HHS Public Access

Author manuscript

Cell Syst. Author manuscript; available in PMC 2018 May 14.

Published in final edited form as:

Cell Syst. 2018 April 25; 6(4): 424–443.e7. doi:10.1016/j.cels.2018.03.012.

A Library of Phosphoproteomic and Chromatin Signatures for Characterizing Cellular Responses to Drug Perturbations

Lev Litichevskiy[1], **Ryan Peckner**[1], **Jennifer G. Abelin**[1,4], **Jacob K. Asiedu**[1], **Amanda L. Creech**[1,4], **John F. Davis**[1], **Desiree Davison**[1], **Caitlin M. Dunning**[1,5], **Jarrett D. Egertson**[3], **Shawn Egri**[1], **Joshua Gould**[1], **Tak Ko**[2], **Sarah A. Johnson**[1], **David L. Lahr**[1], **Daniel Lam**[1], **Zihan Liu**[1], **Nicholas J. Lyons**[1], **Xiaodong Lu**[1], **Brendan X. MacLean**[3], **Alison E. Mungenast**[2,6], **Adam Officer**[1], **Ted E. Natoli**[1], **Malvina Papanastasiou**[1], **Jinal Patel**[1], **Vagisha Sharma**[3], **Courtney Toder**[1], **Andrew A. Tubelli**[1], **Jennie Z. Young**[2], **Steven A. Carr**[1], **Todd R. Golub**[1], **Aravind Subramanian**[1], **Michael J. MacCoss**[3], **Li-Huei Tsai**[2], and **Jacob D. Jaffe**[1,7,*]

[1]The Broad Institute, 415 Main Street, Cambridge, MA 02142, USA

[2]Department of Brain and Cognitive Sciences, Picower Institute for Learning and Memory, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, USA

[3]University of Washington, Department of Genome Sciences, 3720 15th Avenue NE, Seattle, WA 98195, USA

SUMMARY

Although the value of proteomics has been demonstrated, cost and scale are typically prohibitive, and gene expression profiling remains dominant for characterizing cellular responses to perturbations. However, high-throughput sentinel assays provide an opportunity for proteomics to contribute at a meaningful scale. We present a systematic library resource (90 drugs 3 6 cell lines) of proteomic signatures that measure changes in the reduced-representation phosphoproteome (P100) and changes in epigenetic marks on histones (GCP). A majority of these drugs elicited reproducible signatures, but notable cell line- and assay-specific differences were observed. Using the “connectivity” framework, we compared signatures across cell types and integrated data across

*Correspondence: jjaffe@broadinstitute.org.
[4]Present address: Neon Therapeutics, Cambridge, MA, USA
[5]Present address: Waters Corporation, Milford, MA, USA
[6]Present address: Eisai Pharmaceuticals, Andover, MA, USA
[7]Lead Contact

SUPPLEMENTAL INFORMATION

Supplemental Information includes eight figures, one table, and four data files and can be found with this article online at https://doi.org/10.1016/j.cels.2018.03.012.

AUTHOR CONTRIBUTIONS

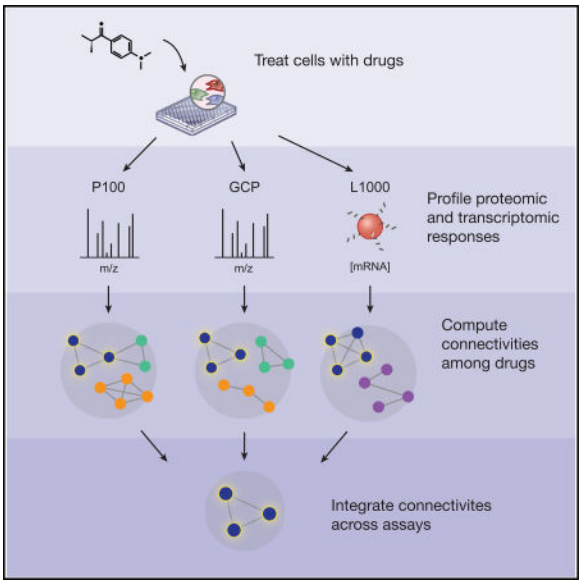
Conceptualization, A.S., L.-H.T., M.J.M., and J.D.J.; Methodology, L.L., R.P., J.G.A., A.L.C., J.D.E., C.M.D., T.K., T.E.N., J.P., and J.D.J.; Software, L.L., B.X.M., V.S., J.K.A., J.G., C.T., Z.L., and J.D.J.; Formal Analysis, L.L., R.P., D.L.L., and J.D.J.; Investigation, L.L., A.L.C., J.F.D., D.D., S.E., C.M.D., T.K., S.A.J., D.L., N.J.L., X.L., A.E.M., A.O., M.P., and J.P.; Resources, L.-H.T. and M.J.M.; Data Curation, L.L., D.L.L., and J.D.J.; Writing – Original Draft, L.L. and J.D.J.; Writing – Review & Editing, T.E.N., S.A.C., and A.S.; Visualization, L.L., A.A.T., and J.D.J.; Supervision, A.E.M., M.P., J.Z.Y., M.J.M., L.H.T., and J.D.J.; Project Administration, J.P., M.P., and J.D.J.; Funding Acquisition, T.R.G., A.S., and J.D.J.

DECLARATION OF INTERESTS

M.J.M. discloses that he is a paid consultant for Thermo Fisher’s life science mass spectrometry group. Additionally the laboratory of M.J.M. has a sponsored research agreement with Thermo Fisher.

assays, including a transcriptional assay (L1000). Consistent connectivity among cell types revealed cellular responses that transcended lineage, and consistent connectivity among assays revealed unexpected associations between drugs. We further leveraged the resource against public data to formulate hypotheses for treatment of multiple myeloma and acute lymphocytic leukemia. This resource is publicly available at https://clue.io/proteomics.

In Brief



A large compendium of cellular responses to drugs as profiled through proteomic assays of phosphosignaling and histone modifications reveals cellular responses that transcend lineage, discovers unexpected associations between drugs, and recognizes therapeutic hypotheses for treatment of multiple myeloma and acute lymphocytic leukemia.

INTRODUCTION

Molecular profiling technologies have enabled tremendous advances in biomedicine ranging from basic mechanistic insights to active guidance of therapeutic choices in precision medicine. For example, a gene expression signature (PAM50) now constitutes one of the major diagnostic classifiers for breast cancer (Parker et al., 2009). Gene expression profiling has long held sway as the technology of choice for generating systematic, comprehensive datasets of sufficient size to power statistical analyses. Early landmark studies (Alizadeh et al., 2000; Bittner et al., 2000; Clark et al., 2000; Golub et al., 1999; Perou et al., 2000) demonstrated the power of molecular profiling, paved the way for its acceptance into the mainstream, and ultimately drove costs down and technology forward. It has even been proposed that gene expression profiling could serve as the "'universal language' with which to describe cellular responses," from which the concept of the Connectivity Map—linking drugs, genes, and phenotypes through expression profiles—has emerged (Lamb et al., 2006; Subramanian et al., 2017).

Yet it is known that mRNA levels alone do not fully capture cell state, and the vocabulary of this universal language may need to be extended to describe some of the critical functions of cellular responses. Early studies recognized apparent discordance between mRNA and protein levels on a large scale (Greenbaum et al., 2003), and current studies suggest that the correlation coefficient between mRNA and protein levels is only ~0.5 (Mertins et al., 2016). More recently, Li et al. (2017) have shown that phosphorylation events show low correlation with mRNA levels from their corresponding genes, and therefore phosphoproteomic data are likely to add value to gene expression measurements. These observations underscore the importance of exploring complementary readouts to gene expression profiling.

Alternative profiling methodologies could potentially fill the gaps in gene expression profiling by measuring analytes that cannot be detected via nucleic acid and reporting on biological processes with time scales distinct from changes in gene expression. Such methods must also be scalable to prosecute systematic studies of sufficient size to power discovery while remaining cost-effective enough to deploy on a routine basis. Furthermore, it would be highly desirable for molecular profiling assays to directly report on cellular processes affected by novel drug candidates' primary modes of action, which usually involve inhibition of protein activity rather than the achievement of a particular transcriptional state. Two emerging classes might particularly benefit from such directed assays: (1) targeted kinase inhibitors, a class of drugs that is rapidly expanding (Wu et al., 2015); and (2) inhibitors of chromatin-modifying enzymes and sensors of chromatin state, which have emerged as exciting new therapeutic modalities (Dawson et al., 2012; Kelly et al., 2010). While these new therapies are extremely promising, deeper insight is required to fully understand their cellular effects in their intended biological contexts, as well as their possible off-target effects in a system-wide manner. At the same time, many established protein-targeting drugs lack clear mechanistic insight and may harbor unexpected phosphosignaling and epigenetic activities that would be useful in repurposing efforts (Gupta et al., 2013; Singhal et al., 1999).

To fill these needs, we set out to develop a reference resource of proteomic signatures elicited in response to drugs, specifically monitoring phosphosignaling and chromatin state, two key areas for therapeutic development. Unlike recent proteomic resources characterizing the proteomic *states* associated with genetic variation in tumors and cell lines (Li et al., 2017; Mertins et al., 2016), this resource characterizes the proteomic *responses* to systematic application of drug perturbagens. Here, we describe the creation and validation of a pilot library containing signatures from our reduced-representation phosphoproteomic assay—a sentinel assay that reports on a broad range of signaling pathways (Abelin et al., 2016)—and our global chromatin profiling assay (Creech et al., 2015). These proteomic assays are highly standardized and rigorously quantitative, yet automated and compact enough to achieve a relatively high throughput and reasonable scale. The relatively large scale of the data (compared with other proteomic perturbation studies) allows for a principled query of proteomic signatures that, for the first time, allows us to extend the Connectivity Map concept (Lamb et al., 2006; Subramanian et al., 2017) to proteomics data and easily integrate proteomic with transcriptomic data. We show that the resource itself contains a wealth of information about drug activity in cells, and also that the resource can be leveraged for analysis of external data to recognize potential therapeutic opportunities. This

data resource contributes to the NIH Library of Integrated Network-based Cellular Signatures (LINCS) program (Keenan et al., 2017), and represents a first-in-class queryable proteomics resource of over 3,400 profiles. Importantly, the resource continues to grow beyond its current size with longitudinal support from LINCS, and opens the door to both utilization by and contributions from the community at large.

RESULTS

Structure and Scope of the Proteomic Signature Resource

We created a library of proteomic perturbational signatures containing more than 3,400 samples (Figure 1A). In this initial dataset generated over 21 months, we profiled 90 small molecules spanning a variety of mechanisms of action (MoAs) with focused subsets of kinase inhibitors and epigenetically active compounds (Table S1). We utilized five widely studied cancer cell models (representing breast, lung, pancreatic, prostate, and skin lineages) and one neurodevelopmental cell model, neural progenitor cells (NPCs), to test a non-cancer model. Samples were profiled with a reduced-representation phosphoproteomic assay (P100) and a global chromatin profiling (GCP) assay, both of which are liquid chromatography-mass spectrometry (LC-MS)-based and previously described (Abelin et al., 2016; Creech et al., 2015). The analytes measured by P100 are ~100 phosphorylated peptides from cellular proteins, and the analytes measured by GCP are ~60 post-translationally modified peptides from histones (e.g., methylated, acetylated, and combinations thereof). The analytes measured by P100 serve as a reduced representation of the phosphoproteome and act as early sentinels of bioactivity of a diverse set of signaling pathways and drug mechanisms. The analytes measured by GCP include nearly every well-studied post-translational modification on the core nucleosomal histone proteins. These modifications convey epigenetic information in cells, and their dysregulation is associated with a wide range of diseases (Araf et al., 2016; Aumann and Abdel-Wahab, 2014; Gräff and Mansuy, 2009; Jaffe et al., 2013; Ntziachristos et al., 2016; Peña et al., 2014). P100 profiles were collected 3 hr after treatment, and GCP profiles were collected 24 hr after treatment (see Drug Treatment in STAR Methods).

Mass spectrometry data are summarized by Skyline software (MacLean et al., 2010) to yield a quantification for each analyte of the amount of peptide that was detected relative to a spiked-in synthetic control. Further data processing is performed by the Proteomics Signature Pipeline (https://github.com/cmap/psp). In brief, data are $\log_2$-transformed, analytes and samples with an excess of missing data are filtered out, a constant offset is applied to each sample to bring its range of values to the same scale as the other samples on a plate, and analytes are median normalized (see Data Processing for GCP and P100 in STAR Methods). All data presented in this study, at multiple levels of processing (Figure S1A), are publicly available alongside extensive analyte and experimental metadata (Key Resources Table). Unless explicitly noted otherwise, all of the following analyses utilized the most processed form of data (level 4).

Analyte-Centric Analysis of the Resource

Proteomic assays can be evaluated either at the level of individual analytes or by considering the entire profile of changes as the relevant unit of measure. While the biological interpretation of the GCP analytes (post-translational modifications of histones) is facilitated by prior research (Jenuwein and Allis, 2001; Martin and Zhang, 2005), the P100 phosphopeptides are not as directly interpretable because they were derived in a data-driven, rather than prior knowledge-driven, manner. Although we demonstrated the general utility of the P100 analytes in a prior publication (Abelin et al., 2016), we used our expanded dataset to look further into direct biological interpretations of the P100 peptides.

We performed an analysis to look for strong changes in individual P100 peptides that were associated with certain drugs or MoA classes. Several vignettes are presented in Figure 1. For example, we directly monitor phosphorylation of S100 on JUND, a downstream target of JNK kinases (Lamb et al., 2003; Yazgan and Pfarr, 2002). Both JNK inhibitors in our set, SP600125 (pyrazolanthrone) and CC-401, cause a decrease in the levels of JUND S100ph (Figure 1B).

In a novel observation, we found that phosphorylation of S3426 on AHNAK is downregulated by the MEK (MAP2K1) inhibitors selumetinib and PD-0325901 (Figure 1C). It is also selectively downregulated by vemurafenib in A375 cells, which are sensitive to this drug by virtue of a mutation in BRAF (see the discussion of multi-assay data integration below). AHNAK is a very large protein (5,890 amino acids) with numerous documented phosphosites (Bian et al., 2014; Zhou et al., 2013), many of which were observed in our prior studies (Abelin et al., 2016). However, in those studies we found that S3426ph behaves differently from most other sites on AHNAK with respect to perturbation response, and it was included in the P100 panel. AHNAK has been implicated in the regulation of MAPK/ERK pathway signaling, and specifically in the regulation of phosphorylation of MEK and RAF (Lee et al., 2008). It has been shown to activate protein kinase C, an upstream kinase of RAF. It seems plausible that there might be feedback from the inhibition of RAF and MEK to phosphosites on AHNAK.

As a parallel example in GCP, we can demonstrate interesting regulation of ubiquitination on lysine 18 on histone H3 (H3K18ub1) in response to inhibition of DNA methylation. Decitabine, a DNA methyltransferase inhibitor, significantly upregulates the abundance of the H3K18ub1K23ac0 histone mark (Figure 1D). In a prior study, Qin et al. (2015) showed that H3K18ub1 is required for DNA methylation. Here, we show a feedback loop where prevention of DNA methylation leads to increased levels of this required histone mark.

We suspect there are many similar vignettes that could be demonstrated for other P100 and GCP analytes, each of which would require systematic investigation and follow-up that we consider to be beyond the scope of this resource. Rather, we wanted to examine the question of whether individual analytes or entire profiles would be more valuable for further analysis in general. We tried to assess this by asking (1) whether responses to individual drugs were dominated by large changes to small numbers of analytes or whether the inverse was true, and (2) whether dominant analytes for drugs were transferable across cell types.

We started by defining “dominant” analytes as those that changed by an absolute magnitude of 3-fold for a given drug response ($|\log_2| \geq 1.58$). Out of 540 potential drug-cell type combinations in each assay, fewer than 33% had even a single dominant analyte (Figure 1E). In contrast, almost all analytes were dominant in at least one drug-cell type condition (70/96 in P100, 51/59 in GCP; Figure 1F), speaking to the orthogonality of the analytes chosen. Yet the signal (total change from 0) derived from dominant analytes was dwarfed by smaller changes from weaker analytes (Figure 1G). We also found that dominant analytes for a given drug were unlikely to be transferable across cell types. Of all potential dominant analyte-drug combinations (70 × 90 for P100, 51 × 90 for GCP), fewer than 1% recurred in four or more of the six cell types, with almost none in P100 (Figure 1H).This latter result illustrates the difficulty in obtaining truly universal biomarkers that can be used across many biological systems. We concluded from these analyses that the bulk of the signal measured in these assays comes from relatively small changes to a wide array of analytes, and even strong analytes were unlikely to be universal. Because of these factors and the difficulty of interpreting analyte-centric data on a case-by-case basis, we focused on using entire profiles of analytes as the unit of comparison among conditions. There are other structural advantages to this approach as well, such as easier integration of data from disparate assays, discussed below.

Connectivity: A Profile-Centric Analysis Framework

We adopted the concept of connectivity as a consistent, assay-independent means of identifying two or more drugs that elicit similar molecular response profiles in cells (Lamb et al., 2006). It enables us to ask the question “which compounds look alike in our dataset?” without focusing on the biological meanings of the individual analytes measured. For example, if we recognized a connection between a drug of known mechanism and one of unknown mechanism, we could formulate a hypothesis about the unknown drug. This is a prime use case for the resource in evaluation of novel chemical matter in the future.

The connectivity score between two compounds is a background-corrected Spearman correlation between their molecular profiles (Figure 2A; see Computing Similarities and Connectivities in STAR Methods). A connectivity score of 1 indicates that two perturbations are more similar to each other than all other pairs of perturbations, 0 indicates that their similarity is unexceptional, and −1 indicates that they are less similar to each other than all other pairs of perturbations (Figures 2B–2D). Drugs with positive connectivity elicit similar underlying cellular states, whereby phosphopeptides serve as a proxy for signaling state in P100 and chromatin modifications serve as a proxy for epigenetic state in GCP. Drugs with negative connectivity elicit very different cellular states with anti-correlated profiles. The connectivity score between two compounds takes into account how well correlated the two compounds are to all other compounds, but its values strongly resemble the underlying correlations (Figure S1B).

Connectivity scores between two or more drugs can be visualized using heatmaps or network views (Figure 2A, far right) and subjected to further quantitative interrogation. Complete matrices of proteomic connectivity data can be browsed and interactively manipulated at https://clue.io/proteomics. Network views are directed or undirected graphs

representing perturbations as nodes and connectivity scores above a user-defined threshold as edges. Unlike heatmaps, network views make it easy to see different modules of biology at a glance.

Proteomic Assays Produce Reproducible Signatures that Capture Diverse Cellular Responses

Before applying the connectivity analysis to the resource, we wanted to ensure that the underlying data were of high quality. We assessed the technical quality of our datasets by quantifying replicate reproducibility at the profile level and comparing with more common analytical metrics of reproducibility. First, we compared the global distributions of replicate correlations with non-replicate (Spearman) correlations (Figure 3A). The distributions were significantly different, indicating that repeated profiling of the same drug treatment produced responses that were similar to each other and distinct from other drug treatments (two-sample Kolmogorov-Smirnov test: P100: $D = 0.67$, $p < 10^{-15}$; GCP: $D = 0.62$, $p < 10^{-15}$). For comparison with another large-scale phosphoproteomic profiling effort, we note that our global replicate correlations resemble those of Li et al. (2017) in their profiling of 650 cell lines using RPPA technology (Figures S2A–S2D).

Next, we asked how many individual perturbagens were reproducible in each cell type. A perturbagen was considered reproducible if its replicates were significantly correlated compared with a permutation null (see Replicate Reproducibility in STAR Methods). At a q-value threshold of 0.05, we found that in each cell line in both assays, at least 53 out of the 90 compounds profiled (average = 69.0 compounds, or 76.7%) were reproducible (Figure 3B). Furthermore, in each cell line at least 48 compounds (53.3%) were reproducible in both assays (shaded portion), indicating that a compound was likely to be reproducible in both assays if it was reproducible in one. These results are in line with the fraction of small molecules determined to be bioactive using other profiling modalities: 68.3% for Cell Painting morphological profiling (Wawer et al., 2014) and 38% for L1000 gene expression profiling (Subramanian et al., 2017).

In each cell line, more compounds were reproducible in P100 than in GCP (mean difference = 17.7 compounds). Downsampling P100 to the same number of analytes as GCP (n = 59) revealed that P100's larger feature space only partially explained the discrepancy in replicate reproducibility; even when down-sampled to 60 analytes, P100 still called an average of 12.1 more compounds reproducible than GCP (Figure S2E). This finding indicates that the larger feature space of P100 does not fully explain the discrepancy in replicate reproducibility. Our downsampling analysis also showed that marginal increases in the number of analytes consistently led to improvements in replicate reproducibility. This observation further emphasizes that the P100 sentinel analytes are biologically meaningful, since incrementally adding more analytes improves our ability to discriminate between replicate and non-replicate drug treatments.

The compounds that were reproducible in P100 but not GCP in at least four cell lines (gossypetin, rolipram, olaparib, TBB, tacrolimus, and everolimus) tended to be annotated with kinase-directed activities. We speculate that a bias in our compound selection toward kinase-modulating drugs may further explain why the phosphoproteomic assay had better

replicate reproducibility. We also note that replicate reproducibility was similar between NPCs and the cancer cell models. This observation emphasizes that both assays work as well in unusual cellular contexts, such as a neurodevelopmental cell line, as in the more typical cancer cell models.

To assess the long-term stability of our assays, we utilized the fact that we profiled the same positive controls on many plates over the course of data generation. The positive control for P100 was staurosporine, a pleiotropic kinase inhibitor, and the positive control for GCP was vorinostat, a potent histone deacetylase (HDAC) inhibitor.

As expected, samples profiled on the same plate were slightly more comparable with each other than samples produced on different plates. Namely, Figure S3A shows that the median correlation among positive control samples profiled on the same plate (P100, staurosporine: 0.89; GCP, vorinostat: 0.87) is higher than the median correlation among samples profiled on different plates but within the same cell line (P100, staurosporine: 0.81; GCP, vorinostat: 0.84). Correlations among samples profiled in different cell lines were lower (P100, staurosporine: 0.69; GCP, vorinostat: 0.79), indicating that cell type-to-cell type variability is greater than plate-to-plate variability.

In addition, we also quantified sample comparability using the coefficient of variation (CV). Figure S3B demonstrates the same trends as those observed using correlations in Figure S3A. That is, analyte CVs of samples on the *same* plate are lower than CVs of samples on different plates, which are lower than analyte CVs of samples profiled in *different* cell lines. Across-plate but within-cell type CV values of 0.35 (P100) and 0.21 (GCP) demonstrate good plate-to-plate variability and stable long-term performance of our assays.

Using Proteomic Connectivities to Detect and Refine Mechanisms of Action

Having validated the technical quality of our datasets, we sought to investigate how well various annotated MoAs were detected by each assay. For this analysis, we define intra-class connectivity as the median connectivity score among compounds belonging to the same MoA (Figure 3C; see Table S1 for compound annotations). In both assays, the majority of MoA classes had positive intra-class connectivities. Several MoA classes, such as bromodomain (BRD) inhibitors and MEK inhibitors, had high connectivity in both assays (Figure 3D, left).

Other MoA classes, such as JNK inhibitors, had low—in fact, negative—intra-class connectivity in both assays (Figure 3D, right). In the case of the two JNK inhibitors that we profiled, both compounds had strong self-connectivity (suggesting high replicate reproducibility and therefore a definitive signal of each in both assays) but failed to connect to each other. This is despite the observation in our data that both JNK inhibitors (CC-401 and SP600125) cause downregulation of a known JNK substrate (JUND, S100ph; Figure 1B). We posit that this lack of connectivity reflects significant off-target effects in one or both of the compounds at the doses tested. Indeed, a screen of 20 kinase inhibitors highlighted one of our two JNK inhibitors—SP600125—as having very poor separation between its affinity for intended targets and off-targets, indicating that this drug has substantial off-target effects (Fabian et al., 2005). Our data show that the next-best

connected compound in our dataset to SP600125 in P100 is the JAK2 inhibitor TG101348. This example illustrates the power of using unbiased assays and our connectivity framework to recognize unappreciated off-target effects, and suggest alternative hypotheses about MoA.

Connectivity analysis can provide refinement to ambiguous MoA classes in an assay-specific manner. We included three compounds thought to modulate the activity of the sirtuin class of histone deacetylases: resveratrol, EX527, and salermide. EX527 and salermide are sirtuin inhibitors, but with differing selectivity (Peck et al., 2010). Resveratrol is typically annotated in literature as a sirtuin activator (Baur and Sinclair, 2006; Wendling et al., 2013), but it has also been argued that resveratrol does not directly act on sirtuins at all (Beher et al., 2009; Pacholec et al., 2010). Because of this ambiguity, we grouped resveratrol and the two known sirtuin inhibitors into a single MoA class. In our analysis, this class had positive connectivity in P100 (0.49) but negative connectivity in GCP (−0.47). Upon investigation, we discovered that the negative connectivity in GCP is explained by resveratrol's strong negative connections to the other two compounds in the class, EX527 and salermide (Figure 3D, middle). The negative connectivity in GCP between resveratrol and the sirtuin inhibitors indicates that these compounds have very different effects on chromatin, while the positive connectivity in P100 among all three compounds suggests that there are common signaling pathways activated regardless of whether sirtuins are activated or inhibited. Looking at GCP profiles, it was evident that resveratrol induced opposite effects from the other two compounds and is most likely a sirtuin activator. However, all three compounds are positively connected to geldanamycin, a heat-shock protein 90 (HSP90) inhibitor (Whitesell et al., 1994), in the P100 data. As a chaperone, HSP90 plays an important role in protein homeostasis, folding, and complex formation (Whitesell et al., 2003). Our results suggest that both positive and negative regulation of sirtuin activity provokes a stress response similar to HSP90 inhibition. Taken together, P100 data recognized that there was a common thread among the compounds involving stress response while GCP data provided the finer details about their opposing mechanisms.

In summary, our analysis of MoA intra-class connectivity has demonstrated the ability of the assays to capture cellular responses to a diverse set of therapeutic and investigational drugs. We show examples in which we are able to recapitulate known relationships as well as examples in which we can detect distinctions among class members.

Global Analysis of Connectivity Profiles Suggests that Diverse Signaling States Converge to a Restricted Set of Chromatin States

Our resource contains over 580,000 potential pairwise connections among 540 distinct drug-cell combinations (in each assay). Connectivity analysis provides an objective measure of associations among perturbations in cells. We can define a connectivity profile for each perturbation: simply a vector of connectivity scores to all other compounds. Because they are quantitative, connectivity profiles can be analyzed in the same manner as raw assay profiles with techniques such as clustering and principal component analysis. However, they have the advantage of reducing assay-specific considerations and provide a framework for comparing and integrating data across multiple assay types (see the discussion of multi-assay data integration below).

We first sought to understand the structure of the connectivities as a whole. As a first step, we projected all within-cell connectivity profiles (allowing drug-drug connections only within a single cell type) using t-stochastic neighbor embedding (t-SNE; Figure 4A and Supplemental Information) (van der Maaten and Hinton, 2008). We noticed that both P100 and GCP connectivity profiles organized into spatial clusters based on cell type, rather than MoA (Figure S4A). This was not true when P100 or GCP (level 4) profiles were projected in the same manner (Figures S4B and S4C). There was far less structure in the projections, and what structure could be found correlated more with drug mechanism. This spatial clustering of connectivity by cell type supports the notion that individual cell types are "wired" differently. Drugs that connect in one cell type may not connect in another, and likewise for entire groups of drugs.

We further noticed that the projection of GCP connectivity seemed more highly structured than that of P100 (Figure 4A, compare right with left), with fewer and tighter spatial clusters, each encompassing more data points. This difference in the number of connectivity profile clusters led us to wonder whether there were more phosphosignaling states than chromatin states that could be adopted by individual cell types in general. To address this question, we hierarchically clustered drugs (1 – Pearson correlation) by their connectivity profiles in each cell type and cut the resulting dendrograms at fixed percentages of their height (depicted schematically in Figure 4B, with results in Figure 4C). This analysis demonstrated that the number of clusters grows more quickly for P100 connectivity profiles than for GCP connectivity profiles as the cut percentage decreases. This alternative means of analysis provides further support for the existence of more phosphosignaling states available to cells than chromatin states.

To examine the cell type-specific properties of connectivity profile clusters, we mapped drugs from phosphosignaling connectivity clusters to chromatin connectivity clusters. The analysis was limited to reproducible compounds in both assays in each cell type as defined in Figure 3B, and dendrograms were cut at 60% of their height (average stability bootstrap probability of ~86%, single member clusters were eliminated; see Figure S5). We counted the number of times a drug from a phosphosignaling cluster was present in a chromatin cluster, and generated connectivity "flow" diagrams that represent this mapping (Figure 4D). Phosphosignaling clusters were annotated by the major signaling-active drug mechanisms in the dataset, and chromatin clusters by the major epigenetically active drug mechanisms analogously. These flow diagrams showed diverse phosphosignaling states channeling into a smaller number of chromatin states in a cell type-specific manner. MCF7 and YAPC cell lines appeared to have the most complicated flow structure, while PC3 and NPC cells were the most simplistic. Frequently, signaling clusters containing cell-cycle inhibitors mapped to chromatin clusters containing EZH2 inhibitors. Interestingly, NPC cells did not develop distinct chromatin connectivity clusters for EZH2 and HDAC inhibitors, although these classes may separate at a deeper dendrogram cut. BRD inhibitor-containing clusters also demonstrated unusual behavior, sometimes associating with HDAC inhibitors, other times associating with EZH2 inhibitors, and occasionally forming their own clusters. When we attempted to create an aggregated connectivity flow diagram across all cell types using the same methods, a very high number of connectivity clusters was formed (data not shown). We suspect that this is due to the intrinsic differences in the range of phosphosignaling and

epigenetic states available to each cell type. Indeed, the connectivity flows visualized in Figure 4D paint a complex regulatory picture where drug perturbations have cell type-specific effects that are potentiated by available responses. The one unifying theme among all connectivity flows was that a variety of different signaling states can be induced in each cell type, and these coalesce into a relatively smaller number of chromatin states.

Comparison of Proteomic and Transcriptomic Connectivities Demonstrates the Strengths of Assays Measuring Different Biological Dimensions

We envision that one major use of this resource will be to add value to, rather than replace, larger-scale profiling efforts using other readouts. To illustrate how transcriptional and proteomic profiling can be utilized in tandem, we produced data for our entire set of perturbagens using the L1000 transcriptional profiling assay (Subramanian et al., 2017). L1000 raw data were processed using the standard L1000 computational pipeline. Next, similarities and connectivities among perturbations in L1000 were computed exactly the same way as for proteomic data using our universal connectivity framework. In this way, we were able to directly compare connectivity results from proteomics (GCP and P100) with those from transcriptomics (L1000; Figure 5A). Replicate reproducibility and connectivity within MoA classes in L1000 were similar to results observed for the proteomic assays (Figure S6).

Our comparison of the three assays began with the following question: "How many of the strongest connections in one assay are also strong connections in another assay?" We defined a strong connection to be within the top 5% of connections for an individual cell line in each assay. We computed the percent overlap of the strongest connections (that is, the number of common connections divided by the number of considered connections) for each pair of assays either in each cell line individually or by aggregating all connectivities across cell lines using the median (Figure 5B). We found that the top 5% of connections had an average of 16.8% overlap for cell-specific results and an average of 31.7% overlap for aggregated results. Therefore, aggregating connectivity scores nearly doubled the percent overlap. This finding suggests that averaging over cell type-specific effects helps to improve agreement among the three methods of profiling. Both of these results are higher than the overlap expected by chance alone (5%). Still, it is clear that different assays are most sensitive to different cellular responses, given that their strongest connections have little agreement, especially for cell-specific results. We repeated this analysis using 1% and 10% cutoffs with similar results (Figures S7A and S7B).

We also computed the percent overlap among all three assays concurrently. At the 5% cutoff, there was 6.0% overlap for cell-specific connectivities and 16.0% overlap for aggregated connectivities (Figure S7C). This percent overlap was lower than for pairwise comparisons (because it is more difficult for a strong connection to show up in all three assays compared with just two of them), but again we observed that overlap improved after aggregating across cell lines.

As an alternative method of comparing assays, we looked at whether individual drugs had similar patterns of connectivity across assays. In contrast to the previous analysis in which we considered the top connections in each cell line, this analysis treated drugs separately.

For each drug, we compared its connectivity profile (see above) across assays (Figure 5C). We utilized an enrichment metric to compare connectivity profiles, but other similarity metrics, such as Spearman correlation, yielded similar results (Figure S7D). We found that an average of 14.8% of drugs had connectivity profiles that were highly similar across assays (recall >0.95; see Recall of Connectivity Profiles in STAR Methods). In brief, recall of 0.95 indicates that a drug's connectivity profile in one assay is more similar to its own connectivity profile in another assay than it is to 95% of the non-self-connectivity profiles (Figure S7E). The percentage of drugs with highly similar connectivity profiles more than doubled (33.7%) when connectivity scores were aggregated across cell lines. These quantifications of assay agreement were comparable with those reported by the overlap analysis (16.8% overlap without aggregation, 31.7% overlap with aggregation), indicating coherence between these two distinct analyses.

Both analyses highlighted NPCs as having especially low agreement across assays (red points in Figures 5B and 5C), yet replicate reproducibility was not appreciably lower in NPCs compared with the cancer cell models (Figure 3B). Our interpretation of these two findings taken together is that drugs treated in NPCs produced robust signatures, but many of these signatures were not distinct enough in the neural progenitor context to produce connectivity profiles that could be easily distinguished from the background. Figure 4D highlights that NPCs have one of the least complex connectivity flows in the dataset, and the restricted availability of signaling and chromatin states in NPCs is consistent with the integrative analysis here.

One novel observation revealed by the second analysis was that P100 and L1000 had considerably greater agreement with each other than with GCP. Considering aggregated connectivity scores, we observed that 53.3% of drugs had highly similar connectivity profiles (recall >0.95) when comparing P100 with L1000, while that number dropped to 26.7% and 21.1% for comparisons with GCP. Considering that P100 measures the reduced phosphoproteome while L1000 measures the reduced transcriptome, it is not surprising that these two assays are more similar to each other than they are to GCP, which measures the more narrow readout of chromatin changes. The fact that this trend was not observed in the overlap analysis indicates that all three assays are capable of reporting on the very strongest biological relationships, but there are individual compounds whose activity is not well captured through chromatin changes.

We sought to understand why there was little agreement among the strongest connections in each assay type. To investigate this question, we produced network views of the top connectivity scores for each assay aggregated (by median) across cell lines (Figure 5D), choosing only the top 0.5% of connections for visual clarity. This top 0.5% in each assay shows examples of strong connectivity among compounds, with mechanistic biases particularly likely to produce responses for the specific assay type. For instance, the GCP network shows five of the six HDAC inhibitors connecting strongly to each other, and all three EZH2 inhibitors connecting strongly to each other. We expected HDAC and EZH2 inhibitors to have strong signals in GCP because they are strong chromatin modifiers, and indeed we see connections between members of these classes in GCP but not in the other 2 assays (at this stringent threshold).

A response to modulation of the cell cycle was captured by all three assays in slightly different ways. All three assays reported the connection of dinaciclib to flavopiridol, which are both cyclin-dependent kinase (CDK) inhibitors, but these compounds connected to a variety of other compounds in each assay. Our interpretation of these results is that different assays detect different components of cell-cycle perturbation. It should also be noted that samples were collected at different time points after treatment, possibly demonstrating a time-dependent evolution in response. This diversity of readouts could be beneficial for researchers interested in one particular type of cellular response; for example, a researcher seeking to discriminate subtle differences among a group of kinase inhibitors at an early time point would likely be most interested in connectivity results reported by P100.

Integration of Proteomic and Transcriptomic Connectivities Reveals Cell Line-Selective Vulnerabilities

In addition to serving as a complementary hypothesis generation tool, proteomic profiling will add value to transcriptomic profiling (and other profiling technologies) by providing reinforcing data. Especially for unexpected connections, we hypothesized that support from more than one assay would increase the likelihood that a connection between two compounds would be confirmed as biologically meaningful. Therefore, we looked for examples of connections common to the three connectivity datasets.

To facilitate finding common connections, we computed the average connectivity score from the three assays; we refer to these data as AVG (Figure 6A). We visualized the strongest AVG connections aggregated across cell lines (again by computing the median of six cell-specific connectivity scores) with a network view (Figure 6B). In contrast to the networks in Figure 5D, this network shows only connections with support from all three assays.

Apart from the large cluster containing two cell-cycle inhibitors and a variety of other MoAs, all other clusters except one contained compounds with the same annotated MoA. The sole exception was the cluster containing TG101348 (circled) and the three BRD inhibitors. TG101348 is commonly annotated as a JAK2 inhibitor (Wernig et al., 2008), which makes its connection to BRD inhibitors perplexing. However, it was recently demonstrated through a BRD binding assay that TG101348 indeed has BRD inhibitor activity (Ciceri et al., 2014). This independent experiment strengthened our hypothesis that an unexpected connection seen by all three profiling assays would be likely to withstand experimental validation.

AVG data also reported compelling cell-specific connections. We immediately saw that vemurafenib (a BRAF inhibitor) was connected to the two MEK inhibitors only in A375 cells, an expected result. A375 cells are highly sensitive to both BRAF and MEK inhibition because they harbor the $BRAF^{V600E}$ mutation that makes them dependent on RAF-MEK-ERK signaling (Wagle et al., 2011). Looking at cell-specific connections in AVG data, we confirmed that there was strong connectivity (0.91 and 0.93) between vemurafenib and the two MEK inhibitors in A375 cells and weak connectivity in all of the other cell lines (Figure 6C). The presence of these expected connections in all three assays indicates that inhibition of BRAF and inhibition of MEK appear similar to each other in all profiling modalities: chromatin, phosphoproteomic, and mRNA changes.

A cell-specific connection even more striking in its specificity than the previous example was that of SCH 900776 (SCH) to SP600125 (SP) in MCF7 (Figure 6D). SCH is a cell-cycle inhibitor that targets CHEK1 (Bridges et al., 2016), while SP is a JNK inhibitor (discussed above). These compounds have different MoAs so it was unclear why they should connect, and it was especially unclear why they should connect in only one cell line. Noting the similarity of this connection to that of vemurafenib and MEK inhibitors in A375, which connect to each other in A375 because of differential cytotoxicity in A375 cells, we hypothesized that SCH connected to SP in MCF7 because these compounds were differentially cytotoxic in MCF7 cells.

We investigated our hypothesis with a follow-up viability experiment (Figure 6E). To account for variable cell growth rates, we quantified drug cytotoxicity using GR_{50} rather than IC_{50} (Hafner et al., 2016). The GR (growth rate) value represents cell count relative to DMSO control, adjusted for the cell line's growth rate. We found that SCH was 13.8-fold more cytotoxic in MCF7 (GR_{50}: 1.03 μM) than in A375 (14.2 μM), and SP was 1.7-fold more cytotoxic in MCF7 (GR_{50}: 14.7 μM) than in A375 (24.9 μM). For comparison, the GR_{50}s of vemurafenib and selumetinib in A375 were 0.58 μM and 0.31 μM, which are comparable with the GR_{50} of SCH in MCF7 (1.03 μM; see Supplemental Information). These results confirmed our hypothesis that MCF7 was differentially sensitive to SCH and, to a lesser extent, SP. Furthermore, they suggest that these inhibitors might be candidate therapeutics for cancers similar to MCF7 in lineage (i.e., breast) or genetic features (e.g., estrogen receptor positive). Indeed, CHEK1 inhibition is a promising therapeutic direction in the treatment of breast and ovarian cancers (Bryant et al., 2014), and JNK signaling has an important, but poorly understood, role in breast cancer (Ashenden et al., 2017).

One modification to our original hypothesis is that CC-401, like SP, turned out to be more cytotoxic in MCF7 than in A375 cells. CC-401 was 2.1 times more cytotoxic in MCF7 (GR_{50}: 16.1 μM) than in A375 (GR_{50}: 34.4 μM). By our connectivity results alone, we might have predicted that only SP, and not CC-401, would be differentially sensitive in MCF7. One explanation for this discrepancy is that the dose of CC-401 (5 μM) was insufficient for perturbing cells in the manner that caused SP (25 μM) to connect to SCH. By including CC-401 in our follow-up experiment, we provided evidence that JNK inhibitors in general, rather than SP in particular, might be effective therapeutics for MCF7-like cancers. This anecdote highlights the caveat of any high-throughput hypothesis generation resource: while attempts are made to optimize as many experimental parameters as possible, follow-up experiments are indispensable for refining and validating therapeutic hypotheses.

Proteomic Connectivity Links Genetics to Function and Identifies Potential Therapeutic Avenues

One powerful application of a library resource of systematic perturbational signatures is the ability to query existing, new, or even computationally derived phosphoproteomic or chromatin profiling data. To illustrate this application, we revisited data from a previously published chromatin profiling study that demonstrated how cancer-associated genetic alterations resulted in unique chromatin signatures (Jaffe et al., 2013). The Cancer Cell Line Encyclopedia (CCLE) is a collection of more than 900 diverse cancer lines representing

different tissues and sites of origin, and encompasses a wide variety of cancer driver mutations and downstream genetic dependencies (Barretina et al., 2012). In our previous work, we identified clusters of chromatin signatures related to gain- and loss-of-function of histone lysine methyltransferases. We set out to query these signatures against our library of systematic drug perturbations to correlate genetic features with drug activities.

To test this strategy, we first queried the chromatin profiles of cell lines belonging to a cluster identified as bearing EZH2 loss-of-function mutations in Jaffe et al. (2013) (Figure 7A). EZH2 is a histone H3 lysine 27 (H3K27) methyltransferase capable of catalyzing the addition of methyl groups up to the fully saturated trimethyl state (Cao et al., 2002). Additionally, it is the core catalytic member of the Polycomb Repressive Complex 2 (PRC2), whose activity has been associated with repression of genes and marking of inactive chromatin (Kuzmichev et al., 2002). The signatures of the five cell lines in Figure 7A are characterized by the loss of H3K27me1, me2, and me3, with a concomitant increase in H3K27me0 and ac1. When these signatures are queried against our library of drug perturbations and ranked by their median connectivity scores, the top 18 positive connections are against EZH2 inhibitor compounds in the six distinct cell backgrounds present. This is, remarkably, the complete set of EZH2 inhibitor signatures present in the library.

We adapted the technique of single-sample gene set enrichment analysis (GSEA, or GSEA-preranked) to look for enrichment of MoAs in the query results (Barbie et al., 2009; Subramanian et al., 2005). Using our previously assigned MoA annotations, we tested for enrichment of each MoA in the rank-ordered connectivity query results. As a proof of concept, the most highly enriched MoA class for connections to EZH2 loss-of-function mutant lines was the EZH2 inhibitor class (Figure 7B, false discovery rate [FDR] = 0.0).

Another key finding by Jaffe et al. (2013) was the recognition that two classes of genetic events—*t4;14* translocation and *NSD2:p.Glu1099Lys* mutation—led to strikingly similar chromatin phenotypes based on the increase in H3K36 methylation levels (e.g., profiles from each class of genetic event did not resolve into distinct clusters but did segregate from other cell lines lacking these events). This observation led to the assignment of *NSD2:p.Glu1099Lys* as a gain-of-function mutation and a dependency in acute lymphoblastic leukemia (ALL) lines that harbored it. Building upon that previous work, we queried the chromatin signatures (obtained from a predecessor of the GCP assay) of *t4;14* and *NSD2:p.Glu1099Lys* cell lines from Jaffe et al. (2013) against our resource of drug perturbation-induced chromatin signatures. Surprisingly, the connectivity profiles were able to perfectly segregate the *t4;14* and *NSD2:p.Glu1099Lys* classes of cell lines into different clusters (Figure 7C) despite the apparent similarity of the chromatin signatures that had been previously observed (Figure S8). This example illustrates that subtle differences in underlying signatures can drive vastly different connectivities and also demonstrates that using the entire profile for comparative analyses can provide more discriminatory power than simply analyzing differentially expressed analytes or typical clustering methods on raw profiles.

To further understand the basis for stratification, we again tested for enrichment of MoA classes in the connectivity query results. This time, we focused on the perturbations that were the most negatively connected to the signatures of the CCLE lines under the rationale that finding MoAs able to effectively reverse the chromatin signature induced by a genetic dependency might be a good means of selectively killing or sensitizing the cells (Figure 7D). Our enrichment analysis indicated that the signatures of the *t4;14* translocated lines were highly anti-connected to the HDAC inhibitor class of perturbations while the *NSD2:p.Glu1099Lys* lines were anti-connected to BRD inhibitors, CDK inhibitors, and mammalian target of rapamycin (mTOR) inhibitors in our dataset (all at FDR < 0.05). These functional enrichments are highly consistent with investigational therapies for the specific cancer subtypes that typically harbor these mutation classes (see below).

The first class of genetic event, *t4;14* translocation, was composed entirely of multiple myeloma cell lines. The initial publication describing the CCLE (Barretina et al., 2012) included drug sensitivity for five of the six *t4;14* lines described here (KMS-28BM was not profiled) along with ~50% of all cell lines in the collection. One of the drugs tested in that study was panobinostat, a potent HDAC inhibitor. The panobinostat EC_{50} was in the 12th percentile or better for all of the *t4;14* multiple myeloma lines tested, and notably the EC_{50} of panobinostat against KMS26 was less than the first percentile at 3.7 nM. The insight provided by the negative connection of these lines to HDAC inhibitors is further supported by the fact that HDAC inhibitors have proved to be potent elements in the treatment of multiple myeloma, especially in combination with other agents (Laubach et al., 2017; Steiner and Manasanch, 2017).

The second class, *NSD2:p.Glu1099Lys*, was composed of mostly acute lymphoblastic leukemia cell lines (a mixture of B-ALL and T-ALL, with the exceptions of SW579, a thyroid cancer line, and RL, a lymphoma line). Unfortunately, most of these lines were not profiled for drug sensitivity in the original CCLE study. However, several studies have recently emerged demonstrating the potential of mTOR inhibitors as therapeutics for ALL, especially in synergy with other agents (Iacovelli et al., 2015; Shi et al., 2016; Tasian et al., 2017; Witzig et al., 2015). One recent study details the synergy between an mTOR inhibitor and a CDK4/6 inhibitor in T-ALL, hitting two of the three classes of compounds that are anti-connected to the *NSD2:p.Glu1099Lys* set of lines. BRD inhibitors have also demonstrated therapeutic potential in B-ALL (Ott et al., 2012). The overlap among the mechanisms nominated by the proteomic connectivity analysis and these ongoing clinical studies is encouraging for the prospect of translational use of the proteomic resource.

Taken together, this evidence suggests that connectivity analysis might be helpful in suggesting potential therapeutic strategies based on anti-connections between genetic alterations and classes of drugs. It is not implausible that, if a disease induces a certain molecular state in a cell, a compound that drives the cell toward an opposite state might help contribute to a resolution of the disease state. Our literature analysis suggests that this “driving” might be most helpful as a potentiator of other therapeutics or a means of overcoming resistance.

DISCUSSION

The improvement in profiling technologies and throughput has enabled construction of systematic resources characterizing cellular responses to drugs (Iorio et al., 2016; Rohban et al., 2017; Subramanian et al., 2017). Historically, proteomic techniques have been difficult to apply to such resource generation due to the heterogeneous sampling nature of mass spectrometry-based techniques as well as the cost and time required. Yet proteins and dynamic changes thereto are important effectors in cells and are frequently the actual targets of drug therapies. Thus, it is valuable to create systematic resources with molecular readouts in the proteomic space to complement other techniques.

Here, we demonstrated the feasibility and value of creating a resource of drug perturbation proteomic signatures for signaling and epigenetics, two mechanistic classes being extensively developed for therapeutics. Many landmark proteomic studies focus on the depth of coverage at the expense of the number of samples profiled. We chose to focus on breadth rather than depth to generate the basis for an expandable resource. We created thousands of profiles from 90 different drugs in six biological model systems. To overcome the time, costs, and stochasticity normally associated with deep proteomics, we employed only compact, targeted assays that characterize high-value analytes. We demonstrated that signatures from these assays are reproducible and that drugs in general are apt to produce strong signals in these molecular spaces, regardless of known or assumed mechanism of action. Because these assays have been automated and standardized, we can continue to build this resource and ensure that new data are comparable with extant data.

In contrast to previously generated systematic resources of proteomic profiling data (Li et al., 2017; Mertins et al., 2016) that focused on experiments of nature (genetic variation), we sought to explore the promise of the proteomic dimension to understand drug mechanisms, optimize therapeutics under development, and recognize potential therapeutic indications based on the induction of desirable proteomic signatures. Our analyses show concrete initial examples of all three of these use cases. First, the cell-specific activities of vemurafenib (known) and SCH 900776 (novel) are revealed through connectivity (Figures 6C–6E). Second, we show the off-target effects of the JAK2 inhibitor TG101348 (also known as fedratinib) crossing over into bromodomain inhibitor space (Figure 6B). Clinical trials of fedratinib were halted by the Food and Drug Administration in 2013 due to toxicity (Grogan, 2013). Perhaps our observations may shed light on the mechanism as fedratinib re-emerges into clinical development (Carroll, 2017). Finally, we show that genetic states can be queried against our resource for potential therapeutic indications in the case of NSD2 mutation classes (Figure 7). Of note, there are over 80 active clinical trials for HDAC inhibitors in multiple myeloma, and 54 active clinical trials for bromodomain, cyclin-dependent kinase, and mTOR inhibitors in acute lymphoblastic leukemia (Citeline clinical trials database, https://citeline.informa.com). The potential efficacy of these drugs for these disease indications is predicted by our analyses.

A key motivation was to use assays that directly report on cellular biochemical processes likely to be directly affected by drugs. One could argue that space of the P100 phosphosignaling assay is relatively small compared with the number of known

phosphosites in existence, and that little is known about the biological function of specific analytes in the assay. This makes it somewhat difficult to draw biological conclusions from the primary observations of changes in specific phosphopeptides (i.e., we know little about the regulation of AHNAK S3426ph but can make some strong speculations), as opposed to the GCP assay, for which there is general knowledge about repressive or activating tendencies of chromatin marks. However, the data-driven approach with which the analytes were selected has demonstrated strong reporting activity on a wide variety of drug perturbations. By intentionally neither selecting nor excluding sentinel markers from known pathways, we sought to avoid bias. We also note that, while drugs and drug MoAs have strong replicate reproducibility in P100, changes in most individual sites are relatively subtle, rarely exceeding 3-fold from median levels (Figures 1E and 1G). Therefore, we believe that the relevant unit of comparison is the entire profile of a drug perturbation with that of other perturbations, rather than looking at the activity of specific markers within that profile.

Creating functional readouts in multiple biological activity spaces is a core goal of the LINCS program (Keenan et al., 2017), under which this resource was developed. Having distinct assays created a challenge of data integration. To address this challenge, we applied the concept of connectivity (Lamb et al., 2006)—principled comparison of whole molecular profiles in large datasets to recognize different biological conditions that induce related biochemical states—to our proteomic data. Connectivity allows a uniform way of asking the questions, such as: Given molecular profile data induced by a drug treatment, what else does this look like? Have we seen something like this before?

Importantly, our adaptation of connectivity is independent of assay type, which allows for facile integration of data from discrete assays by asking whether the same connections are found in each. Here, we used the exact same framework to integrate transcriptional profiling data obtained for the same experimental conditions. The proteomic assays perform on par with or better than other omics readouts (i.e., L1000) in terms of replicate reproducibility and MoA class detection (Figure S6). Yet we found it surprising that only a small number of connections consistently rise to the top in all three activities assayed (Figure 5). This finding supports the notion that there is no “one size fits all” approach to characterizing cellular responses (i.e., by only measuring transcriptional profiles), and that different assays provide unique windows into responses to different classes of perturbagens. For example, our chromatin assay is particularly well suited to characterizing responses to compounds with epigenetic targets, while the signaling assay may be blind to effects of these drugs. Yet we do not discount the complementary information offered by assays that can perhaps unexpectedly provide evidence of connectivity despite not directly assaying a drug’s annotated MoA (e.g., BRD inhibitors have stronger intra-class connectivity in signaling than they do in epigenetics, Figure 3). These observations underscore the need for multiple readout activities to be considered for understanding biological systems and their responses to new milieux.

At the same time, the repeated observation of the same connection(s) among all three assay types discussed here attaches a special importance to the result. There is a tendency in biological research to attempt to generalize results over all of biology, yet at the same time

to seek model-specific methods of manipulating systems (e.g., labeling a drug as a JAK2 inhibitor, yet searching for indications where a JAK2 inhibitor will selectively kill cells driven by aberrant JAK2 signaling activity). The framework of connectivity can help to distinguish the cases of true universality of response (i.e., Figure 6B, where connections are aggregated across assays and cell types) compared with cell type-specific responses (Figures 6C and 6D). The cell type-specific agreement of connections over three assays rediscovers the mechanism of a great recent advance in targeted cancer therapeutics: that a $BRAF^{V600E}$ mutation (present in A375) confers sensitivity to the drug vemurafenib by mimicking the disruption of MEK activity. The connection of this selective BRAF inhibitor to the MEK inhibitors was observed only in A375 (Figure 6C) despite the universality of the MEK inhibition response itself (Figure 6B). We hope that many more such examples can be discovered as we continue to analyze and expand our dataset.

The data from this study are available in a variety of forms (see Key Resources Table and STAR Methods for details), including web apps that enable quick access to connectivity data and tutorials to illustrate common use cases. We emphasize that this resource is the foundation for a systematic repository of proteomic signatures useful for drug characterization and more. Through LINCS, we are committed to its maintenance and expansion. The resource is expandable in many dimensions; for example, by generating profiles for more drugs, genetic perturbations and other stimuli, and profiling more cell types or models of biology (including tissue or patient-derived induced pluripotent stem cell models). All of these activities are already under way in our LINCS Proteomic Characterization Center for Signaling and Epigenetics. Beyond the analysis presented here, we have already released primary signature data for genetic knockdowns by CRISPR, time-course experiments, and drug responses in cardiovascular cell models via our Panorama repository (Key Resources Table).

There are other creative ways to increase the impact of the resource. Future developments in assay technologies might allow for sample multiplexing, raising the rate of resource expansion. More comprehensive mass spectrometry technologies (data-independent acquisition, or “DIA”) might readily expand the number of analytes covered in targeted proteomics assays, assessing biological pathway modulation directly from primary data. And we could also harvest data from other public studies that are likely to contain quantitative information on analytes covered by our assays (e.g., Bhanu et al., 2016; Kulej et al., 2015; Luense et al., 2016; Sidoli et al., 2015) for direct comparison with and integration with our data. All analytes in our assays were selected, in part, for their relatively universal observability, making data harvesting a realistic possibility. One could also envision expanding the resource in the direction of new assay panels with complementary readouts for understanding the protein dynamics of disease including ubiquitination, arginine methylation, and non-histone acetylation or methylation.

Finally, we envision collaborative interactions with users of this community resource, especially those interested in the characterization and development of therapeutics, thus influencing its future directions. We hope to have demonstrated the value of adding systematic proteomic studies to drug characterization and development efforts, and that the use of connectivity is a powerful abstraction that enables integration across data types.

Ultimately, the value of this resource can only be validated by the discoveries and insights that it enables in the future, but we are enthusiastic about the potential applications of this resource at present.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Cancer Cell Lines
  - Neural Progenitor Cells (NPCs)
- METHOD DETAILS
  - Drug Treatment
  - P100 Assay
  - GCP Assay
  - L1000 Assay
  - Data Processing for GCP and P100
  - Viability Follow-up Experiment, Related to Figure 6
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Determining Enrichment of Perturbation Sets, Related to Figures 1B–D, 7B, and 7D
  - Computing Similarities and Connectivities, Related to Figures 2 and S1B
  - Visualizing Connectivities, Related to Figure 2
  - Global Replicate and Non-Replicate Correlations, Related to Figures 3A and S2A–D
  - Replicate Reproducibility, Related to Figures 3B, S2E, and S6A
  - MoA Connectivity, Related to Figures 3C and S6B
  - Examples of MoA Connectivity, Related to Figure 3D
  - t-SNE Analysis, Related to Figures 4A and S4
  - Clustering and Dendrogram Cutting, Related to Figures 4C–D and S5
  - Percent Overlap Algorithm, Related to Figures 5B and S7A–C
  - Recall of Connectivity Profiles, Related to Figures 5C and S7D–E
  - Network Visualizations, Related to Figures 5D and 6B

STAR★METHODS

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Jacob D. Jaffe (jjaffe@broadinstitute.org).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Cancer Cell Lines—After thawing from −80°C storage, cancer cell lines were recovered in standard tissue culture-treated dishes. Cells were allowed to propagate at 37°C and 5% CO_2 for at least three doublings or until we had approximately 0.5 million cells per well of a six-well dish. A375 (female), A549 (male), and YAPC (male) cells were cultured in RPMI 1640 medium (Thermo Fisher). MCF7 cells (female) were cultured in DMEM (Thermo Fisher). PC3 cells (male) were cultured in RPMI 1640 medium containing 1 mM sodium pyruvate and 10 mM HEPES (Thermo Fisher).

DNA fingerprinting was used to authenticate the identity of cancer cell lines. Fingerprinting was performed at the Genomics Platform of the Broad Institute of MIT and Harvard (Cambridge, MA) using Fluidigm technology. The Fluidigm fingerprint panel includes a total of 96 SNPs, including 9 SNPs that overlap with the Affy 6.0 array and have multiple proxy SNPs each, 66 SNPs that overlap with Illumina's 1m and 2.5m arrays and have multiple proxy SNPs each, 32 SNPs in transcribed regions of housekeeping genes that are expressed in most cell types, and 1 gender determining SNP.

Neural Progenitor Cells (NPCs)—Individual colonies of H9 human embryonic stem cells (ESCs; female) were cultured with mTeSR1 media (StemCell Technology) in matrigel (BD Biosciences)-coated plates. For NPC induction, ESC cell colonies of 60–80% confluence were incubated in a 1:1 mixture of N-2 and B-27-containing media (see below) supplemented with 1 μm dorsomorphin (Tocris Bioscience) and 10 μm SB 431542 (Tocris Bioscience). ESCs differentiated to NPCs in a single passage, and NPCs were cultured for nine passages.

N-2 medium consisted of DMEM/F-12 GlutaMAX (Thermo Fisher), 1x N-2 supplement (Thermo Fisher), 5 μg/ml insulin (Sigma), 1 mM L-Glutamine (Thermo Fisher), 100 μm MEM nonessential amino acids solution (Thermo Fisher), 100 μM 2-Mercaptoethanol (Sigma), 50 U/ml penicillin and 50 mg/ml streptomycin (Thermo Fisher).

B-27 medium consisted of neurobasal medium (Thermo Fisher), 1x B-27 supplement (Thermo Fisher), 200 mM L-Glutamine (Thermo Fisher), 50 U/ml penicillin and 50 mg/ml streptomycin (Thermo Fisher).

METHOD DETAILS

Drug Treatment—Cells were plated into six-well dishes 24 hours before treatment with one of 90 drugs from Table S1. Media was changed the morning of treatment. Cells were treated for either 3 (P100), 6 (L1000), or 24 (GCP) hours at 37°C before undergoing one of the following assay protocols. For each drug, at least three biological replicates were performed.

In our previous description of the P100 assay (Abelin et al., 2016), we performed time course studies to establish three hours as an early meaningful time point, with subsequent deepening of the signatures at later time points. We chose not to exceed three hours for P100 because of the acute nature of the responses that we were hoping to capture, and to measure signaling changes that take place prior to wholesale transcriptional changes (6 to 24 hours). As for GCP, we wanted the opportunity for the cells to divide at least once, and most of our cells have a cycle time of ~24 hours. We acknowledge that there is no perfect time for collection of all samples, and more focused studies can be executed on alternative time scales. These times were chosen to achieve the broadest utility for our initial resource.

We chose treatment concentrations consistent with known bioavailability levels in humans where possible. Our algorithm for choosing treatment concentrations was to utilize public drug metabolism and pharmacokinetics (DMPK) and absorption, distribution, metabolism, and excretion (ADME) data to select the reported bioavailable concentrations of the drugs in serum. In the absence of such data, we consulted literature for EC_{50}/IC_{50} values or effective concentrations used in cellular studies. In the absence of prior knowledge or given conflicting evidence, we generally chose 1 μM as a default concentration (Table S1).

P100 Assay—Please see Abelin et al. for a detailed description of the P100 assay (Abelin et al., 2016). Briefly; drug-treated cells were washed with cold PBS, lysed in-plate with urea buffer, and harvested by scraping. Cell lysates were transferred to a 96-well plate, flash-frozen using liquid nitrogen, and stored at −80°C until further processing. Upon thawing,

samples (500 μg) were reduced, alkylated, and digested overnight with trypsin using the BRAVO Automated Liquid Handling Platform (Agilent). Peptides were desalted using a C18 Sep-Pak Cartridge (Waters) prior to immobilized metal affinity chromatography phosphopeptide enrichment. Phosphopeptides were enriched using commercially available Fe-NTA AssayMAP cartridges (Agilent). Salts were removed in a final desalting step using RPS cartridges (Agilent). A mix of isotopically labeled synthetic peptides was added to each sample prior to MS analysis. To ensure consistency of synthetic standard mixtures, a large batch of master mix was prepared using liquid handling automation, aliquoted, and stored at −80°C prior to the outset of these studies. Peptides were separated on a C18 column (EASY-nLC 1000, Thermo Scientific) and subsequently analyzed by mass spectrometry (MS) as described in Abelin et al., or in DIA mode (Q Exactive™-HF Orbitrap™, Thermo Scientific). In DIA, full scans were acquired in the 300–1200 m/z range at 60,000 FWHM resolving power followed by DIA scans spanning m/z 400–1000 at 30,000 FWHM resolving power, using a 22 m/z isolation window and a NCE of 27. Alternating traversal of the DIA m/z range had their center isolation m/zs offset by 50%. Overlapping DIA scans were deconvolved with Skyline (MacLean et al., 2010) and analyzed exactly as in Abelin et al.

GCP Assay—Please see Creech et al. for a detailed description of the GCP assay (Creech et al., 2015). Briefly; drug-treated cells were collected upon centrifugation. Upon lysis of the cells, histones were extracted using sulfuric acid and were precipitated using trichloroacetic acid. Samples (10 ug — a 5x reduction compared with Creech et al.) were propionylated, desalted, and digested overnight with trypsin. A second round of propionylation was employed and samples were subsequently desalted using C18 Sep-Pak Cartridge (Waters). A mix of isotopically labeled synthetic peptides was added to each sample prior to MS analysis. Peptides were separated on a C18 column (EASY-nLC 1000, Thermo Scientific) and analyzed by MS in a PRM mode (Q Exactive™-plus, Thermo Scientific) as described in Creech et al.

Detailed SOPs for P100 and GCP assays, including synthetic peptide master mixture formulation, can be found at https://panoramaweb.org/labkey/wiki/LINCS/Overview%20Information/page.view?name=sops.

L1000 Assay—Please see Subramanian et al. for a detailed description of the L1000 assay (Subramanian et al., 2017). Briefly; drug-treated cells were lysed using TCL Buffer (Qiagen), and mRNA transcripts were captured on oligo-dT-coated plates. Transcripts underwent ligation-mediated amplification (LMA); that is, mRNA was reverse transcribed to cDNA, gene and bead-specific probes were annealed to the cDNA, and probes were ligated and amplified via PCR using biotinylated universal primers. The PCR amplicon was then hybridized to beads with complementary oligonucleotide barcodes. After hybridization, the biotinylated amplicon was stained with streptavidin-phycoerythrin and detected using a Luminex FlexMap 3D system. The intensity of bead fluorescence corresponds to mRNA transcript abundance. Data were computationally processed using the standard Connectivity Map pipeline.

Data Processing for GCP and P100—Mass spectrometry data files were imported into Skyline software (MacLean et al., 2010) in order to extract and integrate MS2 signals for

endogenous (light) and synthetic internal standard (heavy) peptides. After manual quality assurance, Skyline files (.sky) were uploaded to a Panorama server (Sharma et al., 2014), where raw quantitative data were automatically assembled into Gene Cluster Text (GCT) files. The GCT format enables storing metadata and data in the same file. For each analyte (P100: n = 96, GCP: n = 61), the $\log_2$ ratio of the intensity of the endogenous peptide to the intensity of the internal standard peptide is reported. Samples are processed in batches of 96-well plates. We refer to this level of processing as level 2 data.

These data needed to be further normalized to enable comparison of samples within and across plates. The computational pipeline that performed these data processing steps is known as the Proteomics Signature Pipeline (PSP) and is available online at https://github.com/cmap/psp. Importantly, the pipeline is automatically executed by the Panorama server on uploaded Skyline documents so all data processing operations occur independently of human intervention, yielding a reproducible research pipeline. The pipeline is self-documenting by appending a record of processing operations as metadata to each sample.

First, any analytes or samples with an excess of missing data were filtered out. Samples with too few analytes (fewer than 80% of P100 analytes and fewer than 50% of GCP analytes) were rejected in order to make sure that a diversity of analytes were present for each sample. Analytes were also rejected on a plate-by-plate basis when they were measured in fewer than 90% of samples in P100 and 50% in GCP. These thresholds are included in the provenance code for each sample and consistent within an assay.

Next, a constant offset was applied to each sample in order to bring all samples to the same range. For GCP, we measure an analyte that is known to be invariant (histone H3, positions 41–49), so we subtracted this measurement from each sample. For P100, we do not have an invariant analyte, so we computed an analytical offset for each sample that minimized the distance between sample medians on a single plate. This is considered level 3 data.

Finally, we made our signatures differential by subtracting from each analyte its median across a plate. Therefore, each final value is the $\log_2$ ratio of endogenous to internal standard peptide, relative to the median analyte measurement across all samples on a 96-well plate. This is considered level 4 data.

We make our data publicly available at multiple levels of processing (see “Data and Software Availability” below).

Viability Follow-up Experiment, Related to Figure 6—A375 and MCF7 cells were thawed from −80°C storage, recovered in standard tissue culture-treated dishes, and allowed to propagate for at least three doublings at 37°C and 5% CO_2. Cells were then plated into 384-well dishes and, 24 hours later, treated with DMSO or one of the following drugs: vemurafenib, selumetinib, SCH 900776, SP600125, or CC-401. We performed an eight-point dose series in duplicate, where the maximum dose was 10 μm for vemurafenib and selumetinib and 100 μm for the other three drugs. Cells were treated for five days at 37°C.

We quantified cell viability by imaging for total cell counts (see Supplemental Information). Images of cells were acquired using an Opera Phenix High-Content Screening Instrument

(PerkinElmer) at 10x magnification in confocal mode, using brightfield and digital-phase contrast channels. Harmony software (PerkinElmer) was used to identify cells based on digital-phase contrast images and to count total cell numbers within four fields of view. Growth Rate (GR) values were computed, and sigmoid curves were fit using the online GR calculator (Hafner et al., 2016). The GR value represents cell count relative to DMSO control, adjusted for the cell line's growth rate.

QUANTIFICATION AND STATISTICAL ANALYSIS

Determining Enrichment of Perturbation Sets, Related to Figures 1B–D, 7B, and 7D—To determine whether a perturbation or a group of perturbations significantly modulated an analyte, we calculated the enrichment of that perturbation's or group of perturbations' profiles compared to all other perturbations ($n \approx 1{,}700$). Enrichment was assessed using the Benjamini-Hochberg False Discovery Rate (FDR) returned by the gene set enrichment analysis (GSEA) algorithm, but querying against perturbation sets rather than gene sets. Level 4 data were considered for Figures 1B–1D. Connectivity results were considered for Figure 7.

Computing Similarities and Connectivities, Related to Figures 2 and S1B—We utilize the framework of connectivity in order to compare perturbations across different profiling assays. We compute connectivity in two steps, and we do this separately for each cell line in each assay. First, we compute similarities among all profiles. We chose to use Spearman correlation, but other similarity metrics (e.g. Pearson correlation or Euclidean distance) could be substituted. Another similarity metric that we could have utilized is GSEA, where one signature is compared to others by reducing it to up and down gene sets and asking for the enrichment of these gene sets in the other signatures. We chose to use correlation because we wanted to utilize all of our analytes. GSEA heavily weights the extreme tails of a signature, while Spearman correlation does not. Since correlation is sensitive to the length of vectors being correlated and our assays produce vectors (i.e. signatures) with different lengths, we cannot directly compare correlations across assays. For example, the correlations between L1000 signatures, whose lengths are 978, are much smaller than the correlations between GCP signatures, whose lengths are 61.

Therefore, we next convert correlations to connectivity scores by comparing observed correlations to a background distribution of correlations. The background distribution consists of all correlations between one of two drugs being compared and all other drugs profiled. Let us consider an example.

To compute the connectivity between two compounds A and B with three replicates each, we compare the correlations between the replicates of A and the replicates of B (test distribution; $n = 9$) to the correlations of the replicates of B with all other samples (background distribution; $n \approx 270$). We use the two-sample Kolmogorov-Smirnov (KS) test to compare our two distributions. We opted for the non-parametric KS test because the assumptions of most parametric tests (e.g. normality of test distributions) are typically not satisfied by our distributions of correlation coefficients. The test statistic D of the KS test is what we call the connectivity score, except that we artificially add a sign to D in the

following way: if the median of the test distribution is less than the median of the background distribution, the connectivity score becomes negative. Therefore, the range of connectivity scores is −1 (strong negative connection) to 1 (strong positive connection). When the test distribution consists of only one number (e.g. if we are comparing two drugs with only one replicate each), then the test statistic D simply becomes the fractional rank of a single correlation against the background distribution.

An important consequence of connectivity is the ability to integrate data across assays. Because our assays measure different analytes, we are unable to directly compare perturbations' signatures. Instead, we compare their connectivities. In other words, we ask "does drug X have the same connections to other drugs in GCP as it does in P100?" This framework of connectivity allows us to quantitatively compare perturbations both within an assay and across assays, and connectivity scores have the same range regardless of assay.

Visualizing Connectivities, Related to Figure 2—Connectivity scores can be visualized using heatmaps or network views.

Heatmaps can be manipulated as any other form of matrix data. For example, one can visualize the strongest connections to an individual compound by sorting a single column. One can visualize the strongest connections to a set of compounds (e.g. drugs corresponding to the same MoA) by summarizing their individual connectivities (e.g. by median) and sorting by this summary value. Complete matrices of proteomic connectivity data can be browsed and interactively manipulated at https://clue.io/proteomics.

Network views are directed or undirected graphs representing perturbations as nodes and connectivity scores above a user-defined threshold as edges. Unlike heatmaps, network views make it easy to see different modules of biology at a glance. Automated methods for generating network views and input files suitable for Cytoscape (Shannon et al., 2003) visualization are provided as part of the Proteomics Signature Pipeline.

Global Replicate and Non-Replicate Correlations, Related to Figures 3A and S2A–D—For each cell line separately, we extracted all correlations among replicates, excluding the correlation between a sample and itself (because this value is always 1), and among non-replicates. We combined all cell line-specific values to get the global replicate and non-replicate distributions. We used kernel density estimation (with the kdeplot command in the seaborn package for Python) to smooth these distributions.

For Figure 3A, we used level 4 (differential) data. For Figures S2A and S2B, we used level 2 data ($\log_2$ transformed heavy-to-light analyte ratios). For Figures S2C and S2D, we used level 3 (quality-controlled and normalized) data. See "Data Processing for GCP and P100" above for more information about our data levels.

Replicate Reproducibility, Related to Figures 3B, S2E, and S6A—In order to determine whether the replicates of a drug treatment were reproducible, we compared its pairwise replicate correlations to a permutation null. A similar procedure is also described in (Smith et al., 2017). For a given drug X with k replicates, we extracted the pairwise

Spearman correlations ($n = \frac{k * (k-1)}{2}$ correlations per drug) among its replicates and aggregated these values using the median. Next, we created a permutation null by randomly sampling k profiles at random from the ≈ 270 profiles in a given cell line-assay condition, computing the median of their Spearman correlations, and repeating this 1,000 times in order to create a null distribution containing 1,000 values. Each drug was assigned a *p*-value, which was defined as the fraction of values in the null distribution that were greater than the observed median correlation among replicates. For example, a *p*-value of 0.01 for a particular drug means that the replicates of that drug had a higher median pairwise correlation than did 99% of randomly selected, size-matched sets of samples from the same assay and cell line. Finally, we corrected for multiple hypothesis testing by adjusting *p*-values to *q*-values using the Benjamini-Hochberg approach. Here, we considered drugs with *q*-values less than 0.05 to be reproducible.

Generating the permutation null is non-deterministic. To determine the stability of generating the permutation null, we reran this algorithm 10 times for each drug-cell type combination. The error bars represent the 5th and 95th percentiles of the 10 runs.

MoA Connectivity, Related to Figures 3C and S6B—In each cell line, we computed the median of connectivities among compounds belonging to the same mechanism of action (MoA). Singleton MoAs were discarded. Each bar shows the six median connectivities (one for each cell line) for a particular MoA. The height of each bar is the median of these six values, and error bars represent the 25th and 75th percentiles of these six values.

Examples of MoA Connectivity, Related to Figure 3D—Each value in these matrices is a median of the six connectivity scores corresponding to the six cell lines.

t-SNE Analysis, Related to Figures 4A and S4—t-SNE was used to generate two-dimensional projections of our datasets (van der Maaten and Hinton, 2008). The input to t-SNE was either level 4 matrices (Figures S4B and S4C) or connectivity matrices (Figures 4A and S4A). t-SNE was performed using Morpheus (see below) with perplexity = 60 and learning rate = 10. Projections were rendered in R using ggplot2.

Clustering and Dendrogram Cutting, Related to Figures 4C–D and S5—A distance matrix was computed using the “correlation” metric of the Dist function in the R package “amap” for the connectivity profiles for each cell line (Supplemental Information, n = 90 per assay per cell type). This matrix was used to create dendrograms using the hclust function with “average” linkage. The function cutree was used to identify connectivity clusters at fixed percentages of the maximal heights of the resulting dendrogram. Bootstrap analysis to assess cluster stability was performed using the “approximately unbiased” method contained in the R package ‘pvclust’ (Figure S5, Suzuki and Shimodaira, 2006). The average bootstrap value was computed as the mean of bootstraps for branch points above and below the 60% cut line. For Figure 4D, connectivity profiles were limited to the set of reproducible drugs (see “Replicate Reproducibility” above) in each cell line, dendrograms were cut at 60% of their maximal height, and clusters with single members were eliminated. A network graph was constructed where nodes were considered individual connectivity

clusters in P100 or GCP data, and edges were drawn between them with weight equal to the number of drugs shared between clusters. Network graphs were rendered using the “sankey” package of Google visualizations.

Percent Overlap Algorithm, Related to Figures 5B and S7A–C—For each connectivity matrix, the top 5% of connections were extracted, except for self-connections. Self-connections were excluded since they are another way of quantifying replicate reproducibility, which was already addressed by previous analyses. Next, the percent overlap of these lists of connections were computed. For example, if the top 5% (n = 100) of connections in {assay A, cell line C} had 12 connections in common with the top 5% (n = 100) of connections in {assay B, cell line C}, then the percent overlap was 12%. This computation was performed for each cell line separately as well as for a matrix in which cell lines were aggregated by computing the median of six cell-specific connectivity scores. For the cell-specific results, the height of the bar shows the mean, and the error bars represent the 95% confidence intervals. This algorithm was repeated using 1% and 10% cutoffs for Figures S7A and S7B. For Figure S7C, the percent overlap was computed as the joint intersection across all 3 assays divided by the number of connections chosen.

The null percent overlap corresponds to picking the top P% connections at random. For example, if P = 5, a random 5% of connections from one assay compared to a random 5% of connections from another assay would be expected to have an overlap of 0.25% (of all connections), which, when normalized to the portion of connections considered, is 0.25% / 5% = 0.05 or 5%. This analytical null was computationally confirmed by repeating the random sampling many times in all cell lines.

Recall of Connectivity Profiles, Related to Figures 5C and S7D–E—In order to assess whether a drug had the same pattern of connectivity in different assays, we compared connectivity profiles across assays. The connectivity profile for a given drug X is the vector of connectivity scores between drug X and all other drugs in a particular assay and cell line. We compared the connectivity profiles for the same drug in two different assays using Spearman correlation originally, but we found that considering only the tails of the connectivity profiles improved the similarity between connectivity profiles (Figure S7D). In order to look only at the tails of the connectivity profiles, we used a weighted enrichment score (Subramanian et al., 2005), instead of Spearman correlation.

We asked whether the connectivity profile of drug X in assay A could recall its matching connectivity profile in assay B by computing a metric called recall (see Figure S7E). We defined the recall for a given drug X as the fractional rank of the similarity between its matching connectivity profiles compared to its similarities with all other connectivity profiles. For example, recall of 0.95 for drug X means that its connectivity profile in {assay A, cell line C} was more similar to its matching connectivity profile in {assay B, cell line C} than it was to 95% of all other connectivity profiles. A technical note is that the rank differs depending on the directionality of assay comparisons, so the recall for a particular drug is the mean of two ranks (ranking into assay A and ranking into assay B).

This computation was performed for each cell line separately as well as for a matrix in which cell lines were aggregated by computing the median of six cell-specific connectivity scores. For the cell-specific results, the height of the bar shows the mean, and the error bars represent the 95% confidence intervals.

Network Visualizations, Related to Figures 5D and 6B—In each network, a node is a drug. An edge is shown if the connectivity score it represents exceeds the user-defined threshold (e.g. top 5% of connections in each assay). Only positive connectivity scores are considered. Only nodes connected by an edge are included in the networks. The distance between nodes is arbitrary.

Viability Curves, Related to Figure 6E—Growth Rate (GR) values were computed, and sigmoid curves were fit using the online GR calculator (Hafner et al., 2016). Individual points are biological replicates. A dashed line is shown at GR = 0.5 to make it easier to identify the GR_{50} value for each curve.

Querying External Data against Our Resource, Related to Figure 7 and Figure S8—Using our framework of connectivity and the Proteomics Signature Pipeline, we computed connectivity scores between all samples in our resource and the GCP profiles of 115 untreated cancer cell lines from the Cancer Cell Line Encyclopedia (CCLE) (Jaffe et al., 2013). This produced a matrix with 539 rows, representing 90 drugs in six cell lines (one drug had been excluded because of low technical quality), and 115 columns representing 115 cell lines from CCLE. For panel A, we extracted CCLE cell lines with EZH2 mutations (n = 5) and sorted by the median connectivity score to these five cell lines. For panels B and D, we assessed enrichment of MoA classes in the 539 rows of drugs using GSEA (preranked by median connectivity across all samples belonging to the genetic class: n = 5 for EZH2, n = 6 for t4;14,n =9 for $NSD2^{E1099K}$), with compound sets as our queries (as in Figures 1B–1D, see "Determining Enrichment of Perturbation Sets" above).

For panel C, we subsetted the connectivity matrix to cell lines with gain-of-function NSD2 mutations (n = 15), either through t4;14 translocation or the $NSD2^{E1099K}$ point mutation. We clustered the columns of this subsetted matrix using Spearman correlation and the average linkage method. The two types of NSD2 mutations segregated perfectly.

Figure S8 shows clustering of the NSD2-mutated cell lines based on their GCP profiles, rather than on their connectivity profiles. As with the clustering of connectivity profiles, we clustered using Spearman correlation as the distance metric and the average linkage method.

Comparing Connectivity to Correlation, Related to Figure S1B—To demonstrate that connectivity scores roughly track with Spearman correlations, we scattered our 97,015 connectivity scores (all pairwise drug comparisons within the same cell type, for both assays) against the mean of pairwise Spearman correlations between drug replicates. For example, if comparing drug A to drug B (each with 3 replicates), we computed the mean of the 9 pairwise correlations and compared that to the connectivity score previously computed. Because of extreme overplotting, we included a two-dimensional kernel density estimation

(using the kdeplot command in the seaborn package for Python) to demonstrate that the vast majority of points have low correlation and low connectivity.

Downsampling Analysis, Related to Figure S2E—In order to investigate how much the larger feature space of P100 explained its superior replicate reproducibility compared to GCP, we randomly downsampled the number of P100 analytes to 60 in order to approximately match that of GCP ($n = 59$) and reassessed replicate reproducibility as described above. To make our analysis exhaustive, we downsampled P100 to have N = 10, 20, …, 80, 90 analytes. Downsampling entailed randomly picking N analytes from the 96 available P100 analytes and subsetting the level 4 data matrix to just those N analytes. We did this 10 times for each analyte size N to generate error bars.

Computing Spearman Correlations of Control Compounds, Related to Figure S3A—Spearman correlations among replicates were computed first by considering only replicates on the same plate, then by considering replicates on any plate but within the same cell line, and finally by considering all replicates of the same compound, regardless of cell line. Level 4 data were used.

Computing Coefficients of Variation (CVs), Related to Figure S3B—CVs were computed first by considering only replicates of the same compound on the same plate, then by considering replicates on any plate but within the same cell line, and finally by considering all replicates of the same compound, regardless of cell line.

Since CVs must be computed on data with non-negative values, level 3 data after undoing the log2 transformation were used. Level 4 (differential) data could not be used since values are made relative to a plate-wide median and therefore negative. Finally, the GCP BI10052p peptide was removed from consideration because it is only used for quality-control purposes.

DATA AND SOFTWARE AVAILABILITY

Deposited Data—All Level 2 through Level 4 data (for GCP, P100, and L1000) are available via the Gene Expression Omnibus (accession GSE101406). Level 1 (Skyline files), Level 2, and Level 4 data for individual 96-well plates are available on Panorama Web (https://panoramaweb.org/labkey/LINCS.url).

Code for Manipulating a GCT file—Software for manipulating a GCT file, which is the standard file format for the data discussed in this manuscript, is available in a variety of programming languages (i.e. Java, Matlab, Python, and R). These repositories (cmapJ, cmapM, cmapPy, cmapR) can be found under the Connectivity Map team page on Github: https://github.com/cmap.

Proteomics Signature Pipeline—The software that processes proteomic data from level 2 to level 4, computes similarities and connectivities, and produces network visualizations of connectivity matrices is called the Proteomics Signature Pipeline (PSP). It is available at https://github.com/cmap/psp.

Proteomic Apps—In order to interact with the proteomic data discussed in this manuscript and additional data that will be released in the future, we have developed web applications for querying your own data against our resource and for exploring connections within our resource. The landing page for both of these apps is https://clue.io/proteomics.

Morpheus—Heatmaps were produced using a browser app called Morpheus (https://software.broadinstitute.org/morpheus).

Cytoscape—Networks were produced using Cytoscape software (http://cytoscape.org) (Shannon et al., 2003).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors would like to gratefully acknowledge Dr. Anne Carpenter, Dr. Max Macaluso, Dr. Rajiv Narayan, Dr. Jodi Hirschman, and Bang Wong for extremely helpful discussion on the manuscript. We would also like to thank Dr. Maria Kost-Alimova for assistance with image-based cell viability assays. This work was funded in part by the NIH Common Fund's Library of Integrated Network-based Cellular Signatures (LINCS) program by U54HG008097 (J.D.J.) and U54HG008699 (T.R.G. and A.S.).

References

Abelin JG, Patel J, Lu X, Feeney CM, Fagbami L, Creech AL, Hu R, Lam D, Davison D, Pino L, et al. Reduced-representation phosphosignatures measured by quantitative targeted ms capture cellular states and enable large-scale comparison of drug-induced phenotypes. Mol. Cell. Proteomics. 2016; 15:1622–1641. [PubMed: 26912667]

Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature. 2000; 403:503–511. [PubMed: 10676951]

Araf S, Okosun J, Koniali L, Fitzgibbon J, Heward J. Epigenetic dysregulation in follicular lymphoma. Epigenomics. 2016; 8:77–84. [PubMed: 26698557]

Ashenden M, van Weverwijk A, Murugaesu N, Fearns A, Campbell J, Gao Q, Iravani M, Isacke CM. An in vivo functional screen identifies JNK signaling as a modulator of chemotherapeutic response in breast cancer. Mol. Cancer Ther. 2017; 16:1967–1978. [PubMed: 28611109]

Aumann S, Abdel-Wahab O. Somatic alterations and dysregulation of epigenetic modifiers in cancers. Biochem. Biophys. Res. Commun. 2014; 455:24–34. [PubMed: 25111821]

Barbie DA, Tamayo P, Boehm JS, Kim SY, Moody SE, Dunn IF, Schinzel AC, Sandy P, Meylan E, Scholl C, et al. Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. Nature. 2009; 462:108–112. [PubMed: 19847166]

Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, Wilson CJ, Lehár J, Kryukov GV, Sonkin D, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature. 2012; 483:603–607. [PubMed: 22460905]

Baur JA, Sinclair DA. Therapeutic potential of resveratrol: the in vivo evidence. Nat. Rev. Drug Discov. 2006; 5:493–506. [PubMed: 16732220]

Beher D, Wu J, Cumine S, Kim KW, Lu S-C, Atangan L, Wang M. Resveratrol is not a direct activator of SIRT1 enzyme activity. Chem. Biol. Drug Des. 2009; 74:619–624. [PubMed: 19843076]

Bhanu NV, Sidoli S, Garcia BA. Histone modification profiling reveals differential signatures associated with human embryonic stem cell self-renewal and differentiation. Proteomics. 2016; 16:448–458. [PubMed: 26631989]

Bian Y, Song C, Cheng K, Dong M, Wang F, Huang J, Sun D, Wang L, Ye M, Zou H. An enzyme assisted RP-RPLC approach for in-depth analysis of human liver phosphoproteome. J. Proteomics. 2014; 96:253–262. [PubMed: 24275569]

Bittner M, Meltzer P, Chen Y, Jiang Y, Seftor E, Hendrix M, Radmacher M, Simon R, Yakhini Z, Ben-Dor A, et al. Molecular classification of cutaneous malignant melanoma by gene expression profiling. Nature. 2000; 406:536–540. [PubMed: 10952317]

Bridges KA, Chen X, Liu H, Rock C, Buchholz TA, Shumway SD, Skinner HD, Meyn RE. MK-8776, a novel chk1 kinase inhibitor, radiosensitizes p53-defective human tumor cells. Oncotarget. 2016; 7:71660–71672. [PubMed: 27690219]

Bryant C, Rawlinson R, Massey AJ. Chk1 inhibition as a novel therapeutic strategy for treating triple-negative breast and ovarian cancers. BMC Cancer. 2014; 14:570. [PubMed: 25104095]

Cao R, Wang L, Wang H, Xia L, Erdjument-Bromage H, Tempst P, Jones RS, Zhang Y. Role of histone H3 lysine 27 methylation in Polycomb-group silencing. Science. 2002; 298:1039–1043. [PubMed: 12351676]

Carroll J. John Hood bags $90M for his quest to get fedratinib OK'd by FDA and catapult into the market. Endpoints News. 2017

Ciceri P, Müller S, O'Mahony A, Fedorov O, Filippakopoulos P, Hunt JP, Lasater EA, Pallares G, Picaud S, Wells C, et al. Dual kinase-bromodomain inhibitors for rationally designed polypharmacology. Nat. Chem. Biol. 2014; 10:305–312. [PubMed: 24584101]

Clark EA, Golub TR, Lander ES, Hynes RO. Genomic analysis of metastasis reveals an essential role for RhoC. Nature. 2000; 406:532–535. [PubMed: 10952316]

Creech AL, Taylor JE, Maier VK, Wu X, Feeney CM, Udeshi ND, Peach SE, Boehm JS, Lee JT, Carr SA, et al. Building the Connectivity Map of epigenetics: chromatin profiling by quantitative targeted mass spectrometry. Methods. 2015; 72:57–64. [PubMed: 25448295]

Dawson MA, Kouzarides T, Huntly BJP. Targeting epigenetic readers in cancer. N. Engl. J. Med. 2012; 367:647–657. [PubMed: 22894577]

Fabian MA, Biggs WH, Treiber DK 3rd, Atteridge CE, Azimioara MD, Benedetti MG, Carter TA, Ciceri P, Edeen PT, Floyd M, et al. A small molecule-kinase interaction map for clinical kinase inhibitors. Nat. Biotechnol. 2005; 23:329–336. [PubMed: 15711537]

Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science. 1999; 286:531–537. [PubMed: 10521349]

Gräff J, Mansuy IM. Epigenetic dysregulation in cognitive disorders. Eur. J. Neurosci. 2009; 30:1–8. [PubMed: 19508697]

Greenbaum D, Colangelo C, Williams K, Gerstein M. Comparing protein abundance and mRNA expression levels on a genomic scale. Genome Biol. 2003; 4:117. [PubMed: 12952525]

Grogan K. Sanofi pulls plug on fedratinib over brain disorder risk. PharmaTimes. 2013

Gupta SC, Sung B, Prasad S, Webb LJ, Aggarwal BB. Cancer drug discovery by repurposing: teaching new tricks to old dogs. Trends Pharmacol. Sci. 2013; 34:508–517. [PubMed: 23928289]

Hafner M, Niepel M, Chung M, Sorger PK. Growth rate inhibition metrics correct for confounders in measuring sensitivity to cancer drugs. Nat. Methods. 2016; 13:521–527. [PubMed: 27135972]

Iacovelli S, Ricciardi MR, Allegretti M, Mirabilii S, Licchetta R, Bergamo P, Rinaldo C, Zeuner A, Foà R, Milella M, et al. Co-targeting of Bcl-2 and mTOR pathway triggers synergistic apoptosis in BH3 mimetics resistant acute lymphoblastic leukemia. Oncotarget. 2015; 6:32089–32103. [PubMed: 26392332]

Iorio F, Knijnenburg TA, Vis DJ, Bignell GR, Menden MP, Schubert M, Aben N, Gonçalves E, Barthorpe S, Lightfoot H, et al. A landscape of pharmacogenomic interactions in cancer. Cell. 2016; 166:740–754. [PubMed: 27397505]

Jaffe JD, Wang Y, Chan HM, Zhang J, Huether R, Kryukov GV, Bhang H-EC, Taylor JE, Hu M, Englund NP, et al. Global chromatin profiling reveals NSD2 mutations in pediatric acute lymphoblastic leukemia. Nat. Genet. 2013; 45:1386–1391. [PubMed: 24076604]

Jenuwein T, Allis CD. Translating the histone code. Science. 2001; 293:1074–1080. [PubMed: 11498575]

Keenan AB, Jenkins SL, Jagodnik KM, Koplev S, He E, Torre D, Wang Z, Dohlman AB, Silverstein MC, Lachmann A, et al. The Library of Integrated Network-Based Cellular Signatures NIH Program: system-level cataloging of human cells response to perturbations. Cell Syst. 2017; 6:13–24. [PubMed: 29199020]

Kelly TK, De Carvalho DD, Jones PA. Epigenetic modifications as therapeutic targets. Nat. Biotechnol. 2010; 28:1069–1078. [PubMed: 20944599]

Kulej K, Avgousti DC, Weitzman MD, Garcia BA. Characterization of histone post-translational modifications during virus infection using mass spectrometry-based proteomics. Methods. 2015; 90:8–20. [PubMed: 26093074]

Kuzmichev A, Nishioka K, Erdjument-Bromage H, Tempst P, Reinberg D. Histone methyltransferase activity associated with a human multiprotein complex containing the Enhancer of Zeste protein. Genes Dev. 2002; 16:2893–2905. [PubMed: 12435631]

Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, Lerner J, Brunet J-P, Subramanian A, Ross KN, et al. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. Science. 2006; 313:1929–1935. [PubMed: 17008526]

Lamb JA, Ventura J-J, Hess P, Flavell RA, Davis RJ. JunD mediates survival signaling by the JNK signal transduction pathway. Mol. Cell. 2003; 11:1479–1489. [PubMed: 12820962]

Laubach JP, San-Miguel JF, Hungria V, Hou J, Moreau P, Lonial S, Lee JH, Einsele H, Alsina M, Richardson PG. Deacetylase inhibitors: an advance in myeloma therapy? Expert Rev. Hematol. 2017; 10:229–237. [PubMed: 28076695]

Lee IH, Lim HJ, Yoon S, Seong JK, Bae DS, Rhee SG, Bae YS. Ahnak protein activates protein kinase C (PKC) through dissociation of the PKC-protein phosphatase 2A complex. J. Biol. Chem. 2008; 283:6312–6320. [PubMed: 18174170]

Li J, Zhao W, Akbani R, Liu W, Ju Z, Ling S, Vellano CP, Roebuck P, Yu Q, Eterovic AK, et al. Characterization of human cancer cell lines by reverse-phase protein arrays. Cancer Cell. 2017; 31:225–239. [PubMed: 28196595]

Luense LJ, Wang X, Schon SB, Weller AH, Lin Shiao E, Bryant JM, Bartolomei MS, Coutifaris C, Garcia BA, Berger SL. Comprehensive analysis of histone post-translational modifications in mouse and human male germ cells. Epigenetics Chromatin. 2016; 9:24. [PubMed: 27330565]

MacLean B, Tomazela DM, Shulman N, Chambers M, Finney GL, Frewen B, Kern R, Tabb DL, Liebler DC, MacCoss MJ. Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. Bioinformatics. 2010; 26:966–968. [PubMed: 20147306]

Martin C, Zhang Y. The diverse functions of histone lysine methylation. Nat. Rev. Mol. Cell Biol. 2005; 6:838–849. [PubMed: 16261189]

Mertins P, Mani DR, Ruggles KV, Gillette MA, Clauser KR, Wang P, Wang X, Qiao JW, Cao S, Petralia F, et al. Proteogenomics connects somatic mutations to signalling in breast cancer. Nature. 2016; 534:55–62. [PubMed: 27251275]

Ntziachristos P, Abdel-Wahab O, Aifantis I. Emerging concepts of epigenetic dysregulation in hematological malignancies. Nat. Immunol. 2016; 17:1016–1024. [PubMed: 27478938]

Ott CJ, Kopp N, Bird L, Paranal RM, Qi J, Bowman T, Rodig SJ, Kung AL, Bradner JE, Weinstock DM. BET bromodomain inhibition targets both c-Myc and IL7R in high-risk acute lymphoblastic leukemia. Blood. 2012; 120:2843–2852. [PubMed: 22904298]

Pacholec M, Bleasdale JE, Chrunyk B, Cunningham D, Flynn D, Garofalo RS, Griffith D, Griffor M, Loulakis P, Pabst B, et al. SRT1720, SRT2183, SRT1460, and resveratrol are not direct activators of SIRT1. J. Biol. Chem. 2010; 285:8340–8351. [PubMed: 20061378]

Parker JS, Mullins M, Cheang MCU, Leung S, Voduc D, Vickery T, Davies S, Fauron C, He X, Hu Z, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. J. Clin. Oncol. 2009; 27:1160–1167. [PubMed: 19204204]

Peck B, Chen C-Y, Ho K-K, Di Fruscia P, Myatt SS, Coombes RC, Fuchter MJ, Hsiao C-D, Lam EW. SIRT inhibitors induce cell death and p53 acetylation through targeting both SIRT1 and SIRT2. Mol. Cancer Ther. 2010; 9:844–855. [PubMed: 20371709]

Peña CJ, Bagot RC, Labonté B, Nestler EJ. Epigenetic signaling in psychiatric disorders. J. Mol. Biol. 2014; 426:3389–3412. [PubMed: 24709417]

Perou CM, Sørlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, et al. Molecular portraits of human breast tumours. Nature. 2000; 406:747–752. [PubMed: 10963602]

Qin W, Wolf P, Liu N, Link S, Smets M, La Mastra F, Forné I, Pichler G, Hörl D, Fellinger K, et al. DNA methylation requires a DNMT1 ubiquitin interacting motif (UIM) and histone ubiquitination. Cell Res. 2015; 25:911–929. [PubMed: 26065575]

Rohban MH, Singh S, Wu X, Berthet JB, Bray M-A, Shrestha Y, Varelas X, Boehm JS, Carpenter AE. Systematic morphological profiling of human gene and allele function via Cell Painting. Elife. 2017; 6 https://doi.org/10.7554/eLife.24060.

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003; 13:2498–2504. [PubMed: 14597658]

Sharma V, Eckels J, Taylor GK, Shulman NJ, Stergachis AB, Joyner SA, Yan P, Whiteaker JR, Halusa GN, Schilling B, et al. Panorama: a targeted proteomics knowledge base. J. Proteome Res. 2014; 13:4205–4210. [PubMed: 25102069]

Shi P-J, Xu L-H, Lin K-Y, Weng W-J, Fang J-P. Synergism between the mTOR inhibitor rapamycin and FAK down-regulation in the treatment of acute lymphoblastic leukemia. J. Hematol. Oncol. 2016; 9:12. [PubMed: 26892465]

Sidoli S, Lin S, Xiong L, Bhanu NV, Karch KR, Johansen E, Hunter C, Mollah S, Garcia BA. Sequential window acquisition of all theoretical mass spectra (SWATH) analysis for characterization and quantification of histone post-translational modifications. Mol. Cell. Proteomics. 2015; 14:2420–2428. [PubMed: 25636311]

Singhal S, Mehta J, Desikan R, Ayers D, Roberson P, Eddlemon P, Munshi N, Anaissie E, Wilson C, Dhodapkar M, et al. Antitumor activity of thalidomide in refractory multiple myeloma. N. Engl. J. Med. 1999; 341:1565–1571. [PubMed: 10564685]

Smith I, Greenside PG, Natoli T, Lahr DL, Wadden D, Tirosh I, Narayan R, Root DE, Golub TR, Subramanian A, et al. Evaluation of RNAi and CRISPR technologies by large-scale gene expression profiling in the Connectivity Map. PLoS Biol. 2017; 15:e2003213. [PubMed: 29190685]

Steiner RE, Manasanch EE. Carfilzomib boosted combination therapy for relapsed multiple myeloma. Onco. Targets Ther. 2017; 10:895–907. [PubMed: 28243125]

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. USA. 2005; 102:15545–15550. [PubMed: 16199517]

Subramanian A, Narayan R, Corsello SM, Peck DD, Natoli TE, Lu X, Gould J, Davis JF, Tubelli AA, Asiedu JK, et al. A next generation Connectivity Map: L1000 platform and the first 1,000,000 profiles. Cell. 2017; 171:1437–1452.e17. [PubMed: 29195078]

Suzuki R, Shimodaira H. Pvclust: an R package for assessing the uncertainty in hierarchical clustering. Bioinformatics. 2006; 22:1540–1542. [PubMed: 16595560]

Tasian SK, Teachey DT, Li Y, Shen F, Harvey RC, Chen I-M, Ryan T, Vincent TL, Willman CL, Perl AE, et al. Potent efficacy of combined PI3K/mTOR and JAK or ABL inhibition in murine xenograft models of Ph-like acute lymphoblastic leukemia. Blood. 2017; 129:177–187. [PubMed: 27777238]

van der Maaten L, Hinton G. Visualizing data using t-SNE. J. Mach. Learn. Res. 2008; 9:2579–2605.

Wagle N, Emery C, Berger MF, Davis MJ, Sawyer A, Pochanard P, Kehoe SM, Johannessen CM, Macconaill LE, Hahn WC, et al. Dissecting therapeutic resistance to RAF inhibition in melanoma by tumor genomic profiling. J. Clin. Oncol. 2011; 29:3085–3096. [PubMed: 21383288]

Wawer MJ, Li K, Gustafsdottir SM, Ljosa V, Bodycombe NE, Marton MA, Sokolnicki KL, Bray M-A, Kemp MM, Winchester E, et al. Toward performance-diverse small-molecule libraries for cell-based phenotypic screening using multiplexed high-dimensional profiling. Proc. Natl. Acad. Sci. USA. 2014; 111:10911–10916. [PubMed: 25024206]

Wendling D, Abbas W, Godfrin-Valnet M, Guillot X, Khan KA, Cedoz J-P, Baud L, Prati C, Herbein G. Resveratrol, a sirtuin 1 activator, increases IL-6 production by peripheral blood mononuclear cells of patients with knee osteoarthritis. Clin. Epigenetics. 2013; 5:10. [PubMed: 23844973]

Wernig G, Kharas MG, Okabe R, Moore SA, Leeman DS, Cullen DE, Gozo M, McDowell EP, Levine RL, Doukas J, et al. Efficacy of TG101348, a selective JAK2 inhibitor, in treatment of a murine model of JAK2V617F–induced polycythemia vera. Cancer Cell. 2008; 13:311–320. [PubMed: 18394554]

Whitesell L, Mimnaugh EG, De Costa B, Myers CE, Neckers LM. Inhibition of heat shock protein HSP90-pp60v–src heteroprotein complex formation by benzoquinone ansamycins: essential role for stress proteins in oncogenic transformation. Proc. Natl. Acad. Sci. USA. 1994; 91:8324–8328. [PubMed: 8078881]

Whitesell L, Bagatell R, Falsey R. The stress response: implications for the clinical development of hsp90 inhibitors. Curr. Cancer Drug Targets. 2003; 3:349–358. [PubMed: 14529386]

Witzig TE, Reeder C, Han JJ, LaPlant B, Stenson M, Tun HW, Macon W, Ansell SM, Habermann TM, Inwards DJ, et al. The mTORC1 inhibitor everolimus has antitumor activity in vitro and produces tumor responses in patients with relapsed T-cell lymphoma. Blood. 2015; 126:328–335. [PubMed: 25921059]

Wu P, Nielsen TE, Clausen MH. FDA-approved small-molecule kinase inhibitors. Trends Pharmacol. Sci. 2015; 36:422–439. [PubMed: 25975227]

Yazgan O, Pfarr CM. Regulation of two JunD isoforms by Jun N-terminal kinases. J. Biol. Chem. 2002; 277:29710–29718. [PubMed: 12052834]

Zhou H, Di Palma S, Preisinger C, Peng M, Polat AN, Heck AJR, Mohammed S. Toward a comprehensive characterization of a human cancer cell phosphoproteome. J. Proteome Res. 2013; 12:260–271. [PubMed: 23186163]

Highlights

- First-of-its-kind public resource of proteomic responses to systematic drug treatment
- Profiling of phosphosignaling and chromatin states induced by 90 drugs in 6 cell lines
- Extends Connectivity Map concept to proteomics and integrates with transcriptomics
- Enables recognition of cell type-specific activities and therapeutic opportunities

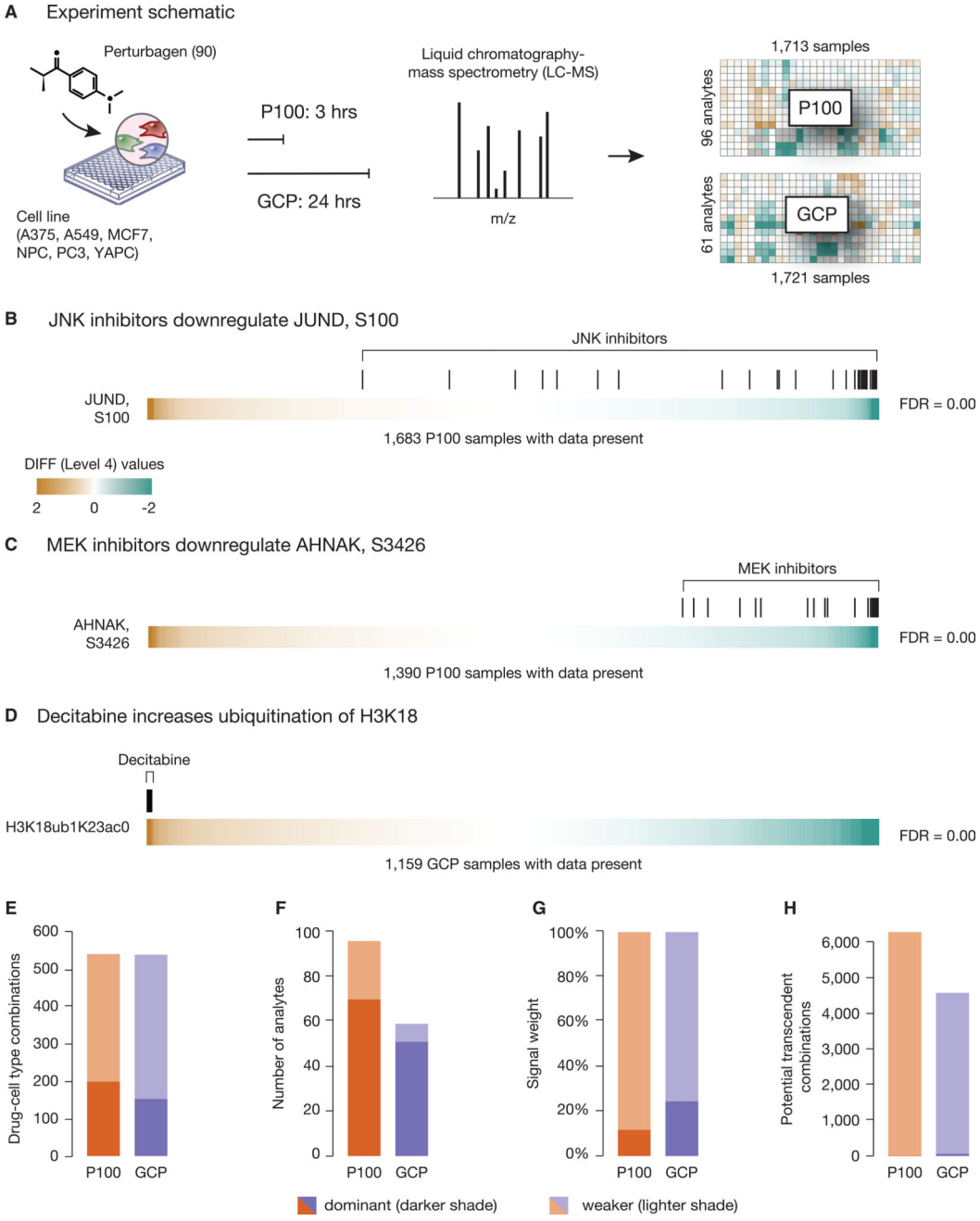


Figure 1. Analyte-Centric Analysis of the Resource

(A) Experiment schematic. Cells were treated with one of 90 small-molecule perturbagens with a minimum of three biological replicates. After 3 (P100) or 24 (GCP) hr of exposure to treatment, cells were lysed and profiled in each of the two assays. After quality control filtering, over 3,400 individual profiles constituted the resource.

(B) JNK inhibitors downregulate phosphorylation of S100 on JUND. Observed abundance of JUND, S100 phosphorylation across all P100 samples (level 4 data) is shown as a single horizontal row of a heatmap. JNK inhibitor profiles are marked with black ticks along the

top of the heatmap. Computing enrichment of the JNK inhibitors across the 1,683 samples with data present yields a GSEA FDR result of 0.00.

(C) MEK inhibitors downregulate phosphorylation of S3426 on AHNAK. Computing enrichment of the MEK inhibitor profiles across 1,390 P100 samples with data present yields an FDR of 0.00.

(D) Decitabine increases ubiquitination of lysine 18 on histone H3. Observed abundance of H3K18ub1K23ac0 across all GCP samples (level 4 data) is shown as a single horizontal row of a heatmap. Computing enrichment of decitabine profiles across 1,159 GCP samples with data present yields an FDR of 0.00.

(E) Number of drug-cell type conditions with at least one “dominant” analyte.

(F) Number of analytes dominant in at least one condition.

(G) Percentage of signal derived from dominant analytes (average of all profiles).

(H) Number of analyte-drug combinations that transcend cell type.

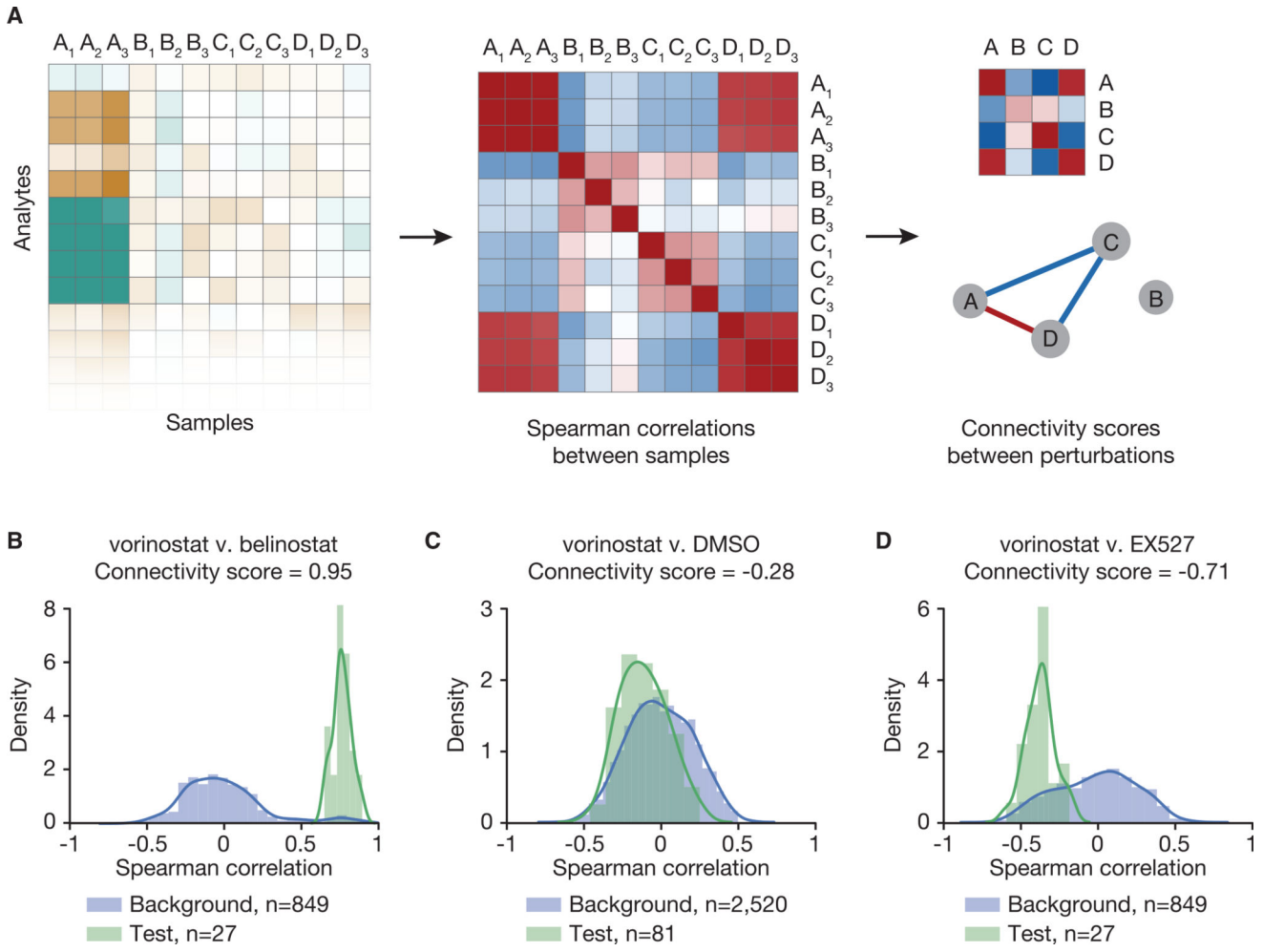


Figure 2. Connectivity Framework

(A) Each sample is represented as a profile of analyte measurements. Spearman correlations are computed between all profiles within a cell line. Finally, we compute connectivity scores by comparing the observed correlations with a background of correlations. Computing connectivity collapses replicates. Connectivity maps may be represented as matrices or networks. Different color scales are used for profile and similarity matrices to make them more distinguishable.

(B) Vorinostat versus belinostat: example of a positive connectivity score close to 1. The background distribution (blue) consists of the correlations between the replicates of belinostat and all other samples. The test distribution (green) consists of the correlations between the replicates of belinostat and the replicates of vorinostat.

(C) Vorinostat versus DMSO: example of a connectivity score close to 0. The background distribution (blue) consists of the correlations between the replicates of DMSO and all other samples. The test distribution (green) consists of the correlations between the replicates of DMSO and the replicates of vorinostat.

(D) Vorinostat versus EX527: example of a negative connectivity score close to −1. The background distribution (blue) consists of the correlations between the replicates of EX527

and all other samples. The test distribution (green) consists of the correlations between the replicates of EX527 and the replicates of vorinostat.
(B) to (D) show data for the GCP assay in A375 cells. See also Figure S1.

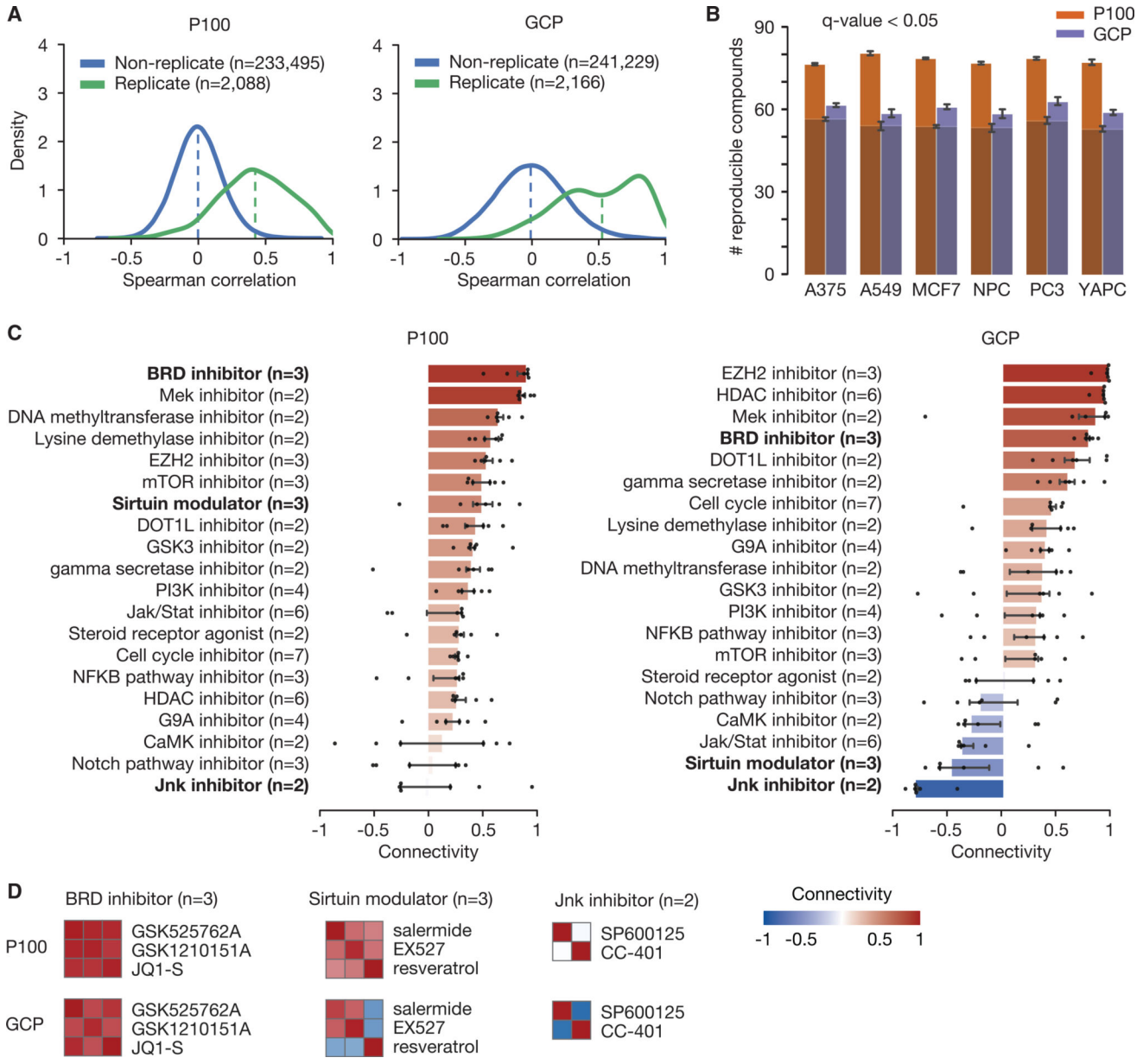


Figure 3. Majority of Compounds Are Reproducible in Both Assays, but Assays Show Different Sensitivities to MoA Classes

(A) Distributions of all Spearman correlations among replicates (green) and among non-replicates (blue).

(B) Bar chart showing the number of compounds considered reproducible in each cell line-assay combination. A compound was considered reproducible if its replicates were significantly correlated compared with a permutation null (q value < 0.05). The algorithm was rerun 10 times to generate error bars; the center is the median, error bars represent 95% confidence interval of 1,000 bootstrapped iterations. The shaded component indicates the overlap of reproducible compounds between GCP and P100.

(C) Bar charts showing the median connectivity of compounds annotated with the same mechanism of action (MoA). The center is the median, error bars represent 50% confidence interval of 1,000 bootstrapped iterations.
(D) Heatmaps showing connectivity among compounds belonging to the BRD inhibitor, sirtuin modulator, and JNK inhibitor MoA classes. The BRD inhibitor class shows high connectivity in both P100 and GCP; the sirtuin modulator class shows high connectivity in P100 but low connectivity in GCP; and the JNK inhibitor class shows low connectivity in both P100 and GCP. Each square is the median of six cell-specific connectivity scores. The labels of the matrices are symmetric; that is, the columns (left to right) have the same annotations as the rows (top to bottom).
Color scale applies to both (C) and (D). See also Figures S2 and S3.

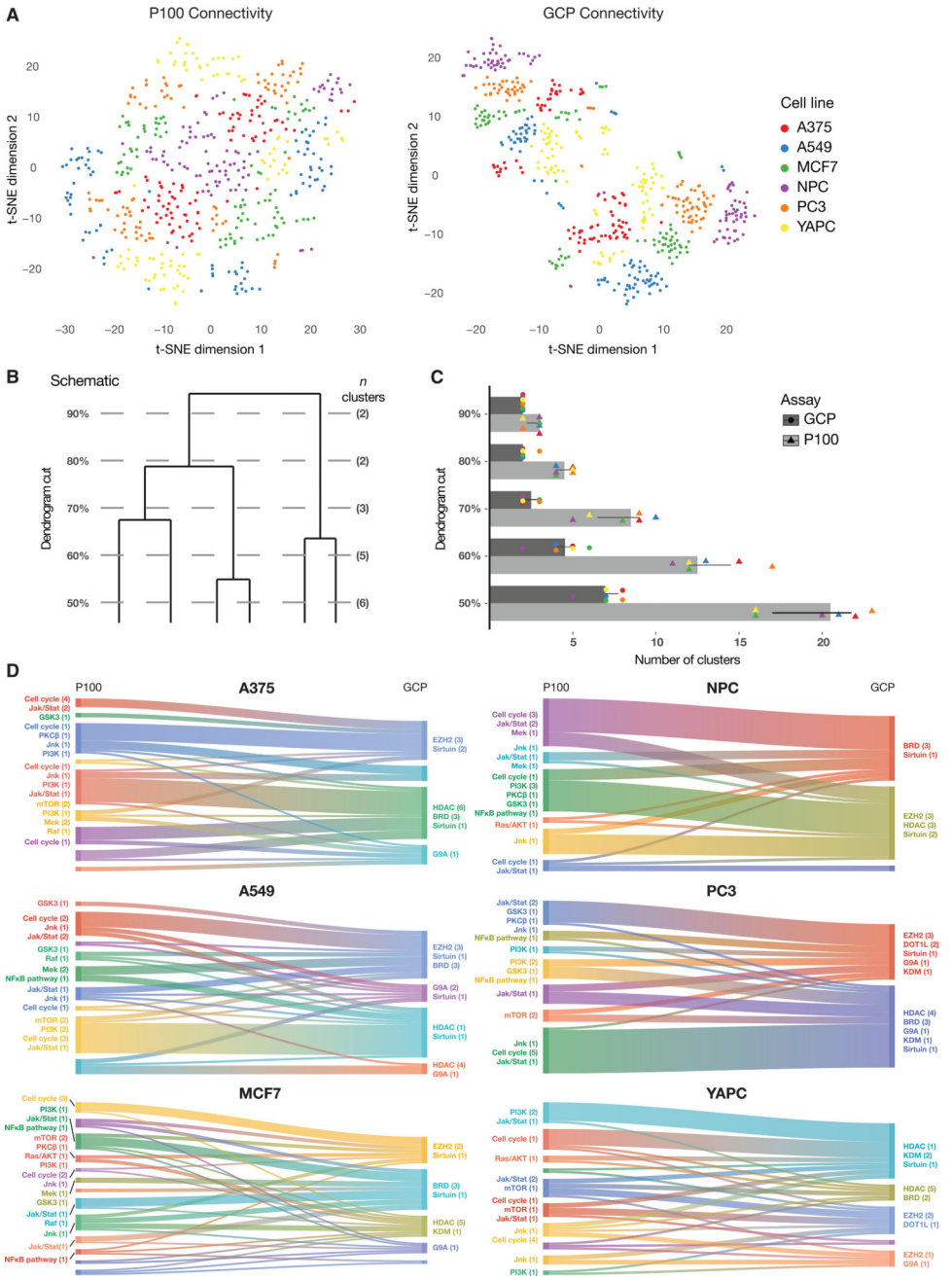


Figure 4. Connectivity Profile Analysis

(A) t-SNE projection of P100 and GCP connectivity profiles for connections within each cell type (distance metric = Pearson correlation, perplexity = 60, learning rate = 10).

(B) Schematic representation of cutting a dendrogram at a fixed percentage of its height and counting resulting clusters, for illustration only.

(C) Number of connectivity clusters formed as a result of cutting dendrograms as depicted in (B). Individual data points (six per assay) are overlaid on the box plots and jittered on the y axis for clarity. The center is the median, error bars represent the 25th and 75th percentiles.

(D) Connectivity flows from P100 connectivity clusters to GCP connectivity clusters. For each cell line, only compounds reproducible in both assays are included. Each cluster is annotated by the major pertinent mechanistic classes for each assay, with the number of drugs in each class shown in parentheses. Colors are arbitrary. Because the analysis was restricted to the reproducible compounds in each cell type and single member clusters were eliminated, the number of clusters at the 60% cut for P100 and GCP may differ slightly from (C).

See also Figures S4 and S5.

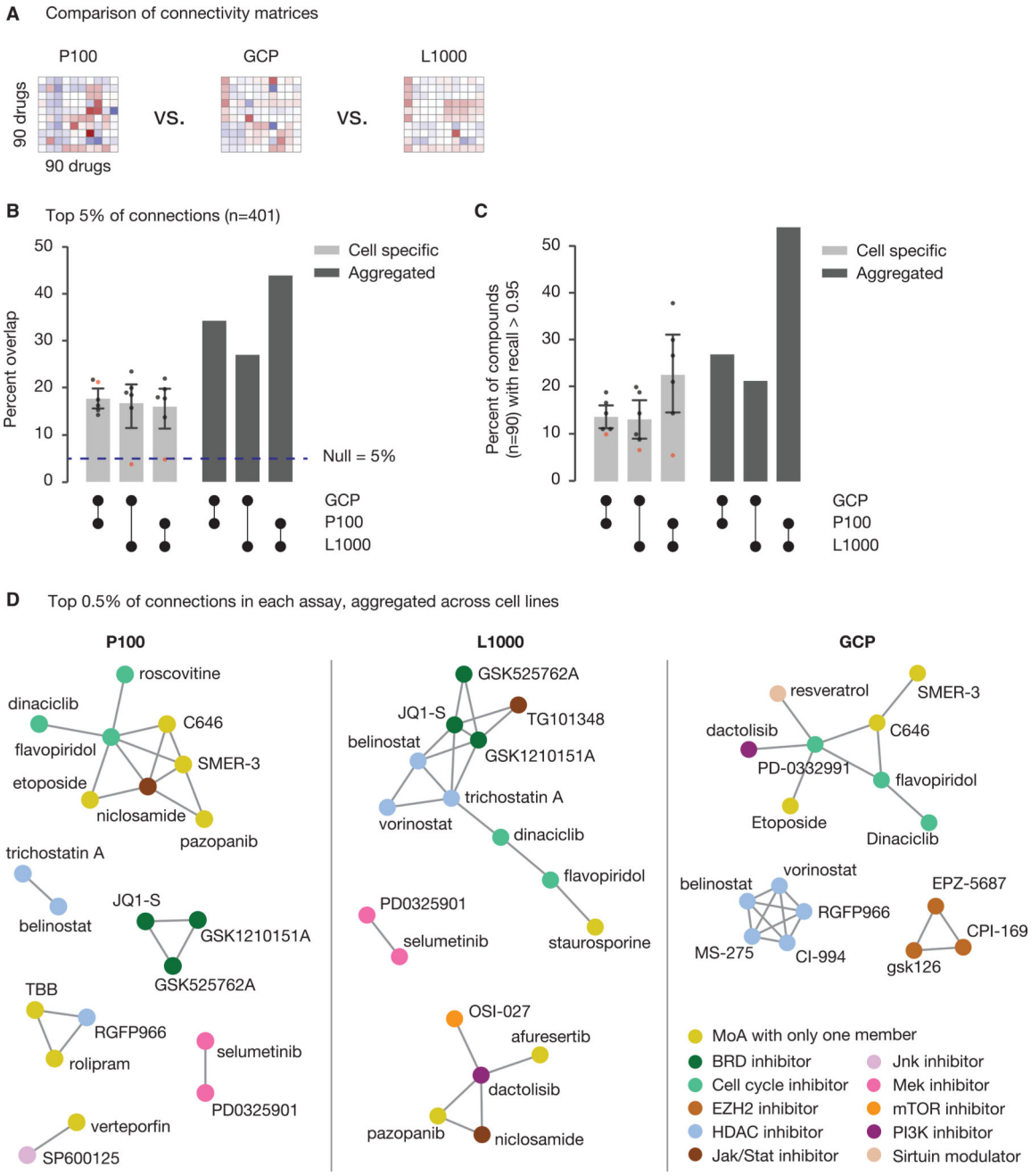


Figure 5. Comparison with Transcriptomic Data Demonstrates Assay-Specific Sensitivities

(A) Schematic of the comparison of connectivity matrices in three assays, including L1000 transcriptomic data.

(B) Percent overlap of the top 5% of connections in the P100, GCP, and L1000 assays. The light-gray bars show percent overlap for cell-specific connectivities, and the dark-gray bars show percent overlap for aggregated connectivities. The dashed line indicates the percent overlap expected by chance. The center is the median, error bars represent 95% confidence interval of 1,000 bootstrapped iterations.

(C) Recall of connectivity profiles across assays. The y axis indicates the percent of compounds (n = 90) that have recall greater than 0.95 for a pairwise comparison. Recall of 0.95 means that the connectivity profile for a particular compound in one assay had higher similarity to its corresponding connectivity profile in another assay than to 95% of other connectivity profiles (see Figure S7E for a schematic of this algorithm). The center is the median, error bars represent 95% confidence interval of 1,000 bootstrapped iterations. Shading as in (B).

(D) Network views of the top 0.5% of connections in each assay. All connectivity scores are positive. Compounds are represented by nodes, and MoA is encoded by the color of the node.

See also Figure S7.

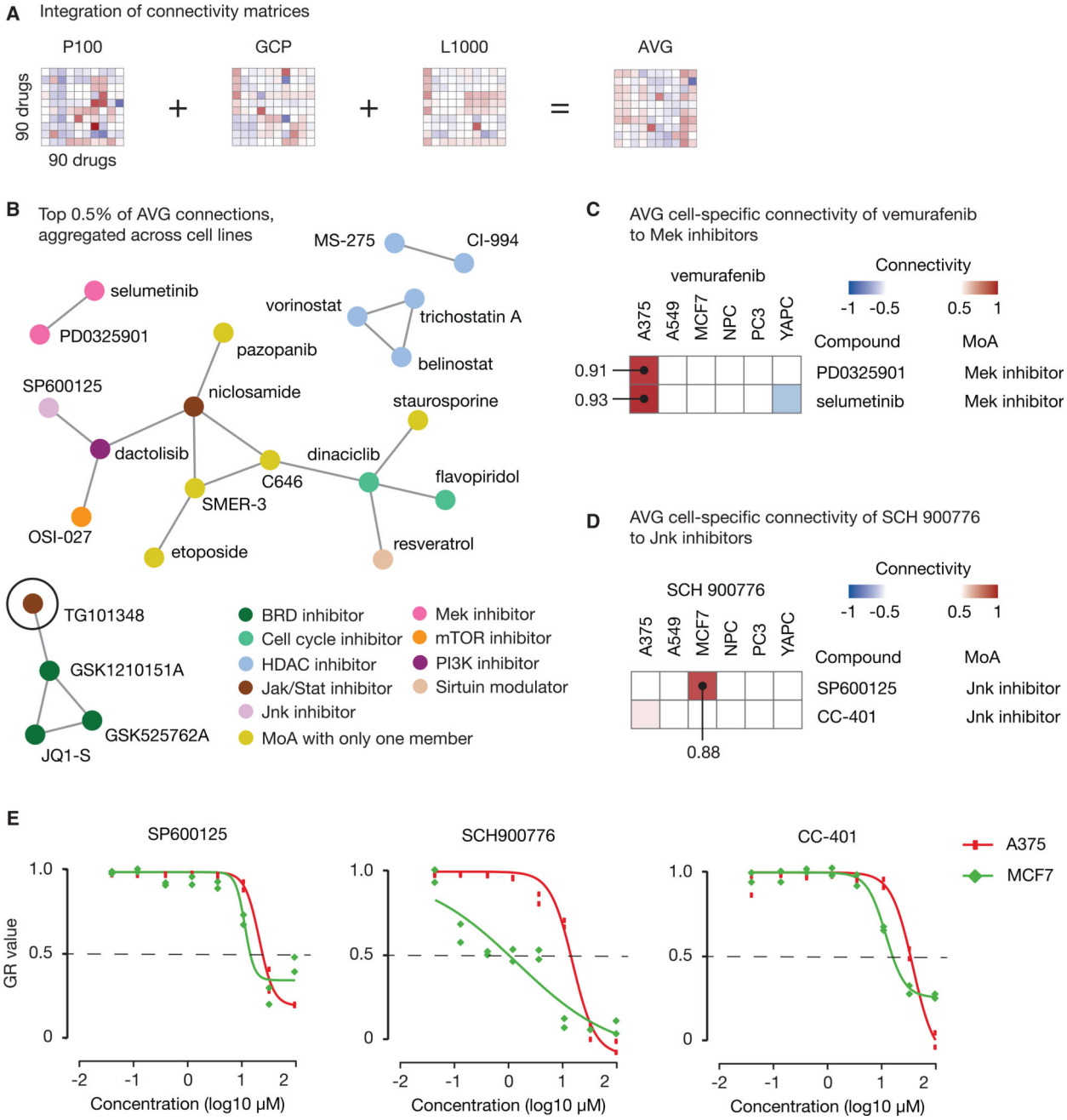


Figure 6. Multi-Assay Data Integration Reveals Cell-Specific Vulnerabilities

(A) Schematic of the integration of connectivity matrices to create AVG data.

(B) Network view of the top 0.5%of connections for AVG data, which is an average of the connectivity scores in P100, GCP, and L1000. All connectivity scores are positive. Compounds are represented by nodes, and MoA is encoded by the color of the node. TG101348 (circled) has unexpected connectivity to the BRD inhibitors.

(C) Heatmap view of the connectivity scores between vemurafenib and the two MEK inhibitors. The connectivity scores in A375 (0.91 and 0.93) are considerably higher than connectivity scores in any other cell line.

(D) Heatmap view of the connectivity scores between SCH 900776 and the two JNK inhibitors. The connectivity score between SCH 900776 and SP600125 in MCF7 (0.88) is an unexpected cell-specific connection.
(E) Results of a 5-day follow-up viability experiment. The y axis shows GR values in A375 (red rectangles) and MCF7 (green diamonds) for SP600125 (left), SCH 900776 (middle), and CC-401 (right).GR values quantify drug cytotoxicity and are insensitive to different cell growth rates. The x axis shows drug concentration on a $\log_{10}$ scale.
See also Supplemental Information.

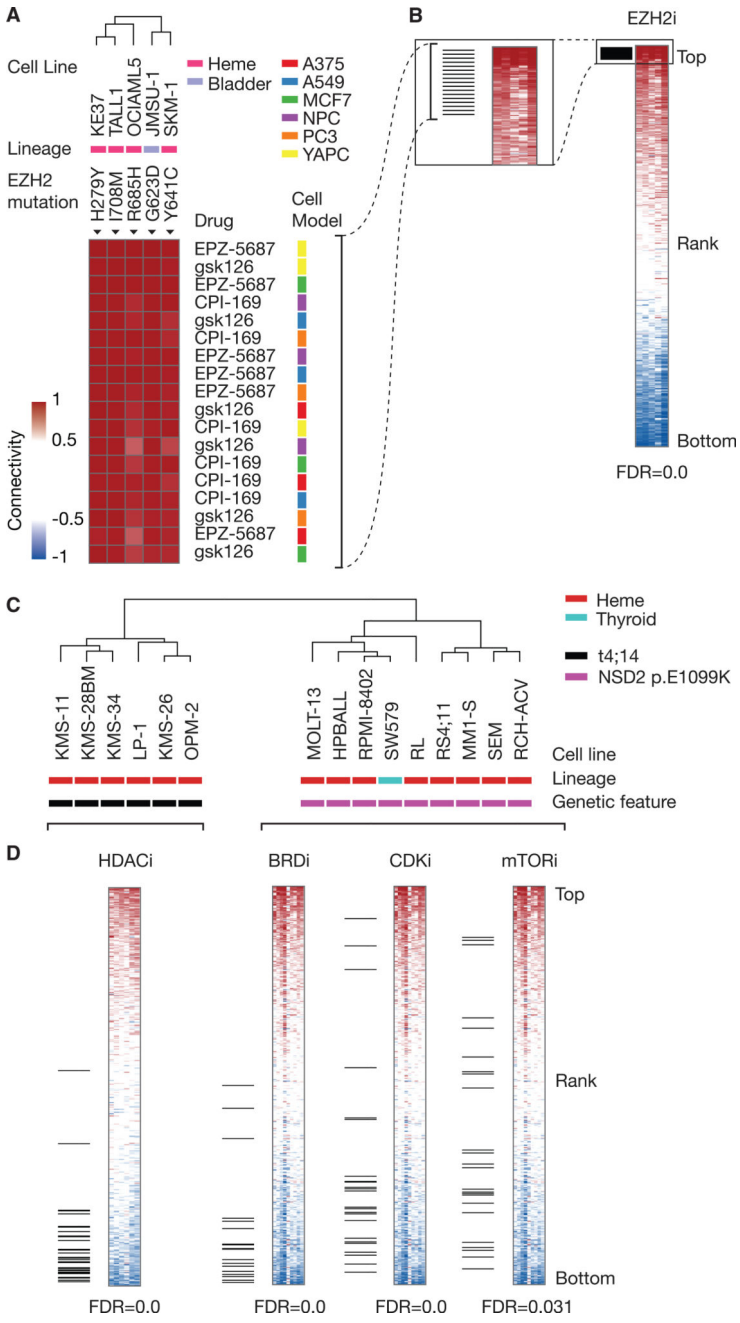


Figure 7. Connectivity Query and Perturbation Set Analysis of a Diverse Set of Cancer Lineages Validates Genetics and Identifies Potential Therapeutic Avenues

(A) Connectivity query of chromatin signatures from EZH2 loss-of-function cell lines from the CCLE. Results are sorted by the median connectivity to the perturbation across the five EZH2 loss-of-function cell lines.

(B) Adaptation of the GSEA algorithm to test for enrichment of MoA classes in connectivity results. The top ranked set is EZH2 inhibitors; all hits to this set are at the top of the list sorted by average connectivity.

(C) Stratification of two sets of NSD2 gain-of-function classes via hierarchical clustering of connectivity query results.

(D) The most highly anti-connected perturbations to the *t4;14* and *NSD2:p.Glu1099Lys* gain-of-function classes of cell lines, when ranked by connectivity for each class. Enriched perturbation sets with FDR of <0.05 are shown for each class. The HDAC inhibitors are the most anti-connected perturbations to the *t4;14* subtype while BRD, CDK, and mTOR inhibitors are all anti-connected to the *NSD2:p.Glu1099Lys* subtype.
See also Figure S8.

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Chemicals, Peptides, and Recombinant Proteins		
See Table S1 for list of compounds profiled	Table S1	N/A
P100 reagents	Abelin et al., 2016	N/A
GCP reagents	Creech et al., 2015	N/A
L1000 reagents	Subramanian et al., 2017	N/A
C18 Sep-Pak Cartridge	Waters	Cat# 186002319
AssayMAP Fe(III)-NTA cartridges	Agilent	Cat# G5496-60085
AssayMAP reversed phase (RPS) cartridges	Agilent	Cat# G5496-60033
mTeSR1 (cGMP) media	StemCell Technology	Cat# 85850
Matrigel Matrix	BD Biosciences	Cat# 356231
Dorsomorphin dihydrochloride	Tocris Bioscience	Cat# 3093
SB 431542	Tocris Bioscience	Cat# 1614
DMEM/F-12 GlutaMAX	Thermo Fisher	Cat# 10565-018
N-2 Supplement	Thermo Fisher	Cat# 17502-048
Insulin	Sigma-Aldrich	Cat# I9278-5ML
L-Glutamine	Thermo Fisher	Cat# 25030-024
MEM Non-Essential Amino Acids Solution	Thermo Fisher	Cat# 11140-050
2-Mercaptoethanol	Sigma-Aldrich	Cat# M7522
Penicillin-Streptomycin	Thermo Fisher	Cat# 15140-122
Neurobasal Medium, minus phenol red	Thermo Fisher	Cat# 12348-017
B-27 Supplement, minus vitamin A	Thermo Fisher	Cat# 125870-10
Deposited Data		
Extracted ion chromatogram MS data	This study	panoramaweb.org/labkey/LINCS.url
Proteomic signatures	This study	GEO: GSE101406
Connectivity matrices	This study	clue.io/proteomics
Experimental Models: Cell Lines		
Human A375 (malignant melanoma)	ATCC	Cat# CRL-1619; RRID:CVCL_0132
Human A549 (non-small-cell lung carcinoma)	ATCC	Cat# CCL-185; RRID:CVCL_0023
Human MCF7 (breast adenocarcinoma)	ATCC	Cat# HTB-22; RRID:CVCL_0031
Human NPC	Differentiated from ESCs as described here	N/A
Human PC3 (prostate adenocarcinoma)	ATCC	Cat# CRL-1435; RRID:CVCL_0035
YAPC (pancreas carcinoma)	DSMZ	Cat# ACC 382; RRID:1794
H9 ESCs	WiCell	Cat# WA09; RRID:CVCL_9773
Software and Algorithms		
Skyline	MacLean et al., 2010	proteome.gs.washington.edu/software/skyline
Proteomics Signature Pipeline	This study	github.com/cmap/psp
cmapPy	N/A	github.com/cmap/cmappy

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Gene Set Enrichment Analysis (GSEA)	Subramanian et al., 2005	software.broadinstitute.org/gsea