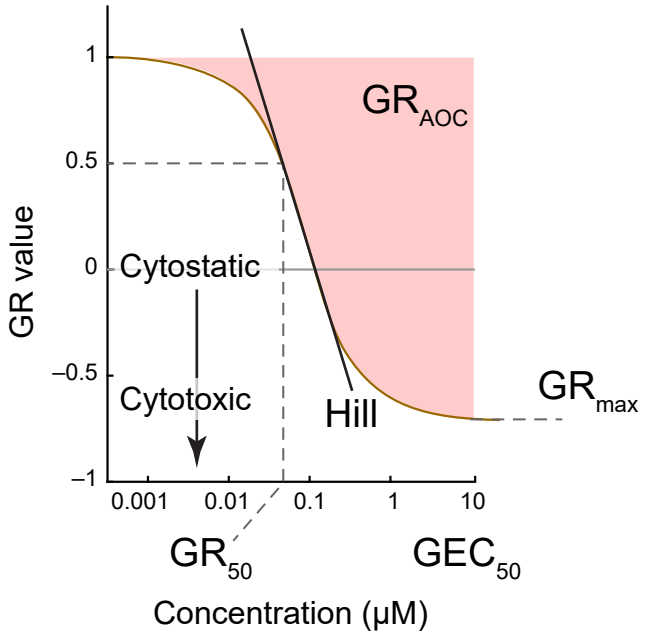


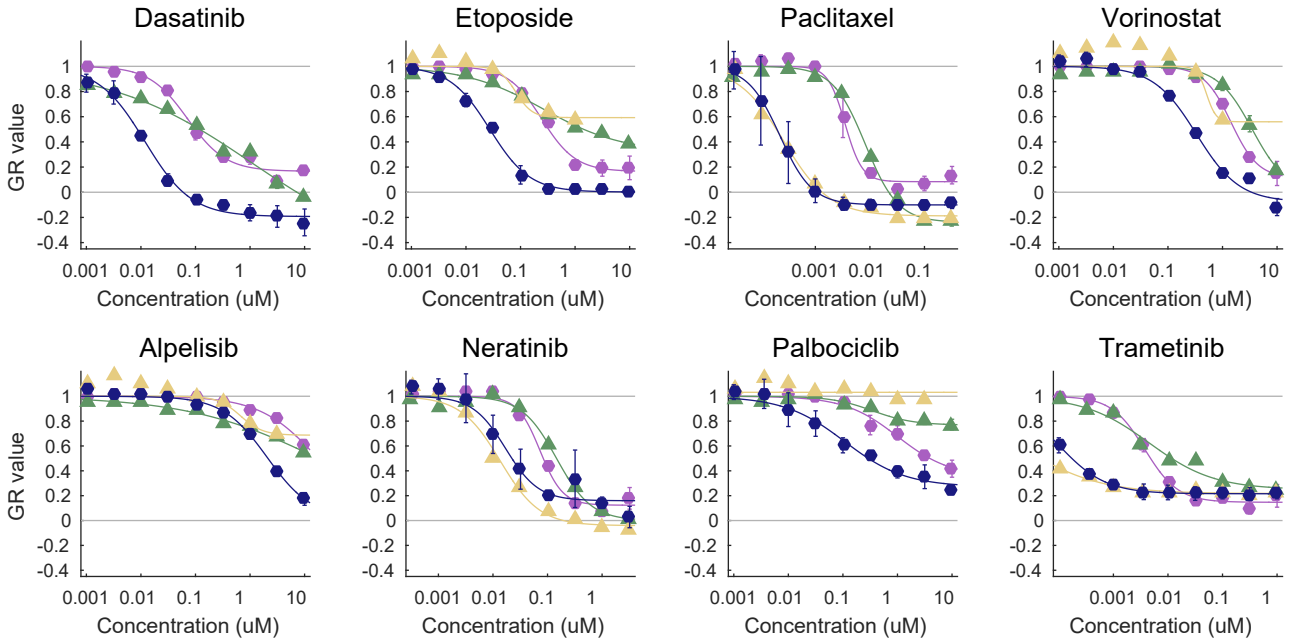
Figure 5

bioRxiv preprint doi: https://doi.org/10.1101/213553; this version posted November 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



Supplemental Figure 1

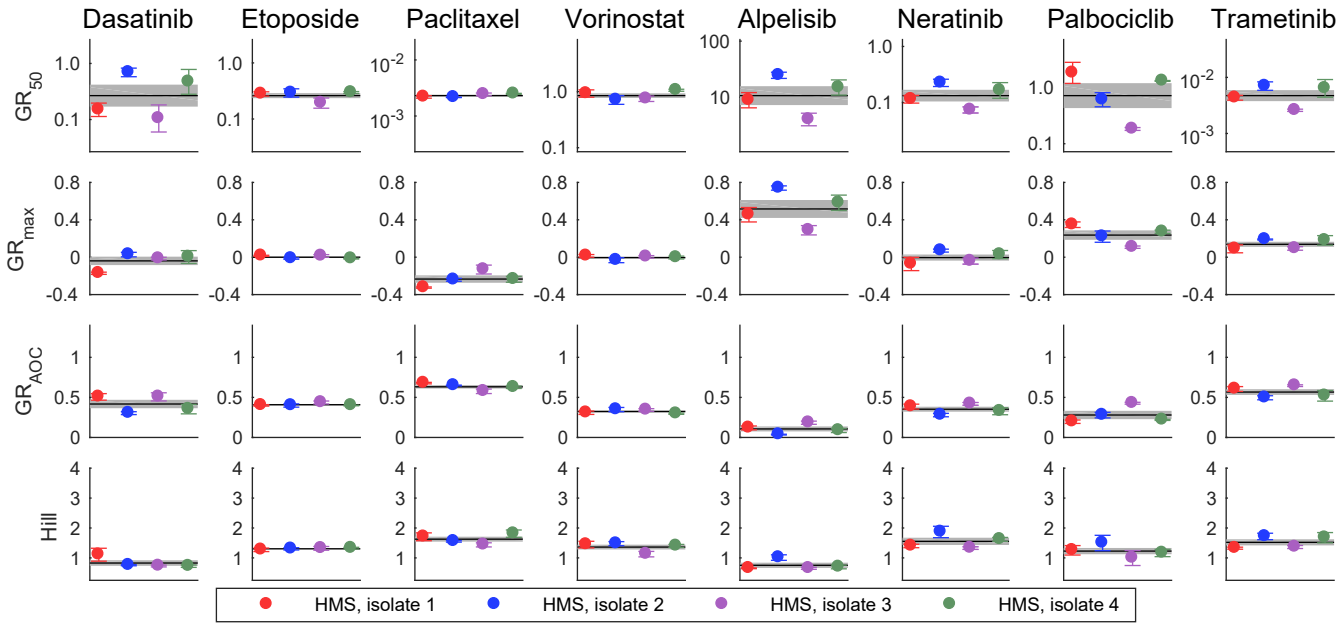
bioRxiv preprint doi: https://doi.org/10.1101/213553; this version posted November 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



	Drug dilution performed at:	Cell culture performed at:	Readout assay:	
Preliminary 1	Center 3	Center 3	Imaging	
Preliminary 2	Center 3	Center 4	CTG	
Center 3	Center 3	Center 3	Imaging	
Center 4	Center 4	Center 4	CTG	

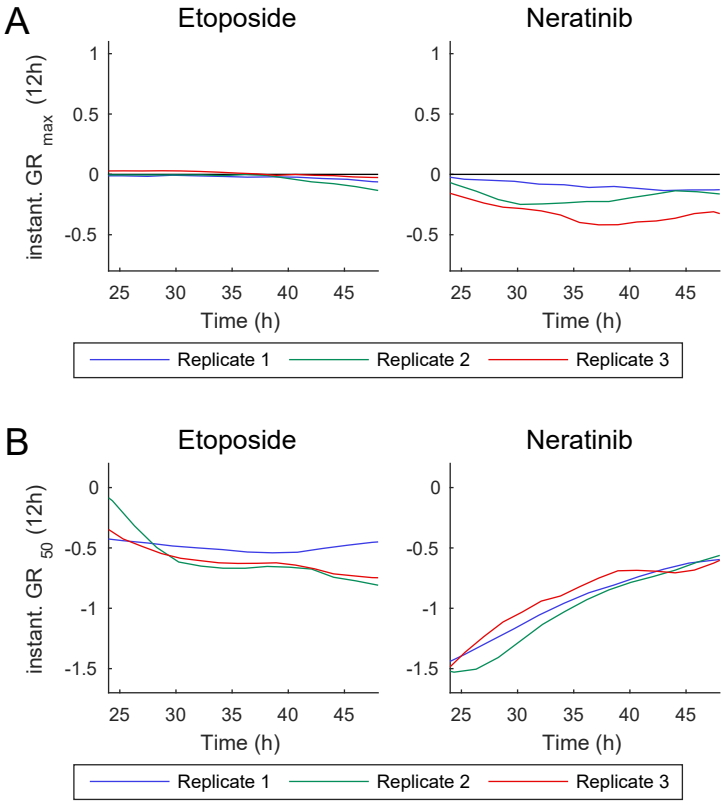
Supplemental Figure 2

bioRxiv preprint doi: https://doi.org/10.1101/213553; this version posted November 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



Supplemental Figure 3

bioRxiv preprint doi: https://doi.org/10.1101/213553; this version posted November 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



Supplemental Figure 4