Published in final edited form as:

Nat Biotechnol. 2016 January ; 34(1): 70–77. doi:10.1038/nbt.3419.

Improving drug discovery with high-content phenotypic screens by systematic selection of reporter cell lines

Jungseog Kang[#1,2], **Chien-Hsiang Hsu**[#1,3,5], **Qi Wu**[1], **Shanshan Liu**[1], **Adam D. Coster**[1], **Bruce A. Posner**[4], **Steven J. Altschuler**[1,5,#], and **Lani F. Wu**[1,5,#]

[1]Green Center for Systems Biology, University of Texas Southwestern Medical Center, Dallas, TX 75390, USA.

[2]Department of Biology, Arts and Science, New York University-Shanghai, Shanghai, 200122, China.

[3]Simmons Cancer Center, University of Texas Southwestern Medical Center, Dallas, TX 75390, USA.

[4]Department of Biochemistry, University of Texas Southwestern Medical Center, Dallas, TX 75390, USA.

[5]Department of Pharmaceutical Chemistry, University of California, San Francisco, San Francisco, CA 94158, USA.

[#] These authors contributed equally to this work.

Abstract

High-content, image-based screens enable the identification of compounds that induce cellular responses similar to those of known drugs but through different chemical structures or targets. A central challenge in designing phenotypic screens is choosing suitable imaging biomarkers. Here we present a method for systematically identifying optimal reporter cell lines for annotating compound libraries (ORACLs), whose phenotypic profiles most accurately classify a training set of known drugs. We generate a library of fluorescently tagged reporter cell lines, and let analytical criteria determine which among them—the ORACL—best classifies compounds into multiple, diverse drug classes. We demonstrate that an ORACL can functionally annotate large compound libraries across diverse drug classes in a single-pass screen and confirm high prediction accuracy via orthogonal, secondary validation assays. Our approach will increase the efficiency, scale and accuracy of phenotypic screens by maximizing their discriminatory power.

[#]To whom correspondence should be addressed: ; Email: steven.altschuler@ucsf.edu, ; Email: lani.wu@ucsf.edu

Author contributions

J.K., Q.W., and S.L. generated the reporter library; A.D.C. built pSEG; J.K. designed the experiments; J.K. and C.H.H. performed the experiments; B.A.P. helped perform the HTS experiments; C.H.H. performed the data analysis; J.K., C.H.H., L.F.W. and S.J.A. wrote the manuscript; and L.F.W. and S.J.A. guided all aspects of this study.

Competing Financial Interests Statement

S.J.A. and L.F.W. have submitted a patent application.

Introduction

Advances in molecular biology have led to an unprecedented ability to profile the genetic- and pathway-level changes that occur in disease[1-4]. However, most of this information has yet to be exploited in drug development, particularly for drugs that are targeted to patient subpopulations, that reduce the side effects of existing drugs and that provide second-line treatment if drug resistance emerges[5,6]. One strategy for discovering such drugs is to search existing large chemical libraries[7-13] for new leads whose activity profiles are similar, but not identical, to those of proven drugs. These compounds may have distinct chemical structures and operate through different mechanisms. The main challenge when using large chemical libraries is how to search them efficiently in ways that scale with the size of the library and the desired number of new drug classes. An efficient approach would be able to classify compounds into different drug classes targeting distinct cellular pathways in a single screening pass.

Purely computational approaches have been used to perform virtual screens across multiple mechanisms of action[14,15], but predictions of chemical mechanism may poorly or non-specifically predict biological activity (e.g. a predicted kinase inhibitor could affect receptor signaling, cell growth, cytoskeletal structure and many other biological processes). Current biochemical screening approaches[16] are not designed for diversifying the repertoire of compounds within or across cellular processes in a single-pass screen; rather, multiple passes would be required to screen a large compound library, with each pass focused on a different target. Likewise, many current “low-dimensional” phenotypic screening approaches use readouts that are either too specific (e.g. single target[17]) or broad (e.g. cell proliferation or death[18]) to distinguish simultaneously among different mechanistic modes of action in a single-pass screen.

High-content phenotypic screens hold promise for identifying lead compounds across multiple drug classes at a single-pass screen. Multi-parametric measures of cellular responses are captured and summarized succinctly as “phenotypic (or cytological) profiles”[19] or “fingerprints”[20,21] and used to group compounds by similarity of their induced cellular responses. Phenotypic profiles have proven their usefulness in partitioning drug libraries into functional classes and predicting mechanism of action using guilt-by-association[19,22-25]. However, assay costs for current approaches based on transcriptomics[26,27] or proteomics[28-30] are too expensive to be scaled routinely to libraries with tens or hundreds of thousands of compounds[31,32],.

High-content imaging[13,19,25,33-35] is an appealing modality due to its relatively lower costs and ability to monitor systems-level responses in individual cells. A key step in every phenotypic screen is the selection of biomarkers (e.g. antibodies, chemical dyes or genetically encoded fluorescent tags). In fluorescent microscopy, only a relatively small number of biomarkers can be monitored simultaneously in each cell. Multiplexing biomarkers and/or performing additional replicate experiments can increase the number of readouts used to probe cellular responses and provide useful information[36,37]. However, increasing the number of biomarkers can lead to greatly increased costs and time for screening. Notably, there is currently no established strategy for systematically identifying a

minimal biomarker set that can accurately classify compounds across multiple, specified drug classes.

The identification of “optimal” drug classification biomarkers could be addressed for either fixed- or live-cell imaging assays. Fixed-cell assays have the advantage that a wide selection of immunofluorescent (IF) probes are available that can report on the expression or activity of a protein. Additionally, sample preparation and image acquisition steps can be decoupled. On the other hand, live-cell assays avoid time-consuming fixation steps, costly IF probes and the need to perform replicate experiments across multiple time points.

In our current study, scalability is a central goal; hence, we chose to focus on phenotypic profiling based on live-cell reporters. The key challenge, then, is how to identify reporter cell lines whose phenotypic profiles best enable accurate classification of compounds across diverse drug classes. We address this challenge in three steps (Fig. 1). First, we construct a library of live-cell reporter cell lines that are fluorescently tagged for genes involved in a wide variety of biological functions. Next, we use analytical criteria to identify the reporter cell line in this library whose phenotypic profiles most accurately classify training drugs across multiple drug classes. (In this study, we focused on cancer-related drug classes.) Finally, we demonstrate that this single reporter cell line, in a single-pass screen, can accurately identify lead compounds across diverse drug classes. We refer to this informative reporter cell line as an ORACL, for “Optimal Reporter cell line for Annotating Compounds Libraries,” as classifying compounds into specified drug classes effectively provides functional annotation for a drug library.

Results

Construction of reporter cell line library for live-cell drug screening

To enable high-content profiling of large-scale compound libraries, we made use of a collection of triply-labeled live-cell reporter cell lines (Fig. 2a, left). The first two labels facilitated automated identification of cellular regions and extraction of information about their morphology. This was established by using a plasmid for cell image Segmentation (pSeg; Online Methods), which demarked the whole cell (mCherry fluorescent protein (RFP)) and nucleus (Histone H2B fused to cyan fluorescent protein (CFP)). The third label allowed each reporter cell line to monitor the expression of a different protein. This was established using Central Dogma (CD)-tagging[38], a genomic-scale approach for randomly labeling different full-length proteins (marked by inserting yellow fluorescent protein (YFP) as an extra exon; Online Methods). CD-tagged proteins typically express at endogenous levels and have preserved functionality, though for our profiling purposes they are only required to serve as reliable biomarkers of cellular responses to compounds.

The collection of reporter cell lines was built from the A549 non-small cell lung cancer cell line, which is amenable both to transfection assays (high rates of transfection) and imaging studies (cells do not tend to clump and can be more easily identified by computer). The collection was constructed by selecting a pSeg-tagged parent A549 clone expressing the nuclear and cellular reporters. (Stable pSeg-tagged A549 clones were grown for tens of passages without signs of reduced expression.) With this clone, ~600 of triply-labeled A549

reporter clones were generated. The identities of the CD-tagged genes were identified by 3' RACE. From this large collection, we selected 93 reporters that were tagged for distinct proteins, placed in diverse GO-annotated functional pathways (Fig. 2a, right), and had detectable YFP levels by microscopy.

As an initial examination of our library of 93 reporter cell lines, we selected six reporter cell lines that represent diverse functional pathways and display distinct spatial localization patterns (Supplementary Fig. 1). These reporter cell lines were treated for 48 hours with six compounds that targeted pathways related to our reporters. Microscopy images revealed that each reporter cell line displayed diverse responses (cell morphology and protein localization) from drug to drug (Supplementary Fig. 1). Previous studies with fixed-cell assays demonstrated that quantitative measurements of cellular responses can be used to predict targeted cellular pathways for compounds across diverse drug classes[19,39]. Could our live-cell reporters also be used in a similar fashion? Would compounds from the same (or different) class(es) elicit similar (or dissimilar) responses from our reporter cell lines? Presumably some reporter cell lines will be better than others for making predictions. Could we identify an "optimal" reporter cell line, whose predictions are highly accurate at classifying compounds into multiple drug classes?

Computation of phenotypic profiles

To address these questions, we made use of "phenotypic profiling," an image-informatics approach previously developed for analyzing high-content microscopy screens[19]. Phenotypic profiles effectively transform compounds into vectors whose entries summarize the responses of cells in a population to the perturbation. This transformation occurred in three main steps (Supplementary Fig. 2 and Online Methods). First, images of perturbed cells were transformed into collections of feature distributions: ~200 features of morphology (e.g. shape of the nuclear and cellular domains) and protein expression[39] (e.g. intensity, localization and texture properties of the tagged protein) are measured for each cell (Supplementary table 1). Second, feature distributions for each condition were transformed into numerical scores: for each feature, differences in cumulative distribution functions (CDF) between perturbed and unperturbed conditions were summarized by a Kolmogorov-Smirnov (KS) statistic[19]. Third, scores were transformed into phenotypic profile vectors: for each perturbation, KS scores were concatenated across features to form a phenotypic profile. The resulting phenotypic profile succinctly summarized the effects of a compound, and could be further extended by concatenating profiles from multiple time points, compound concentrations or even responses from multiple reporter cell lines.

We first investigated whether compounds from the same class would produce relatively similar profiles and whether distinct drug classes would result in dissimilar profiles. We treated our six selected reporter cell lines (Supplementary Fig. 1) with a small panel of "test" drugs (31 conditions = 5 compounds × 6 different drug classes + 1 DMSO control; Supplementary table 2) and imaged cellular responses every 12 hours for 48 hours. 100 DMSO profiles were generated from randomly selecting cells in control conditions (Online Methods). Heat map representations of phenotypic profiles, built by concatenating data across our six reporter cell lines, revealed a strong similarity of compounds from similar

drug classes and, likewise, dissimilar profiles for compounds of dissimilar classes (Supplementary Fig. 3). Thus, as with profiles built from fixed-cell assays and antibody readouts, we observed that profiles based on live-cell readouts produced informative signatures of drug classes.

To select a small number of time points for more scalable screening, we visualized our profiles as time-varying curves by projecting our collection of profiles at each time point into three dimensions (Online Methods). The resulting time traces showed the unperturbed (DMSO-treated) cells remaining in a tight "ball" (Supplementary Fig. 4, gray curves). By contrast, time traces for different classes of drugs moved in different directions away from the DMSO "origin", with different members of each class moving in similar directions (Supplementary Fig. 4). Further, time traces were quite similar across replicate experiments (solid lines vs. dash lines). The divergence of these time traces from one another also suggested that time points of 24 and/or 48 hours were sufficient for discriminating among drug classes (Supplementary Fig. 5). Thus, our results suggested that phenotypic profiles from our reporter cell lines can be used to predict drug classes and that only a small number of time points might be needed for screening.

Identification of the ORACL

To economize large-scale screens, we next investigated to what degree a single reporter cell line and time point could be used to accurately discriminate among our different drug classes. This time we treated all 93 reporter cell lines with our panel of "test" drugs, imaged cellular responses, and then computed phenotypic profiles for each of the reporter cell lines individually using only the final 48 hour time point after drug treatment. As we used only a single time, each drug profile was a single point in our high-dimensional phenotype space. To assess prediction accuracy, we used a cross-validation approach in which we randomly removed six "test" drug profiles (one from each of the six classes), computed the centroid of the remaining four drug profiles in each class, and assigned each of the six test profiles to its nearest centroid (Online Methods). Prediction accuracy was determined by repeating this process 100 times and averaging the results across two duplicates of the experiment.

We found that prediction accuracy varied dramatically from reporter to reporter ("random" guesses from 1 DMSO + 6 drug classes is expected to be ~14% accurate; Fig. 2b, left). As might be expected, the "untagged" (no CD tag) reporter cell line, labeled only with cellular region markers, displayed the lowest prediction accuracy (~60%). (We note this accuracy is already more than four times better than random guessing, which confirmed recent results that morphology carries considerable information for predicting drug classes[25].) Nevertheless, our results also confirmed the intuition that additional information from tagged proteins would improve prediction accuracy. The top reporter cell line—tagged with XRCC5, a nuclear-localized protein that functions in DNA double strand break repair—displayed a high prediction accuracy (94%). This cell line, when compared to others, exhibited more similar phenotypic responses for drugs within the same class (Fig. 2b middle and right). An interesting question, outside the scope of this current study, is determining why certain reporter cell lines are more informative than others. We referred to this best

reporter cell line as an Optimal Reporter cell line for Annotation of Compounds Libraries, or "ORACL."

Identification of multi-classes hits with ORACL

We next used our ORACL to perform a large-scale phenotypic screen of small-molecule compound libraries. These libraries included: the NCI approved oncology drug set IV (101 compounds), the NCI diversity set IV (1596 compounds), the NCI natural product set III (117 compounds), the Prestwick FDA approved drug set (1100 compounds), and the UTSW 8K set (8,000 compounds). All compound libraries were assayed at three different concentrations, except the Prestwick and UTSW sets, which were assayed at a single concentration due to their large sizes. Finally, we included our "reference" drug set. Additionally, to test the ability of the ORACL to identify "novel" drug classes that were not used in its selection, we added a small number of drugs in four new drug classes, for a total of 10 drug classes affecting diverse biological processes (Supplementary table 2; Online Methods). All 38 reference drugs were used at eight, 5-fold dilutions. Finally, to increase our chances of identifying compound effects, given the limited number of compound concentrations selected, cells were imaged at both 24 and 48 hours (Supplementary Fig. 6). In total, profiles were built from ~62,000 3-channel images of ~20,000 conditions (derived from 10,914 compounds and 38 reference drugs at different concentrations as well as control conditions), ~60,000,000 identified cellular regions, and ~230 features per cell, yielding a total of $\sim 1.4 \times 10^{10}$ data points. Our final compound profiles were built by merging data across the 24 and 48 hr time points. As before, these compound profiles can be viewed as points in a high-dimensional feature space.

We took a multi-step strategy to identify hits (Online Methods). First, we transformed feature space and reduced dimensionality to maximally separate our reference drug classes from one another. Linear discriminant analysis (LDA)[40] was applied to our collection of reference drug profiles to identify an "optimal" transform that increased separation of profiles across drug classes while decreasing separation of profiles within each class. Second, we assigned our unknown compounds to the nearest reference drug class. A nearest-neighbor approach was applied to the LDA-transformed space to assign each unknown compound to the class of its nearest reference drug. Third, we calculated confidence scores, ranging from 0 (low) to 1 (high), for each prediction. Scores were estimated based on the collection of intra- and inter-drug class distances among our reference drug profiles. Compounds were re-annotated as "unclassified" if their predictions had low confidence scores (a threshold score of 0.1 was heuristically chosen based on calibration with the NCI approved oncology drug of known mechanism; Supplementary Fig. 7, Online Methods and Supplementary table 3). Finally, we identified "hit" compounds. Hits were defined as compounds not annotated by the control class "DMSO"; that is, hits are "bioactive", but may not necessarily be near to known drug classes. Taken together, our approach allowed us to predict which compounds have activity different from DMSO, predict whether they belong to known or novel drug classes, and prioritize compounds for validation based on confidence scores.

Using our strategy, we identified 429 “primary hit” compounds from our diverse compound libraries (Fig. 3; Supplementary Fig. 8). (We note that there is always a tradeoff between favoring precision *vs.* recall for identifying hits in screens. Though favoring recall has the potential advantage of missing fewer candidate compounds, for this study we chose to favor precision to demonstrate the ability of our ORCACL to identify high quality—rather than high numbers of—hits in different drug classes.) To filter out primary hits that might have induced weak phenotypes, these 429 primary hits were rescreened (secondary screen) at only the highest five concentrations of each reference drugs. After this step, 175 “high-confidence” secondary hits remained (Fig. 3b, middle pie chart; Online Methods), which comprised: 69 unclassified compounds; and 106 compounds classified across 6 of our 10 reference drug classes (49 DNA inhibitors, 45 MT (microtubule) inhibitors, 5 mTOR inhibitors, 4 proteasome inhibitors, 2 HDAC inhibitors, 1 Hsp90 inhibitor; Fig. 3b, bottom pie chart).

Validation of identified hits from the screening

We next investigated the accuracy of our secondary hits. We began with our two smallest predicted classes: Hsp90 and HDACs. Gratifyingly, identified hits in these two classes both had literature support. In the Hsp90 class, the compound NSC330500 (macbecin II) was shown previously to be an Hsp90 inhibitor[41]. In the HDAC class, compounds NSC701852 (from the NCI oncology drug set) and Vorinostat (from the Prestwick library) were different names for the same compound (SAHA), a known HDAC inhibitor[42].

To test prediction accuracy in our remaining four drug classes, we performed experimental validation (Fig. 4). We used all of our reference drugs and 175 secondary hits so that we could both calibrate readout thresholds using reference drugs and estimate false discovery rate of our predictions (Online Methods). Specifically, for each validation, we chose a readout threshold so that 90% of the reference drugs above the threshold belong to the class to be validated (FDR_{Ref} = 0.1, Fig. 4 top). Then we calculated the false discovery rate of our secondary hit predictions (FDR_{hits}) as the percentage of predicted compounds that failed to pass the readout threshold.

To test DNA damaging activity, we carried out immunofluorescence (IF) staining experiments to detect the level of phospho-H2AX, whose level increases rapidly in response to DNA damaging agents. The DNA damaging ability of each reference drug/compound was measured by the median of averaged phospho-H2AX intensities in nucleus regions. In Fig. 4, at the threshold of 60.74 (FDR_{Ref} = 0.1), 76% (37/49) of our predicted DNA compounds (Fig. 4, blue dots) passed the validation (24% failed, FDR_{hits} = 0.24).

To test microtubule perturbing activity, we carried out live cell imaging with TUBA1C CD-tagged reporter cell line to examine mitotic arrest. We calculated mitotic index (the percentage of cells undergoing mitosis) based on cell morphology and tubulin intensity (Online Methods). At the threshold of 0.06 (FDR_{Ref} = 0.1), 96% (43/45) of our predicted MT compounds (Fig. 4, yellow dots) passed the validation (4% failed, FDR_{hits} = 0.04).

To test proteasome inhibiting activity, we used an ubiquitin-fused R-GFP clone of HeLa cells[43]. Under normal conditions, R-GFP would be degraded through the

ubiquitinproteasome system by the N-end rule pathway, and show no fluorescent intensity. In contrast, proteasome inhibitors cause an increase in R-GFP signal (Fig. 4, drug PS-341 at bottom). At the threshold of 166 (FDR_{Ref} = 0.1), all our 4 predicted proteasome hits were validated (Fig. 4, red dots). Among those, two (Carfilzomib and Bortezomib) were identified previously in the literature[44,45]. The other two (NSC26113 and NSC33570) were not previously known as proteasome inhibitors and are new predictions.

Finally, to test mTOR inhibiting ability, we chose to measure the level of the phosphorylated ribosomal protein S6 by immunostaining. Under normal condition, mTOR constitutively phosphorylates the ribosomal protein kinase (S6K), which in turn phosphorylates S6. Therefore, mTOR inhibitors will decrease the level of phosphorylated S6. However, we found that our Hsp90 reference drugs also decreased the phosphorylation level of S6 (Fig. 4), which was consistent with previous literature[46]. Given the fact that both mTOR and Hsp90 inhibitors reduce the level of phosphorylated S6, we combined these two classes when calculating FDR_{Ref} for reference drugs. At the threshold of −4.7 (FDR_{Ref} = 0.1), our predicted Hsp90 inhibitor (macbecin II; Fig. 4, magenta dots) and three predicted mTOR inhibitors (Fig. 4, green dots) passed the threshold. Two of the three validated mTOR compounds were previously known in the literature as rapamycin and everolimus; the other one (NSC176324) is a new prediction.

Taken together, these results suggest that our predictions of drug classes, derived from a single-pass phenotypic screen using our ORACL, had high accuracy across diverse functional classes.

Identification of compounds in novel drug classes

We also tested whether our approach could also group compounds in our screen belonging to drug classes that were not included in our reference drugs. We used both an unsupervised and a supervised approach (Online Methods). We first used unsupervised hierarchical clustering to group the secondary screening profiles of all 429 compounds that showed bioactivity (away from DMSO) in our primary phenotypic screen (Fig. 5). We found 23 significant clusters (p-value < 0.01, permutation test). As expected, six clusters reflected our initial training classes used to identify our ORACL. Two new clusters were identified because they each grouped together a small number of drugs we had added to our reference set for the primary screen (but were not used for our selection of the ORACL): ER (2 reference drugs and 1 “unknown”) and Aurora Kinase inhibitors (1 reference drug and 2 “unknown”). Notably, three of the remaining 15 novel clusters were identified (by literature searches) as: (1) glucocorticoid steroids (26 members, including Betamethasone, Flunisolide, and Halcinonide); (2) Na^+/K^+ ATPase inhibitors (5 members, including Digoxin, Lanatoside C, and Proscillaridin); and (3) dihydrofolate reductase inhibitors (3 members, consisting of NSC740 (Methotrexate, NCI oncology drug set), NSC382035 (methylbenzoprim, NCI diversity set) and Amethopterin (Methotrexate, Prestwick library)).

We next used a supervised approach to perform a “virtual screen” of the Prestwick and UTSW 8K compound sets (Online Methods). We re-trained new classifiers using the original drug set together with two new drug classes (using 15 glucocorticoid compounds and 5 Na^+/K^+ ATPase inhibitors; Supplementary table 5) that were identified from the

previous cluster analysis (dihydrofolate reductase inhibitors were excluded due to their small numbers). Cross-validation showed that our ORACL recalled: 15/15 “left-out” literature-supported glucocorticoid compounds, and 3/5 Na^+/K^+ ATPase inhibitors (predictions for 2/5 fell below our confidence threshold of 0.1). For our virtual screen, we identified a total of 17 new compounds that were classified into the glucocorticoid class (6 were supported by literature search, 11 had no annotation; Supplementary table 6). These results suggested that our ORACL has the potential to discover compounds in drug classes other than those used for its selection.

Discussion

We address a key challenge of designing phenotypic screens, namely how to select ‘optimal’ biomarkers. Biomarkers are typically chosen by experts based on prior knowledge or availability of reagents. Here, we took a different approach. Rather than hand specifying biomarkers that are specific to a single target or pathway, we developed an objective procedure for selecting a maximally informative reporter cell line whose phenotypic profiles are optimized for distinguishing the effects of multiple cancer drug classes (Fig. 1). We made use of a diverse collection of reporter cell lines, and let analytical criteria determine which among them—the ORACL—could best classify compounds into multiple, diverse drug classes (Fig. 2). We then demonstrated that the ORACL's information-rich, high-dimensional phenotypic profiles could be used in a single-pass screen to predict compound leads accurately across multiple drug classes (Figs. 3-5).

We initially used our ORACL to identify compounds whose activity profiles are similar, but not identical, to current drugs. These new compounds may have similar or distinct chemical structures and operate through different mechanisms (Supplementary table 4). Thus, such an approach provides an opportunity to expand and improve upon classes of drugs that are already known to be effective. Guided by our choice to focus on high-quality (rather high numbers of) hits, we identified 175 compound leads. Our ORACL classified 106 of these into our six reference drug classes. In addition to identifying compounds with DNA(49) and microtubule(45) activity—classes considered “low-hanging fruit”[47] for drug discovery—we identified mTOR(5), proteasome(4), HDAC(2) and Hsp90(1) inhibitors. Many of these predictions were subsequently validated through literature or experiment (90/106; Fig. 4). It is possible that we underestimated the true accuracy of our method. Indeed, our drug classes were defined fairly broadly, and compounds might have activities within a class that would not be detected by our choice of validation assay. For example, to validate DNA compounds we used phospho-H2AX, which reports on DNA double-strand breaks and could miss compounds that affect DNA through alternative mechanisms, such as inhibition of DNA synthesis. Nevertheless, our high overall validation rate of 85% demonstrates that our approach can produce high quality compound “leads” across diverse drug classes.

There is no guarantee that an ORACL can identify compounds in classes other than ones for which it was trained. However, for our particular ORACL and compound libraries we were able to identify clusters of “unclassified” compounds belonging to classes that we did not originally include in the selection of the ORACL, namely Glucocorticoids (26), Na^+/K^+ ATPase inhibitors (5) and DHFR inhibitors (3). We used these grouping to bootstrap *in silico*

predictions of other unknown compound into these classes. This case study suggested that ORACLs have the potential to identify compounds targeting different cellular pathways that were not included in the original experimental design.

It is not to be expected that a small number of reporter cells lines could be used in a large-scale screen to classify compounds accurately across diverse drug classes. In fact, it is a notable finding that a single reporter cell line could be identified to do so. Our results suggest that the strategy of objectively identifying the 'right' reporter cell lines may be as important as (or at least complement) strategies of increasing the numbers of multiplexed biomarkers[36,37].

Our selection of an ORACL was based on an analytical search procedure, which depended on two key sets of inputs. The first set of inputs is the drug training set. How "smart" a reporter cell line depends on user-provided definitions of drug classes and choices of specific training drugs in each class (as well as experimental parameters such as chosen doses, treatment times, etc.). If a provided training drug class lumps together multiple mechanisms of action (e.g. microtubule stabilizers or destabilizers), then the smart reporter cell lines would be expected to ignore these differences; conversely, if a drug class is broken into more refined subclasses, then the smart reporter cell lines would be expected to distinguish among them. The second set of inputs is the library of reporter cell lines. The more diverse the reporter cell line library, the more likely it is that an ORACL can be identified whose classification accuracy is high enough to be useful for screening. Of course, the success of any reporter cell line (in our or any study) is due to some ineffable combination of cell type, clone, biomarker, cell features (e.g. biomarker expression and cell morphology), and so on. Although a reporter cell line tagged for the biomarker XRCC5 was identified as an ORACL in our study, it is reasonable to expect that for a different study a different reporter cell line, tagged with a different gene and optimized for a different set of perturbations, might be selected as the ORACL. Our study raises the question as to which pathways and readouts are most informative for distinguishing different drug classes, and whether computational approaches, including active learning[48], could be adapted to predict ORACLs *a priori*.

The general procedure we describe for finding ORACLs can be incorporated into the design of any phenotypic screen. Desired collections of discovery drug classes can be matched objectively to ORACLs, which will increase the efficiency, scale and accuracy of future phenotypic screens. The identification of ORACLs provides means to classify very large compound libraries across diverse drug classes. As large chemical libraries become increasingly available, methods that efficiently screen for promising compound leads across multiple drug classes might substantially expand our drug repertoire for diseases such as cancer.

Online Methods

Experimental assays

Generation of reporter library—We constructed our live-cell reporter library using the adenocarcinoma cell line A549. As described below, we did this in two steps: 1) we

generated a parent clone containing distinct fluorescent labels for cellular regions; 2) we performed random genomic tagging on this parent clone to build our reporter library.

For visual demarcation of individual cellular and nuclear regions, we constructed a “pSeg” plasmid (short for: “plasmid for image Segmentation”), using a combination of standard molecular cloning techniques and Gibson assembly (New England Biolabs, Inc. #E2611), in which a pMYs retrovirus expression vector (Cell Biolabs, Inc. #RTV022 and #RTV023) was modified to express *Drosophila* histone H2B-fused CFP (Cyan Fluorescent Protein) in the nucleus and mCherry protein in the whole cell. pSeg plasmids were then transfected into the HEK293T-based Platinum-A retroviral packaging cell line (Cell Biolabs, Inc. #RV-102) using Lipofectamine 2000 (Life Technologies #1168027) according to manufacturer protocol. Retrovirus-containing supernatant was then added to A549 cells for genomic integration. After two days of integration at 32°C, medium was replaced and cells were incubated at 37°C for 24 hrs to increase viability. CFP- and mCherry-positive cells were then sorted onto 384-well plate by FACS, and their fluorescence localization was validated by Nikon TE-2000 E2 epifluorescence microscope (Nikon, Inc.). One of the pSeg clones was selected for subsequent generation of our reporter library based on its relative fluorescence intensity and stability compared to other clones.

For random genomic labeling, CD tagging[49] was carried out as in our pSeg labeling step except using a CD tag plasmid (kind gift from Uri Alon). Identification of tagged genes by 3’ RACE was performed as in Sigal et al. (2007).

Cell culture and drug screening assays—A549 adenocarcinoma cells were cultured in RPMI1640 media containing 10% fetal bovine serum, 2mM glutamine, 50 units/ml penicillin, and 50 μg/ml streptomycin (all from Life Technology, Inc.), at 37°C, 5% CO2, and 100% humidity. For screens, cells were grown in 10 cm culture plates overnight, detached by trypsin, counted by TC10 automated cell counter (Bio-Rad Laboratories, Inc.) and seeded onto 384-well plate at a density of 3000 cells/well by Matrix electronic multichannel pipette (Matrix Technologies Co.). After 24 hrs at 37°C, drugs were added by Beckman Coulter BioMek FX liquid handlers (Beckman Coulter, Inc.), and the plate was covered by Breath-Easy sealing membrane (Sigma-Adrich, Inc.) and incubated at 37°C for two days.

Drug libraries—The drug libraries were screened in two batches. Batch 1: the approved oncology drug set IV (101 compounds), the diversity set IV (1596 compounds), and the natural product set III (117 compounds) (all acquired from NCI). Batch 2: the Prestwick library (1100 compounds) (purchased from Prestwick Chemical), and the University of Texas Southwestern 8K diversity subset (~8,000 compounds approximating the chemical diversity of the 230K institutional library that was purchased from ChemDiv, ChemBridge, Comgenix, and TimTek). In Batch 1, compounds were screened at 3 concentrations (10, 1, and 0.1 μM); in Batch 2, compounds were screened at a single concentration (2.5 μM).

Screens—All experiments below were performed in 384-well plates.

93-reporter dataset—93 different reporters were treated with 30 different drugs spanning 6 drug classes. Plate columns: 23 reporters + 1 control reporter (EIF4A1) were assayed. Plate rows: 15 drugs + DMSO control (row H) (drug positions across drug classes were randomized across plates). Two replicates were performed on different plates. In total, 16 plates were used to generate this dataset. (EIF4A1 was included on every plate so that in subsequent analytical steps we could confirm that there were no significant quality control issues, such as plate-to-plate differences or pipetting artifacts.)

Screening dataset—The XRCC5 reporter was used to screen 10,914 chemical compounds. For all plates, the 2nd and 23th columns were treated with only DMSO control and the 1st and 24th columns were treated with positive control drugs Gemcitabine or PS-341. The rest of each plate (Row A to P, Column 3 to 22) was used for compound screening. There were two types of plates: reference plates, which contained 38 known drugs at 8 serial 5-fold concentrations used in analytical steps to predict drug classes of unknown chemical compounds; and compound plates, which contained unknown chemical compounds. Our screening was performed in two batches. To avoid normalization issues, we included reference plates in each batch so that profiles and hits could be computed and identified independently. The first batch contained 2 reference plates and 18 compound plates; the second batch contained 5 reference plates and 29 compound plates. The two batches were imaged and analyzed independently.

Validation assays—Secondary assays were performed to validate predictions in drug classes: (a) DNA, (b) mTOR, (c) mitosis and (d) proteasome.

(a), (b). Immunofluorescence experiments for validation of putative drug mechanisms were carried out in 384-well plates. After 24 hrs of drug treatment, cells were fixed in 4% PFA in PBS and permeabilized by TBS (20mM Tris, pH 7.4, and 0.9% NaCl) containing 0.2% Triton X-100. Primary antibodies were incubated overnight at 4°C in TBS containing 0.1% Triton X-100 and 3% BSA. For validation of DNA inhibitors, phospho-H2A.X antibody (Cat. # 9718, Cell Signaling Tech.) was used in 1:400 dilution. For validation of mTOR inhibitors, phospho-S6 antibody (Cat. #4858, Cell Signaling Tech.) was used in 1:100 dilution. FITC conjugated secondary antibody (Molecular Probes) was incubated for 2 hrs at RT together with rhodamin conjugated Phalloidin (Molecular Probes) at appropriate dilutions. At the final washing step, Hoechst 33341 was added to stain DNA.

(c), (d). For validation of mitosis inhibitors, live-cell imaging of TUBA1C-CD tag clone of A549 was carried out in a 384-well plate (Sanger sequencing confirmed that TUBA1C was tagged with YFP); images were taken at 3 hrs, 24 hrs, and 48 hrs post drug treatment. For validation of proteasome inhibitors, live-cell imaging of Ub-R clone of HeLa cells[43] (kind gift from Dr. DeMartino) was carried out in 384-well plate; after 24 hrs of drug treatment, the medium was exchanged with PBS containing Hoechst 33341 and images were taken.

Image acquisition—Images were acquired by using a Nikon TE-2000 E2 epifluorescence microscope equipped with integrated Perfect-Focus (PFS), Nikon Plan Apochromat 10x objective lens, and CoolSNAP HQ camera (Photometrics) (93-reporter dataset) and Zyla 5.5 sCMOS camera (Andor Technology) (Screening dataset), using 2×2 camera binning. All

image acquisition was controlled by Metamorph software (Universal Imaging). One image was acquired for each well. Images with obvious anomalies (e.g. out of focus, abnormal fluorescent patterns caused by dust, scratches on the plate, or fluorescent compounds) were discarded by manual inspection.

Computational analysis

Image analysis and feature extraction (all datasets)—In brief, image background subtraction was performed using National Institutes of Health ImageJ Rolling Ball Background Subtraction algorithm[50]. Cells were automatically identified using our in-house watershed-based algorithm[39], which retrieves nuclear region using nuclear marker then combines cytoplasmic marker to identify cell boundary. 234 features were measured for each identified cell[39]: 30 intensity features, 92 object morphology features, 5 object moment features, 49 Zernike moment features, 26 Haralick texture features, and 32 cell morphology features. Most treatment conditions did not induce cell death, and typically more than 1000 cell objects in each condition were captured and used to build phenotypic profiles. (Interestingly, even for the few compounds in which a fraction of dying or dead cell objects were captured —typically more than 500—these objects still provided useful information for classification.)

Phenotypic profiles—The phenotypic profile of a treatment (compound) was based on the Kolmogorov-Smirnov (KS) statistic[19]. Let $F_{trt,i}$ be the cumulative distribution function (cdf) of feature i and treatment trt, and $F_{ctrl,i}$ be the cdf of feature i of control (DMSO). $KS_{trt,i}$ computes $F_{ctrl,i} - F_{trt,i}$ at the point where $|F_{ctrl,i} - F_{trt,i}|$ attains its maximum. A positive KS roughly indicates a shift of increased feature values compared to DMSO control. Our phenotypic profiles were then simply vectors of all k KS values: $P_{trt} = (KS_{trt,i}, ..., KS_{trt,k})$. Here, the phenotypic profile of a treatment at a given time point (e.g. 48 hours) using one reporter cell line and all features had length 234. We note that all phenotypic profiles were built on a plate-by-pate and time point-by-time point basis, i.e. responses from treated wells were only compared to DMSO controls on the same plate at the same time point. We also note that phenotypic profiles can easily be expanded by concatenating profiles from different time points (or from different reporter cell lines).

For the 93-reporter dataset (used to select an ORACL), phenotypic profiles were built with features at 48 hours. For each 384-well plate in the 93-clone dataset, the whole DMSO well of a reporter was used to build $F_{ctrl,i}$'s when calculating phenotypic profiles of compound treatments. To calculate phenotypic profiles of DMSO, the cell population of the DMSO well was randomly split into two subpopulations, one for building $F_{ctrl,i}$'s and the other was considered as "mock compound treatment" to build $F_{trt,i}$'s. We repeated this procedure 100 times to obtain 100 DMSO phenotypic profiles for each CD-tagged reporter.

For the screening dataset, phenotypic profiles were built by concatenating profiles at 24 and 48 hours (combining these two time points gives, on average, higher prediction accuracy; Supplementary Fig. 6). These resulting compound profiles had length 468 (= 234 + 234). For each 384-well plate in the screening dataset, cells in DMSO wells at I2, J2, K2, L2, M2, N2, O2, P2, A23, B23, C23, D23, E23, F23, G23, H23 were pooled together to build $F_{ctrl,i}$'s.

Other DMSO wells were considered as "mock compound treatments" to calculate phenotypic profiles for DMSO.

Data visualization (all datasets)—The profile for each compound can be considered as a point embedded in a D-dimensional space (the dimension D is the same as the length of the profile). To visualize the similarity and dissimilarity between compounds, multidimensional scaling (if the number of compounds < 300) or principle component analysis (if the number of compounds >= 300) was performed using Matlab (R2015a) functions *mdscale* or *pca*, respectively.

Prediction accuracy of each reporter (93-reporter dataset)—To choose a "smart" reporter, we used prediction accuracy to assess the ability of reporters to discriminate drugs from different drug classes. For each reporter, 6 drugs (1 for each drug class) were randomly chosen to be testing data. The remaining data (24 drugs + 100 DMSO) were used to calculate the centroids of each drug class (including DMSO). Testing data were assigned to the drug class of the nearest centroid. Prediction accuracy was defined as the fraction of correct assignments. The procedure was repeated 100 times to obtain a population of prediction accuracy for each reporter. Mean and standard deviation were reported.

Feature selection (screening dataset)—Our next step was designed to discard "unreliable" features whose KS scores varied too much from one reference plate to another. We applied linear regression analysis using the reference drugs for each of the two batches of screened libraries independently. If a feature is reproducible, we reasoned that its KS scores from one reference plate should be predictive of the KS scores on another reference plate. For each feature, a linear regression model was fit between data using all 38 reference drugs at the highest four concentrations from two reference plates. For batch 1 we had two replicates of our reference drugs, and features with a coefficient of determination (R^2) < 0.8 were discarded; for batch 2 we had five replicates of our references drugs, all 10 pairwise coefficients of determination were computed and features that had (R^2) < 0.8 for eight or more comparisons were discarded. After feature selection, 8 or 16 were dropped (in batch 1 or 2), leaving 460 or 452 features (respectively) to build our profiles.

Identify effective concentrations of reference drugs (screening dataset)—For the reference plates, there were 8 serial 5-fold dilutions for each drug. Of course, not every concentration had effects on cells and not every drug had the same effective concentration. Therefore, for each drug we identified a concentration above which cells exhibited responses different from DMSO-treated cells. Our strategy was to estimate the range of cellular responses observed in the DMSO treatments, then to search for the lowest drug concentrations that caused phenotypic changes that greatly exceeded this control range.

More specifically, let ($KS_{trt,1}$, ..., $KS_{trt,k}$) be the phenotypic profile of a treatment condition (denoted *trt*) given the collection of KS scores for its *k* features. The strength of cellular responses, R_{trt}, induced by this treatment were quantified by $R_{trt}=\sum_{i=1}^{k}((KS_{trt,i}-\mu_i)/\sigma_i)^2$, where μ_i and σ_i are the average and standard deviation (respectively) of KS values of the DMSO controls for feature *i*. (The total number of DMSO

treatments was 665 or 1166 for batch 1 or 2). To evaluate the significance of observed cellular responses, we estimated the distribution of control cellular responses to DMSO, $\{R_{ctr}\}$, (calculating using each DMSO as a treatment and computing μ_i and σ_i using all other DMSO treatments). Then, for each drug, we chose the effective concentration to be the lowest dose that satisfied the stringent criteria $R_{trt} \geq 2R^*$, where R^* is the 99% quantile of the collection of control response values $\{R_{ctr}\}$. For each drug, only the concentrations that were equal or higher than its effective concentration were used to predict drug classes of unknown compounds.

Predict drug classes of unknown compounds by linear discriminant analysis (LDA) and nearest neighbor classification (screening dataset)—We applied LDA[40,51] with regularization (shrinkage)[52] to find a projection of feature space that places profiles of reference drugs with the same drug class close to each other but far from profiles of drugs with different drug classes. After the projection was learned using our reference drug profiles, all data (reference drugs and compounds) were projected into the subspace.

The distance between a compound and a drug class was then defined to be the shortest distance to any reference drug that belongs to the drug class. Compounds were predicted to belong to the nearest drug class. (This approach is equivalent to prediction based on the drug class of a compound's nearest neighbor in our reference set.) 10-fold cross-validation suggested that the accuracy of this prediction procedure was ~90%.

Confidence (screening dataset)—We associated each prediction with a confidence value. Intuitively, the greater the distance between a compound and its predicted drug class, the less confident we were of the prediction being correct. Specifically, we defined $Confidence = \Pr(R_{c,m} = 1 \mid D_{c,m})$, where $R_{c,m} = 1$ if compound c belonged to drug class m or 0 otherwise, and $D_{c,m}$ is the distance between c and m. By Bayes's theorem,

$$Pr\,(R_{c,m}=1|D_{c,m}) = \frac{Pr\,(D_{c,m}|R_{c,m}=1) \times Pr\,(R_{c,m}=1)}{Pr\,(D_{c,m}|R_{c,m}=1) \times Pr\,(R_{c,m}=1) + Pr\,(D_{c,m}|R_{c,m}=0) \times Pr\,(R_{c,m}=0)}$$

We estimated each term on the right hand side using our collection of reference drugs, where $\Pr(D_{c,m} \mid R_{c,m} = 1)$ and $\Pr(D_{c,m} \mid R_{c,m} = 0)$ were estimated by fitting an exponential distribution and Gaussian kernel smoothing, respectively (matlab function: *fitdist*). Confidence ranged from 0 to 1, with 1 being the most confident that a prediction was correct.

Unclassified drug class, primary hits, and secondary hits (screening dataset)—Compounds were predicted to be in one of our reference drug classes, which included DMSO. If the associated confidence value was lower than 0.1, we defined the predicted drug class of the compound to be “Unclassified”. We defined primary hits to be compounds that were not predicted as DMSO in the primary screen. Secondary hits were compounds that had the same predicted drug class in the secondary screen as the primary screen for at least one of the concentrations. In total, 175 compounds were identified as secondary hits.

Analysis of narrowing primary to secondary hits—From our 429 primary hits, 175 (41%) had the same predictions in both the primary and secondary screens. The remaining compounds can be separated into three categories: (i) 17 (4%) had bad images; (ii) 78 (18%) showed bioactivity but had inconsistent predictions from the primary screens; and (iii) 159 (37%) became inactive. A main reason why some compounds may have appeared inactive in the secondary screen is that we intentionally removed the lowest three concentrations of reference drugs used in the primary screen to eliminate compounds that induced weak phenotypes close to DMSO. It is also possible that we encountered typical issues such as compounds becoming inactive due to multiple freeze/thaw cycles, etc.

Benchmark of our prediction using NCI oncology drug set—The NCI oncology library contains 101 well characterized drugs. We analyzed the results of our primary screen for this library and offer the following rationale for our hit/non-hit calls.

1) Quality Control. 3/101 drugs had bad images (out-of-focus or artifact fluorescent patterns), leaving 98 drugs for subsequent analysis.

2) Hits. 50/98 drugs were identified as hits (bioactive compounds): 42 with classified predictions (83% accuracy) and 8 as “unclassified” (the majority of which belong to classes that were not included in our reference drugs, such as RTK inhibitors, ALK inhibitors, and histamine N-methyltransferase inhibitors).

3) Non-hits. 48/98 were identified as non-hits. We identified three main reasons for these being non-hits. (i) Resistance. Our reporter cell line (derived from A549) has been reported to be resistant to several non-hit drugs in this library, including Cisplatin[53], Gefitinib[54] and Erlotinib[55]. (ii) Biological relevance. Some non-hit drugs may act through mechanisms that are irrelevant to the biology of our particular non-small cell lung cancer-derived reporter cell line, such as: Imiquimod, used as an immune response modifier; Zoledronic acid, used to slow down bone resorption; and Exemestane, used to inhibit aromatase, which synthesizes estrogen. (iii) Dose. An issue for all large-scale drug screens is dose. Our screen was no exception, and in several cases chosen concentrations were insufficient to elicit a response that differed from control (DMSO). For example, Oxaliplatin was screened in the NCI oncology library at 10, 1 and .1 uM, which is below the minimal concentration (25 uM) we found was needed to induce a strong non-DMSO response in our reference drug set, and thus was deemed a non-hit. A similar situation occurred for Pemetrexed.

Readouts of validation experiments—For DNA validation experiments, the effect of a compound was measured by the median of the average nucleus phosphorylated H2AX intensity of all cells in the well. For mTOR validation experiments, the effect of a compound was measured by the negative log of the median of the average cytoplamic pS6 intensity of

all cells in the well. For proteasome validation experiments, the effect of a compound was measured by the mean of the average nucleus GFP intensity of all cells in the well. For MT validation experiments, a mitotic index (the proportion of cells undergoing mitosis) was calculated for each compound. A mitotic cell was required to meet each of the following (empirically determined) criteria: Cell solidity > 0.95 (convex shape), Cell eccentricity < 0.6 (close to a circle), Nucleus solidity > 0.95, Nucleus eccentricity < 0.6, Median of cytoplasmic mCherry intensity > 800 (living cells), and Total TUBA1C intensity > 75000.

False discovery rate (FDR) in validation experiments—We adopted a two-step strategy to evaluate the quality of our predictions. First, we used the "ground truth" reference drugs to calibrate readout thresholds for each drug class to be validated (DNA, MT, Proteasome, or mTOR). In particular, we computed FDR_{Ref} as a function of readout threshold for each validation experiment. Here, FDR_{Ref} was defined to be the percentage of "ground truth" reference drugs that were not in this drug class out of all reference drugs that passed the chosen threshold. In Fig. 4, all thresholds were chosen to give a 0.1 FDR_{Ref}. Second, we evaluated the quality of our prediction procedure on our secondary hits based on chosen readout thresholds. Define: True Negatives (Positives) to be compounds that were not (were) predicted to be in a drug class that fell below (above) our validation threshold; and False Negatives (Positives) to be compounds that that were not (were) predicted to be in an drug class but fell above (below) our validation threshold. (We abbreviate these as TN, TP, FN, and FP as per usual convention.) We then defined $FDR = FP/(TP + FP)$.

As an example, 49 DNA compounds were originally predicted. We then tested all 175 secondary hits for disruption of DNA activity. Based on calibration with our reference dataset, a threshold of 60.74 for phosphorylated H2AX staining gave 0.1 FDR_{Ref}. Based on that threshold, 12 of 49 predicted DNA compounds were below the threshold (FP = 12, TP = 37); we therefore computed an FDR of 0.24 (12/49).

Hierarchical clustering and cluster assignment—To identify novel drug classes, we performed the average linkage hierarchical clustering. We made use of the pair-wise distances between our reference drugs to decide the threshold for cluster assignment. We first separated the pair-wise distances between our reference drugs into two groups: within-class distance and between-class distance. Then we fitted a logistic regression model and chose the threshold to be the distance at which the probability of being within-class distance equals to the probability of being between-class distance. This threshold was used as the cutoff for the hierarchical clustering tree to determine clusters. A standard permutation test (randomization of cluster labels implemented in Matlab 2015a) was used to evaluate the significance of clusters

Supervised approach to assess the ability of our ORACL to identify novel drug classes—From our hierarchical clustering of the 429 hit compounds, 15/26 compounds in the glucocorticoid class and 5/5 compounds in Na+/K+ ATPase inhibitors had literature support for their function. These compounds were then added as new reference drugs (Supplementary table 5) to the original set of 10 reference drug classes. We used a 5-fold cross-validation strategy to evaluate both the recall and the ability to identify new compounds within the Prestwick and UTSW 8K libraries. In each iteration, we randomly

selected a subset of the new reference drugs (12/15 literature supported compounds in the glucocorticoid class, and 4/5 in the Na+/K+ ATPase inhibitor class) and re-preformed the whole prediction process (which resulted in new combinations of cell phenotype features for each in silico screen). To evaluate recall, we evaluated our ability to re-identify compounds "left-out" in each iteration of cross validation (3 for glucocorticoid class and 1 for Na+/K+ ATPase inhibitors; Supplementary table 6). To evaluate the ability to identify new compound candidates, we pulled out all compounds that were classified into these two new drug classes (Supplementary table 6; blue text: compounds contained in our original 429 hits; yellow highlight: literature-supported predictions; all others: novel hits from the in silico screen not contained in our original 429 hits).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank: members of the Altschuler and Wu lab for critical feedback; Uri Alon and members of his lab for providing the CD tag plasmid and guidance on its use; George DeMartino for useful conversations and reagents for the proteasome validation; and Shuguan Wei for help with HTS experiments. This research was partially supported by the National Institute of Health grants CA133253 (S.J.A.), R01CA184984 (L.F.W.), and the Institute of Computational Health Sciences (ICHS) at UCSF (SA, LW).

References

1. van 't Veer LJ, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature. 2002; 415:530–536. [PubMed: 11823860]
2. Thomas RK, et al. High-throughput oncogene mutation profiling in human cancer. Nature genetics. 2007; 39:347–351. [PubMed: 17293865]
3. Kolch W, Pitt A. Functional proteomics to dissect tyrosine kinase signalling pathways in cancer. Nature reviews. Cancer. 2010; 10:618–629. [PubMed: 20720570]
4. Griffin JL, Shockcor JP. Metabolic profiles of cancer cells. Nature reviews. Cancer. 2004; 4:551–561. [PubMed: 15229480]
5. Zhang J, Yang PL, Gray NS. Targeting cancer with small molecule kinase inhibitors. Nature reviews. Cancer. 2009; 9:28–39. [PubMed: 19104514]
6. Kelloff GJ, Sigman CC. Cancer biomarkers: selecting the right drug for the right patient. Nature reviews. Drug discovery. 2012; 11:201–214. [PubMed: 22322254]
7. Sundberg SA. High-throughput and ultra-high-throughput screening: solution- and cell-based approaches. Current opinion in biotechnology. 2000; 11:47–53. [PubMed: 10679349]
8. Mayr LM, Bojanic D. Novel trends in high-throughput screening. Current opinion in pharmacology. 2009; 9:580–588. [PubMed: 19775937]
9. Koehn FE. High impact technologies for natural products screening. Progress in drug research. Fortschritte der Arzneimittelforschung. Progres des recherches pharmaceutiques. 2008; 65:175, 177–210. [PubMed: 18084916]
10. Lachance H, Wetzel S, Kumar K, Waldmann H. Charting, navigating, and populating natural product chemical space for drug discovery. Journal of medicinal chemistry. 2012; 55:5989–6001. [PubMed: 22537178]
11. Nielsen TE, Schreiber SL. Towards the optimal screening collection: a synthesis strategy. Angewandte Chemie. 2008; 47:48–56. [PubMed: 18080276]
12. CJ OC, Beckmann HS, Spring DR. Diversity-oriented synthesis: producing chemical tools for dissecting biology. Chemical Society reviews. 2012; 41:4444–4456. [PubMed: 22491328]

13. Caie PD, et al. High-content phenotypic profiling of drug response signatures across distinct cancer cells. Molecular cancer therapeutics. 2010; 9:1913–1926. [PubMed: 20530715]
14. Klebe G. Virtual ligand screening: strategies, perspectives and limitations. Drug discovery today. 2006; 11:580–594. [PubMed: 16793526]
15. Schneider G. Virtual screening: an endless staircase? Nature reviews. Drug discovery. 2010; 9:273–276. [PubMed: 20357802]
16. Inglese J, et al. High-throughput screening assays for the identification of chemical probes. Nature chemical biology. 2007; 3:466–479. [PubMed: 17637779]
17. Chen B, et al. Small molecule-mediated disruption of Wnt-dependent signaling in tissue regeneration and cancer. Nature chemical biology. 2009; 5:100–107. [PubMed: 19125156]
18. Wilson CJ, et al. Identification of a small molecule that induces mitotic arrest using a simplified high-content screening assay and data analysis method. Journal of biomolecular screening. 2006; 11:21–28. [PubMed: 16234339]
19. Perlman ZE, et al. Multidimensional drug profiling by automated microscopy. Science. 2004; 306:1194–1198. [PubMed: 15539606]
20. Lamb J, et al. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. Science. 2006; 313:1929–1935. [PubMed: 17008526]
21. Potts MB, et al. Using functional signature ontology (FUSION) to identify mechanisms of action for natural products. Science signaling. 2013; 6:ra90. [PubMed: 24129700]
22. Young DW, et al. Integrating high-content screening and ligand-target prediction to identify mechanism of action. Nature chemical biology. 2008; 4:59–68. [PubMed: 18066055]
23. MacDonald ML, et al. Identifying off-target effects and hidden phenotypes of drugs in human cells. Nature chemical biology. 2006; 2:329–337. [PubMed: 16680159]
24. Houle D, Govindaraju DR, Omholt S. Phenomics: the next challenge. Nature reviews. Genetics. 2010; 11:855–866.
25. Futamura Y, et al. Morphobase, an encyclopedic cell morphology database, and its use for drug target identification. Chemistry & biology. 2012; 19:1620–1630. [PubMed: 23261605]
26. King KR, et al. A high-throughput microfluidic real-time gene expression living cell array. Lab on a chip. 2007; 7:77–85. [PubMed: 17180208]
27. Stegmaier K, et al. Gene expression-based high-throughput screening(GE-HTS) and application to leukemia differentiation. Nat Genet. 2004; 36:257–263. [PubMed: 14770183]
28. Kawatani M, et al. Identification of a small-molecule inhibitor of DNA topoisomerase II by proteomic profiling. Chemistry & biology. 2011; 18:743–751. [PubMed: 21700210]
29. Muroi M, et al. Application of proteomic profiling based on 2D-DIGE for classification of compounds according to the mechanism of action. Chemistry & biology. 2010; 17:460–470. [PubMed: 20534344]
30. Bantscheff M, et al. Quantitative chemical proteomics reveals mechanisms of action of clinical ABL kinase inhibitors. Nature biotechnology. 2007; 25:1035–1044.
31. Feng Y, Mitchison TJ, Bender A, Young DW, Tallarico J.a. Multi-parameter phenotypic profiling: using cellular effects to characterize small-molecule compounds. Nature reviews. Drug discovery. 2009; 8:567–578. [PubMed: 19568283]
32. Roti G, Stegmaier K. Genetic and proteomic approaches to identify cancer drug targets. British journal of cancer. 2012; 106:254–261. [PubMed: 22166799]
33. Fuchs F, et al. Clustering phenotype populations by genome-wide RNAi and multiparametric imaging. Mol Syst Biol. 2010; 6:370. [PubMed: 20531400]
34. Bakal C, Aach J, Church G, Perrimon N. Quantitative morphological signatures define local signaling networks regulating cell morphology. Science. 2007; 316:1753–1756. [PubMed: 17588932]
35. Taylor DL. Past, present, and future of high content screening and the field of cellomics. Methods in molecular biology. 2007; 356:3–18. [PubMed: 16988391]
36. Gustafsdottir SM, et al. Multiplex cytological profiling assay to measure diverse cellular states. PloS one. 2013; 8:e80999. [PubMed: 24312513]

37. Wawer MJ, et al. Toward performance-diverse small-molecule libraries for cell-based phenotypic screening using multiplexed high-dimensional profiling. Proceedings of the National Academy of Sciences of the United States of America. 2014; 111:10911–10916. [PubMed: 25024206]
38. Cohen AA, et al. Dynamic proteomics of individual cancer cells in response to a drug. Science. 2008; 322:1511–1516. [PubMed: 19023046]
39. Loo LH, Wu LF, Altschuler SJ. Image-based multivariate profiling of drug responses from single cells. Nature methods. 2007; 4:445–453. [PubMed: 17401369]
40. Johnson, RA.; Wichern, DW. Applied multivariate statistical analysis. 3rd.. Prentice Hall; Englewood Cliffs, N.J.: 1992.
41. Martin CJ, et al. Molecular characterization of macbecin as an Hsp90 inhibitor. Journal of medicinal chemistry. 2008; 51:2853–2857. [PubMed: 18357975]
42. Reddy P, et al. Histone deacetylase inhibitor suberoylanilide hydroxamic acid reduces acute graft-versus-host disease and preserves graft-versus-leukemia effect. Proceedings of the National Academy of Sciences of the United States of America. 2004; 101:3921–3926. [PubMed: 15001702]
43. Wojcik C, et al. Valosin-containing protein (p97) is a regulator of endoplasmic reticulum stress and of the degradation of N-end rule and ubiquitin-fusion degradation pathway substrates in mammalian cells. Molecular biology of the cell. 2006; 17:4606–4618. [PubMed: 16914519]
44. Kuhn DJ, et al. Potent activity of carfilzomib, a novel, irreversible inhibitor of the ubiquitin-proteasome pathway, against preclinical models of multiple myeloma. Blood. 2007; 110:3281–3290. [PubMed: 17591945]
45. Chen D, Frezza M, Schmitt S, Kanwar J, Dou QP. Bortezomib as the first proteasome inhibitor anticancer drug: current status and future perspectives. Current cancer drug targets. 2011; 11:239–253. [PubMed: 21247388]
46. Kim TS, et al. Interaction of Hsp90 with ribosomal proteins protects from ubiquitination and proteasome-dependent degradation. Molecular biology of the cell. 2006; 17:824–833. [PubMed: 16314389]
47. Moffat JG, Rudolph J, Bailey D. Phenotypic screening in cancer drug discovery - past, present and future. Nature reviews. Drug discovery. 2014; 13:588–602. [PubMed: 25033736]
48. Kangas JD, Naik AW, Murphy RF. Efficient discovery of responses of proteins to compounds using active learning. BMC bioinformatics. 2014; 15:143. [PubMed: 24884564]
49. Sigal A, et al. Dynamic proteomics in individual human cells uncovers widespread cell-cycle dependence of nuclear proteins. Nature methods. 2006; 3:525–531. [PubMed: 16791210]
50. Schneider CA, Rasband WS, Eliceiri KW. NIH Image to ImageJ: 25 years of image analysis. Nature methods. 2012; 9:671–675. [PubMed: 22930834]
51. Sugiyama M. Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis. J Mach Learn Res. 2007; 8:1027–1061.
52. Schafer J, Strimmer K. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Stat Appl Genet Mo B. 2005; 4
53. Wu J, Hu CP, Gu QH, Li YP, Song M. Trichostatin A sensitizes cisplatin-resistant A549 cells to apoptosis by up-regulating death-associated protein kinase. Acta pharmacologica Sinica. 2010; 31:93–101. [PubMed: 20048748]
54. Ono M, et al. Sensitivity to gefitinib (Iressa, ZD1839) in non-small cell lung cancer cell lines correlates with dependence on the epidermal growth factor (EGF) receptor/extracellular signal-regulated kinase 1/2 and EGF receptor/Akt pathway for proliferation. Molecular cancer therapeutics. 2004; 3:465–472. [PubMed: 15078990]
55. Chen MC, et al. The HDAC inhibitor, MPT0E028, enhances erlotinib-induced cell death in EGFR-TKI-resistant NSCLC cells. Cell death & disease. 2013; 4:e810. [PubMed: 24052078]

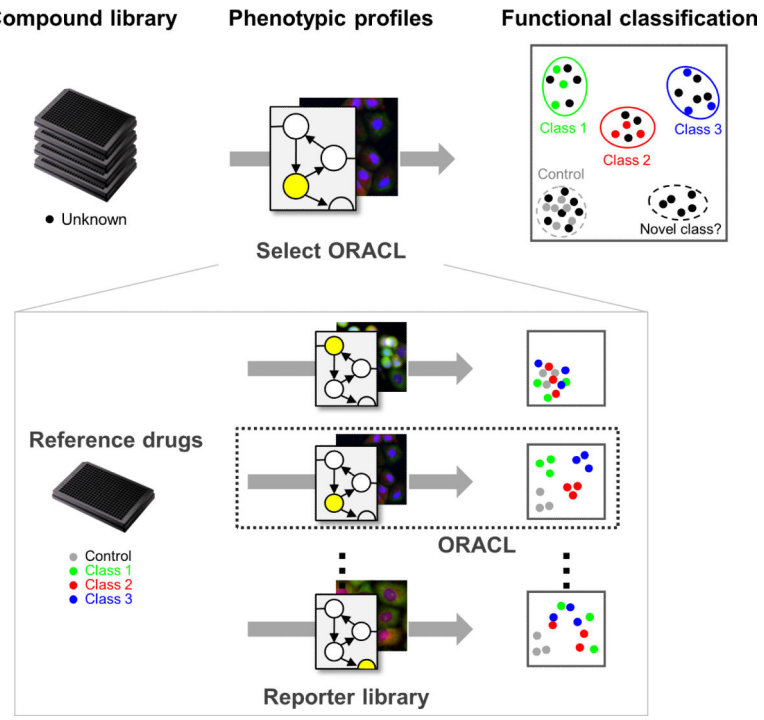

Figure 1. Overview of method

Overview of image-based phenotypic screening steps: Libraries of compounds (left) are applied to cells labeled with biomarkers (middle); cellular responses are extracted from images and used to construct a phenotypic profile (right; cartooned in two dimensions as black points); and compounds are functionally classified (i.e. annotated) based on comparison to phenotypic profiles of known, reference drugs (colored circles). Overview of approach for selecting an Optimal Reporter cell line for Annotating Compounds Libraries, called an “ORACL.” We profile a collection of reference drugs using reporter cell lines labeled for diverse biomarkers. Our ORACL is defined as the reporter cell line whose phenotypic profiles give the highest classification accuracy of the reference drugs. We select this ORACL for large- scale phenotypic screens of unknown compound libraries.

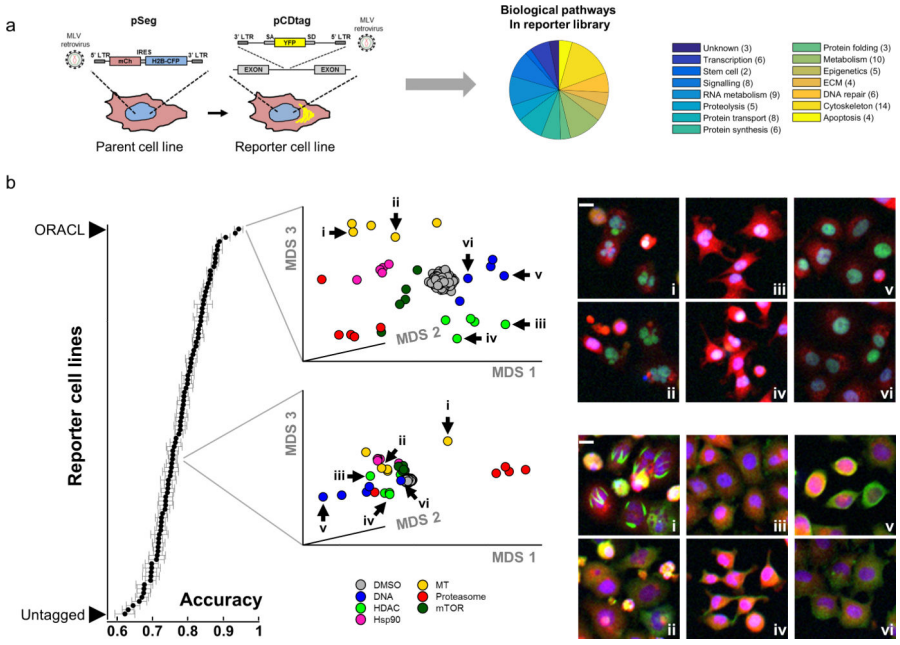

Figure 2. An ORACL is identified that best distinguishes among drug classes

(a) A parent A549 cell line was built with a construct (pSeg) to express cytosolic (mCherry) and nuclear (H2B-CFP) fluorescent proteins to aid in automated cellular region identification. A library of diverse reporter cell lines were built from this parent line using a strategy (CD tagging) that randomly incorporated YFP into different proteins (one per reporter cell line). “Untagged” refers to the parental pSeg-tagged line that lacks a CD tag. (b) **Left:** Drug classification accuracies for each of our 93 CD-tagged reporters. Mean (black dots) and standard deviation (gray bar) of prediction accuracies were calculated from 100 cross-validation studies (Online Methods). **Middle:** drug-response profiles of the ORACL and a “mediocre” reporter cell line were visualized by MDS plot (top and bottom, tagged for XRCC5 or SEPT11 respectively). Each drug (or DMSO) profile is represented by a point and colored according to the drug classes. **Right:** Representative cellular response images for the indicated drugs in the MDS plots at left. The ORACL shows consistent phenotypes within drug classes, whereas the “mediocre” reporter cell line shows inconsistent phenotypes within the same drug classes. Fluorescent reporters: Blue: CFP-nuclear label; Red: mCherry-cytosolic label; Green: YFP-CD tag (intensity scale is the same for Blue and Red, but is adjusted for Green per reporter cell line). Scale bar: 10 μm. Drugs: **i:** Epithilone B; **ii:** Nocodazole; **iii:** Apicidin; **iv:** Oxamflatin; **v:** CPT; **vi:** Etoposide.

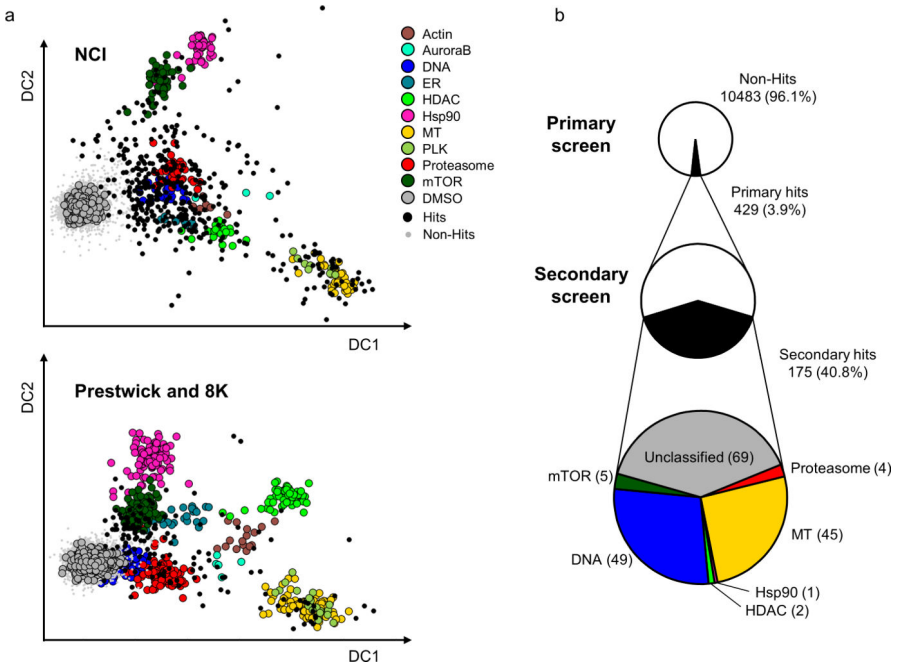

Figure 3. Compound hits across multiple drug classes are identified from a single-pass screen

(a) Shown are LDA projections of phenotypic profiles for reference drugs and compounds in batch 1 (NCI) and batch 2 (Prestwick and 8K). Profiles were computed by concatenating data from 24 and 48 hrs. Each point represents the projected profile for a tested compound and concentration. Reference drugs are colored according to drug classes. Hits and non-hits are shown as black or grey dots, respectively. **(b)** Summary of screen: proportion of primary (top) or secondary "high confidence" (middle) hits, and distribution of predicted drug classes for hits (bottom). DC: discriminant component.

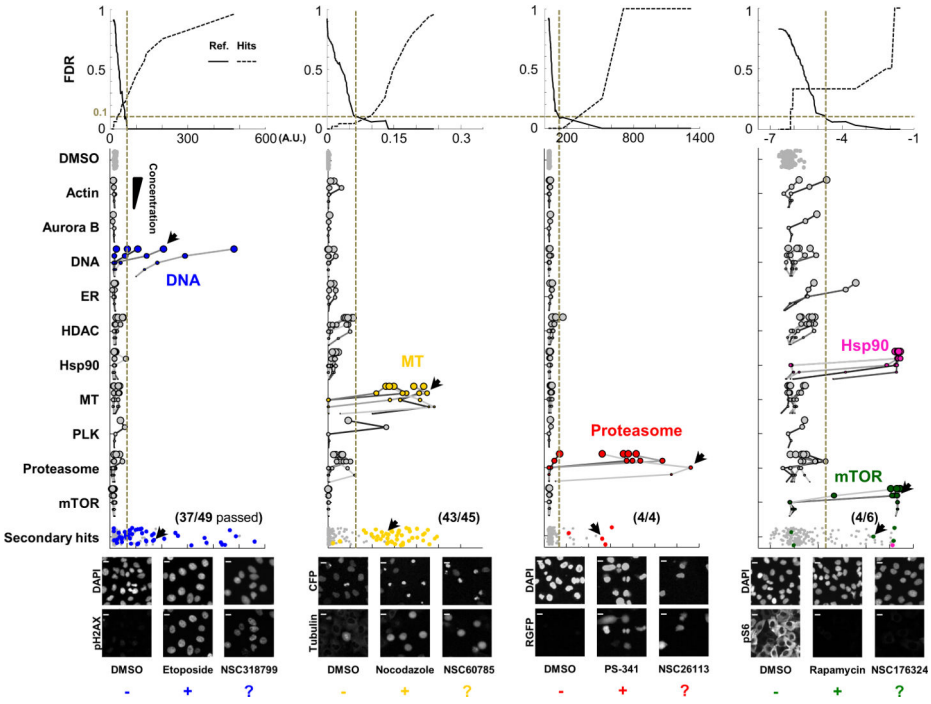

Figure 4. Secondary studies validate predictions across diverse drug classes

Top: False discovery rates (FDR; y-axes) were calculated for 38 reference drugs (FDR_{Ref}, solid line) or 175 high-confidence hits (FDR_{hits}, dashed line) at different thresholds for readouts selected in each validation assay (x-axes; Online Methods). Vertical dark gold dashed lines: readout thresholds at FDR_{Ref} = 0.1 (horizontal dark gold dash line). **Middle:** Readout values (x-axes) of DMSO, reference drugs (at five, 5-fold serial dilutions), and 175 high-confidence hits were shown for each validation experiment. Reference drugs were grouped according to drug classes; each line represents the dose response of one drug. Circle size reflects the concentrations (larger size indicates higher concentration). High-confidence hits that were predicted to belong to the class being validated were highlighted with corresponding colors. **Bottom:** representative images of cells treated with DMSO (−), positive control reference drugs (+), and secondary hits (?) indicated by black arrows in middle panel. Scale bar = 10 μm.

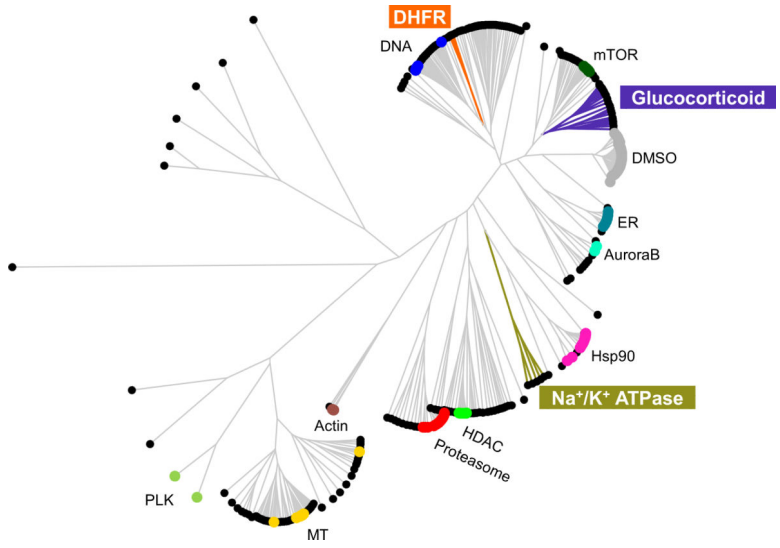

Figure 5. The ORACL can identify novel compound groupings

Compound clusters were identified by hierarchical clustering (see Online Methods). Colored dots correspond to reference drugs. Colored labels and lines indicate examples of clusters that contain multiple, consistently annotated compounds in drug classes not used in the selection of the ORACL.