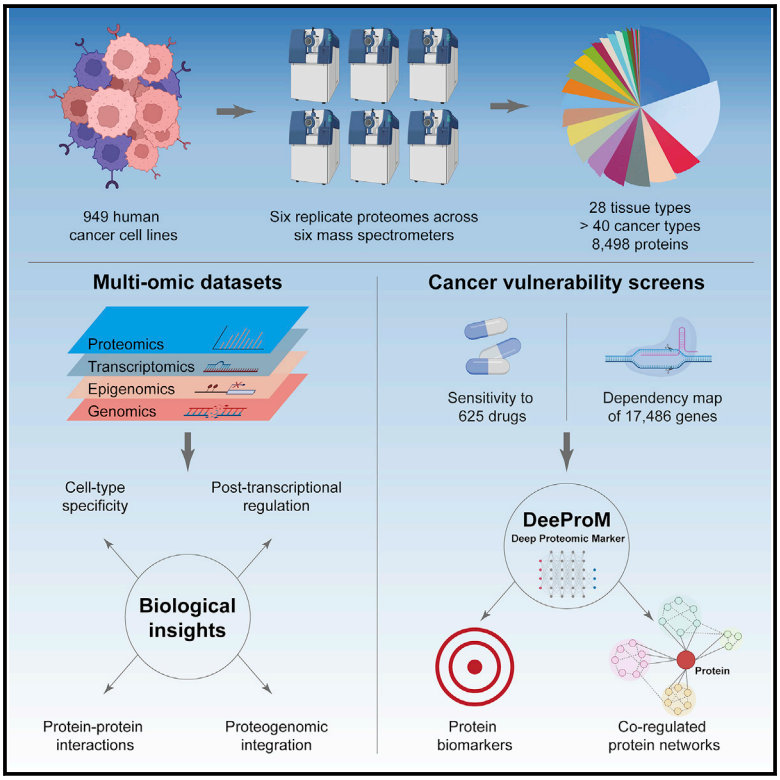
Cancer Cell

Article

Pan-cancer proteomic map of 949 human cell lines

Graphical abstract



Highlights

- Pan-cancer proteomic map of 949 human cancer cell lines across over 40 cancer types
- There are 8,498 proteins with evidence of cell types and broad post-transcriptional regulation
- Deep learning-based pipeline finds biomarkers of drug response and gene essentiality
- Random downsampling reveals highly connected and co-regulated protein networks

Authors

Emanuel Gonçalves, Rebecca C. Poulos, Zhaoxiang Cai, ..., Qing Zhong, Mathew J. Garnett, Roger R. Reddel

Correspondence

probinson@cmri.org.au (P.J.R.), qzhong@cmri.org.au (Q.Z.), mg12@sanger.ac.uk (M.J.G.), rreddel@cmri.org.au (R.R.R.)

In brief

Gonçalves et al. generate a comprehensive proteomic map of 949 human cancer cell lines across more than 40 cancer types. Proteomic data are integrated with multi-omic, drug response, and CRISPR-Cas9 gene essentiality datasets. Deep learning is used to identify biomarkers of cancer vulnerabilities, providing evidence for highly connected protein networks.

Gonçalves et al., 2022, Cancer Cell *40*, 835–849
August 8, 2022
https://doi.org/10.1016/j.ccell.2022.06.010

CellPress

Cancer Cell

CellPress
OPEN ACCESS

Article

Pan-cancer proteomic map of 949 human cell lines

Emanuel Gonçalves,[1,2,3,5] Rebecca C. Poulos,[4,5] Zhaoxiang Cai,[4,5] Syd Barthorpe,[1] Srikanth S. Manda,[4] Natasha Lucas,[4] Alexandra Beck,[1] Daniel Bucio-Noble,[4] Michael Dausmann,[4] Caitlin Hall,[1] Michael Hecker,[4] Jennifer Koh,[4] Howard Lightfoot,[1] Sadia Mahboob,[4] Iman Mali,[1] James Morris,[1] Laura Richardson,[1] Akila J. Seneviratne,[4] Rebecca Shepherd,[1] Erin Sykes,[4] Frances Thomas,[1] Sara Valentini,[1] Steven G. Williams,[4] Yangxiu Wu,[4] Dylan Xavier,[4] Karen L. MacKenzie,[4] Peter G. Hains,[4] Brett Tully,[4] Phillip J. Robinson,[4,*] Qing Zhong,[4,*] Mathew J. Garnett,[1,*] and Roger R. Reddel[4,6,*]

[1]Wellcome Sanger Institute, Wellcome Genome Campus, Cambridge CB10 1SA, UK
[2]Instituto Superior Técnico (IST), Universidade de Lisboa, 1049-001 Lisboa, Portugal
[3]INESC-ID, 1000-029 Lisboa, Portugal
[4]ProCan®, Children's Medical Research Institute, Faculty of Medicine and Health, The University of Sydney, Westmead, NSW, Australia
[5]These authors contributed equally
[6]Lead contact
*Correspondence: probinson@cmri.org.au (P.J.R.), qzhong@cmri.org.au (Q.Z.), mg12@sanger.ac.uk (M.J.G.), rreddel@cmri.org.au (R.R.R.)
https://doi.org/10.1016/j.ccell.2022.06.010

SUMMARY

The proteome provides unique insights into disease biology beyond the genome and transcriptome. A lack of large proteomic datasets has restricted the identification of new cancer biomarkers. Here, proteomes of 949 cancer cell lines across 28 tissue types are analyzed by mass spectrometry. Deploying a workflow to quantify 8,498 proteins, these data capture evidence of cell-type and post-transcriptional modifications. Integrating multi-omics, drug response, and CRISPR-Cas9 gene essentiality screens with a deep learning-based pipeline reveals thousands of protein biomarkers of cancer vulnerabilities that are not significant at the transcript level. The power of the proteome to predict drug response is very similar to that of the transcriptome. Further, random downsampling to only 1,500 proteins has limited impact on predictive power, consistent with protein networks being highly connected and co-regulated. This pan-cancer proteomic map (ProCan-DepMapSanger) is a comprehensive resource available at https://cellmodelpassports.sanger.ac.uk.

INTRODUCTION

Precision medicine relies on the identification of specific molecular alterations that can stratify patients and guide the choice of effective therapeutic options. Cancer vulnerabilities, such as synthetic lethalities, can be systematically studied using functional genetic and small molecule screens. To circumvent the limitations of using patient tissue samples for this type of approach, biomarkers of cancer vulnerabilities have been analyzed using cancer cell lines, together with deep molecular characterization, functional genetic and pharmacological screens (Ghandi et al., 2019; Iorio et al., 2016; Tsherniak et al., 2017; Frejno et al., 2020; Behan et al., 2019). The direct measurement of proteins provides insights into the dynamic molecular behavior of cells and can improve our understanding of genotype-to-phenotype relationships (Liu et al., 2016). Despite the development of precision oncology therapeutics, the complexity of cancer and the inability of genomics to accurately predict the proteome indicate that genomics alone is often insufficient to inform and guide the clinical care of many patients. Measurement of the proteome has the potential to expand our understanding of cancer phenotypes and to improve diagnosis and treatment choices.

Technological and methodological advances have enabled the standardized quantification of thousands of proteins across dozens to hundreds of cell lines (Gholami et al., 2013; Lawrence et al., 2015; Coscia et al., 2016; Roumeliotis et al., 2017; Nusinow et al., 2020) and the profiling of clinical samples derived from minute tissue biopsies (Zhang et al., 2014; Edwards et al., 2015; Pozniak et al., 2016; Zhang et al., 2016; Mertins et al., 2016; Frejno et al., 2017; Vasaikar et al., 2019; Clark et al., 2019; Tully 2020). Using a data-independent acquisition (DIA)-mass spectrometry (MS) approach (Gillet et al., 2012; Guo et al., 2015; Ludwig et al., 2018; Lucas et al., 2019), together with a sample processing workflow with novel data processing methods, it is now possible for proteomes to be acquired reproducibly at scale (Tully et al., 2019; Poulos et al., 2020). The generation and distribution of large-scale proteomic datasets have the potential to drive new computational approaches, including deep learning-based algorithms, to investigate the impact of molecular changes on cancer vulnerabilities. This will enable proteomics to contribute important clinical advances for cancer therapeutic applications.

Cell lines have been invaluable models for our understanding of cellular processes and the molecular drivers of carcinogenesis

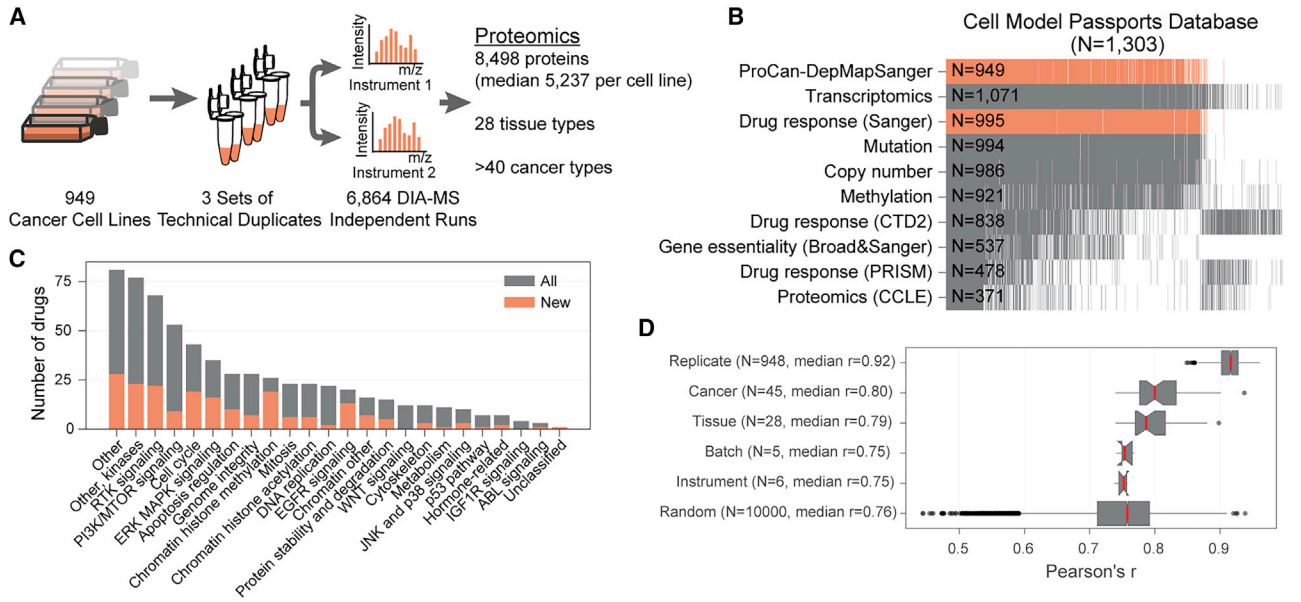


Figure 1. A pan-cancer proteomic map of 949 human cancer cell lines

(A) Methodology overview for pan-cancer characterization of 949 human cell lines using a DIA-MS workflow.

(B) Proteomic measurements were integrated with independent molecular and phenotypic datasets spanning 1,303 cancer cell lines as part of the Cell Model Passports Database. Data include proteomics (ProCan-DepMapSanger) presented here, transcriptomics, drug response (Sanger), mutation, copy number, methylation, drug response (CTD2), CRISPR-Cas9 gene essentiality (Broad&Sanger), drug response (PRISM), and proteomics (CCLE). Each gray slice denotes a unique cell line, and the total number of cell lines per dataset is indicated. The proteomic data (ProCan-DepMapSanger) generated in this study are shown in orange, as well as the expanded drug response (Sanger) dataset.

(C) Number of drugs included in the drug response (Sanger) screen, with the orange bar highlighting the additional number of unique drugs presented in this study compared to previous studies. Drugs are grouped by the pathway of their canonical targets.

(D) Pearson's correlations of the proteomes for each set of six technical replicates, as well as each cancer type, tissue type, batch and instrument. Random indicates the correlation between random unmatched sets of replicates. Median Pearson's *r* for each group is reported. Box-and-whisker plots indicate interquartile range (IQR) with a line at the median. Whiskers represent the minimum and maximum values at 1.5 × IQRs.

See also Figure S1, Tables S1 and S2.

(Garnett et al., 2012; Iorio et al., 2016; Barretina et al., 2012; Tsherniak et al., 2017; Meyers et al., 2017; Behan et al., 2019; Ghandi et al., 2019; Picco et al., 2019), and for identifying cancer cell vulnerabilities to both genetic (McDonald et al., 2017; Tsherniak et al., 2017; Hart et al., 2015; Meyers et al., 2017; Behan et al., 2019) and pharmacological (Garnett et al., 2012; Iorio et al., 2016; Corsello et al., 2020; Barretina et al., 2012; Seashore-Ludlow et al., 2015; Rees et al., 2016) perturbations. However, proteomic quantifications for cancer cell lines are either limited in the range of cancer types or number of samples analyzed, or are largely unavailable (Nusinow et al., 2020; Gholami et al., 2013; Li et al., 2017; Frejno et al., 2020; Guo et al., 2019). For this reason, biomarkers for cancer vulnerabilities have so far been primarily based on genomic and transcriptomic measurements (Iorio et al., 2016; Tsherniak et al., 2017; Ghandi et al., 2019). Consequently, little is known about the contribution of the proteome to cancer vulnerabilities or how the cancer proteome is regulated in diverse tissues and genetic contexts.

This study reports a pan-cancer cell line proteomic map quantifying 8,498 proteins across 949 cell lines. The generation and analysis of this rich resource involved the development of a workflow with rapid sample processing and minimal complexity, followed by the application of a deep neural network-based computational pipeline to uncover cancer targets. Integration of our proteomic data (referred to as the ProCan-DepMapSanger dataset) with existing molecular and phenotypic datasets from the Cancer Dependency Map (Boehm et al., 2021), showed that protein networks are more strongly co-regulated than are transcriptomics and functional genomics. Our approach identified biomarkers of well-established cancer vulnerabilities and, more important, highlighted those that cannot be identified with genomics or transcriptomics alone. The proteome measured in our study had an equivalent performance to the total measured transcriptome in predicting cancer phenotypes. Furthermore, random subsets of 1,500 proteins downsampled from the complete dataset achieved 88% of the power to predict drug responses. These results have broad implications for the design of future studies, ranging from basic research to clinical applications.

RESULTS

A resource of 949 cancer cell line proteomes

To construct a pan-cancer proteomic map, proteomes of 949 human cancer cell lines from 28 tissues and more than 40 genetically and histologically diverse cancer types were quantified (Figures 1A and S1A, Table S1). The proteome for each cell line was acquired by DIA-MS from six replicates using a workflow that enables high throughput and minimal instrument downtime (see STAR Methods, Figure S1B). The resulting

dataset was derived from 6,864 DIA-MS runs acquired over 10,000 MS h (Table S1), including peptide preparations derived from the human embryonic kidney cell line HEK293T that were used throughout all data acquisition periods and instruments for quality control. These data, together with the spectral library, were deposited in the Proteomics Identification Database (Perez-Riverol et al., 2019) with dataset identifier PXD030304. Raw DIA-MS data were processed with DIA-NN (Demichev et al., 2020), using retention time-dependent normalization and with a spectral library generated by DIA-NN. For full details of data processing steps and parameters, see STAR Methods and Table S1. MaxLFQ (Cox et al., 2014) was then used to quantify a total of 8,498 proteins (Table S2, Figure S1C), with a median of 5,237 proteins (min-max range: 2,523–6,251) quantified per cell line (Table S1, Figure 1A). The ProCan-DepMapSanger dataset significantly expands the existing molecular characterizations of this broad range of cancer cell line models (Figure 1B). Pharmacological screens of anti-cancer drugs tested against this cell line panel were also expanded in this study, increasing the number of unique drugs tested by 48% over our prior work (n = 625 drugs and investigational compounds; Figure 1C), with a total of 578,238 half-maximal inhibitory concentrations (IC_{50}) experimentally measured.

High correlations were observed between replicates of each cell line, yielding a sample-wise median Pearson's correlation coefficient (Pearson's *r*) of 0.92 (Figures 1D and S1A). Correlations between unmatched samples from the same instrument or batch were similar to random (median Pearson's $r = 0.75$, Figure 1D). Although integration of outputs from different proteomics platforms is acknowledged to be an unsolved challenge, we have compared our data with previously published proteomic datasets comprising smaller subsets of the same cell lines (Lawrence et al., 2015; Roumeliotis et al., 2017; Guo et al., 2019; Nusinow et al., 2020; Frejno et al., 2020) and have shown comparable levels of correlation among all datasets (Figure S1D). Nonlinear dimensionality reduction using Uniform Manifold Approximation and Projection (UMAP) (McInnes et al., 2018) showed no evidence of instrument or batch effects (Figure S1E). Proteins that were detected had higher mean RNA expression across the cell line panel than proteins that were not detected in this study ($p < 0.0001$ by the Mann-Whitney U test) (Figure S1F), suggesting some bias toward abundant proteins. Overall, this study generated a high-quality and biologically reproducible pan-cancer proteome map of human cancer cell lines.

Proteomic profiles reveal cell type of origin

Next, we defined a stringent set of protein quantifications that were supported by measuring more than one peptide (n = 6,692 human proteins) (Table S2). Visualization of these protein intensities via UMAP showed groupings by cell type of origin, such as distinct clusters of hematopoietic and lymphoid cells and skin cells (Figure 2A). Hematopoietic and lymphoid cells appeared to exhibit further subgroups, and we found that this cell type could be segregated into different cell lineages (Figure 2B). This high-level dimensionality reduction suggested a profile of protein expression that relates to cell type of origin. To investigate this further, a set of 279 proteins that are enriched in certain cell types was selected (Table S3). These cell type-enriched proteins were defined as any protein quantified in 50% or more of cell lines from no more than two tissue types and 35% or less of cell lines from all remaining tissues, considering only tissues represented by at least 10 cell lines (Figure 2C). Cell lines from hematopoietic and lymphoid, peripheral nervous system, and skin cell types showed the greatest numbers of these proteins (Figure 2D). Proteins encoded by genes representing gene ontology terms for lymphocyte activation, neuron projection, and pigmentation were identified in each of these cell types, respectively. Further, the cell type-enriched proteins had a higher correlation between the transcriptome and proteome than did other proteins, suggesting that these represent cell type-specific processes that are more highly conserved between transcription and translation (Figure 2E). Overall, this analysis demonstrates a general alignment of the proteomic data with cell lineage, revealing patterns of protein expression that are consistent with some cancer cell types of origin.

Post-transcriptional regulation in diverse cancer cell types

We next sought to identify the key drivers of the distinct protein expression patterns observed across the cell line panel and to investigate how these integrate with other molecular and phenotypic measurements. Multi-omics factor analysis (MOFA) (Argelaguet et al., 2018, 2020) was used to integrate the proteomic measurements with a range of molecular (promoter methylation, gene expression and protein abundance) and phenotypic (drug response) datasets available for most cancer cell lines (Table S4 and Figure 3A). MOFA uses a Bayesian group factor analysis framework to enable unsupervised integration of datasets to infer a set of factors (latent variables) that account for biological and technical variability in the data (Argelaguet et al., 2018, 2020). Epithelial-to-mesenchymal transition (EMT) canonical markers, vimentin and E-cadherin, and EMT gene set enrichment analysis enrichment scores (Subramanian et al., 2005; Liberzon, 2011) were found to be associated with the first two factors (F1 and F2) corresponding with large portions of the variability across all datasets (Figure 3A). Technical aspects like cell line media and growth conditions explained little or no variability, while cell size and growth rate were moderately related with some factors (Figure S2A). Cancer cell lines from the same tissue of origin showed gradients of EMT markers (Figure 3B), which may indicate that the cell lines were established from cancers derived from different epithelial or mesenchymal lineages, or that the cancers had undergone EMT. EMT markers are known to be associated with different stages of cancer progression including initiation, metastasis, and the development of therapy resistance (Brabletz et al., 2018). We observed that some factors capture tissue-specific processes, with their loadings being enriched toward the cell type-enriched proteins defined earlier (Figure 3A). For example, a MOFA analysis highlighted an association between factor 12 and skin-derived cell lines (Figure S2B), which in this study are primarily from melanomas. Factor 12 also related to phenotypic measurements that are typical of melanomas, correlating with CRISPR-Cas9 gene essentiality scores for BRAF (Figure 3C) and with its inhibitor, dabrafenib (Figure 3D), both of which are strongly associated with cell lines harboring BRAF mutations that are very common in cutaneous melanomas.

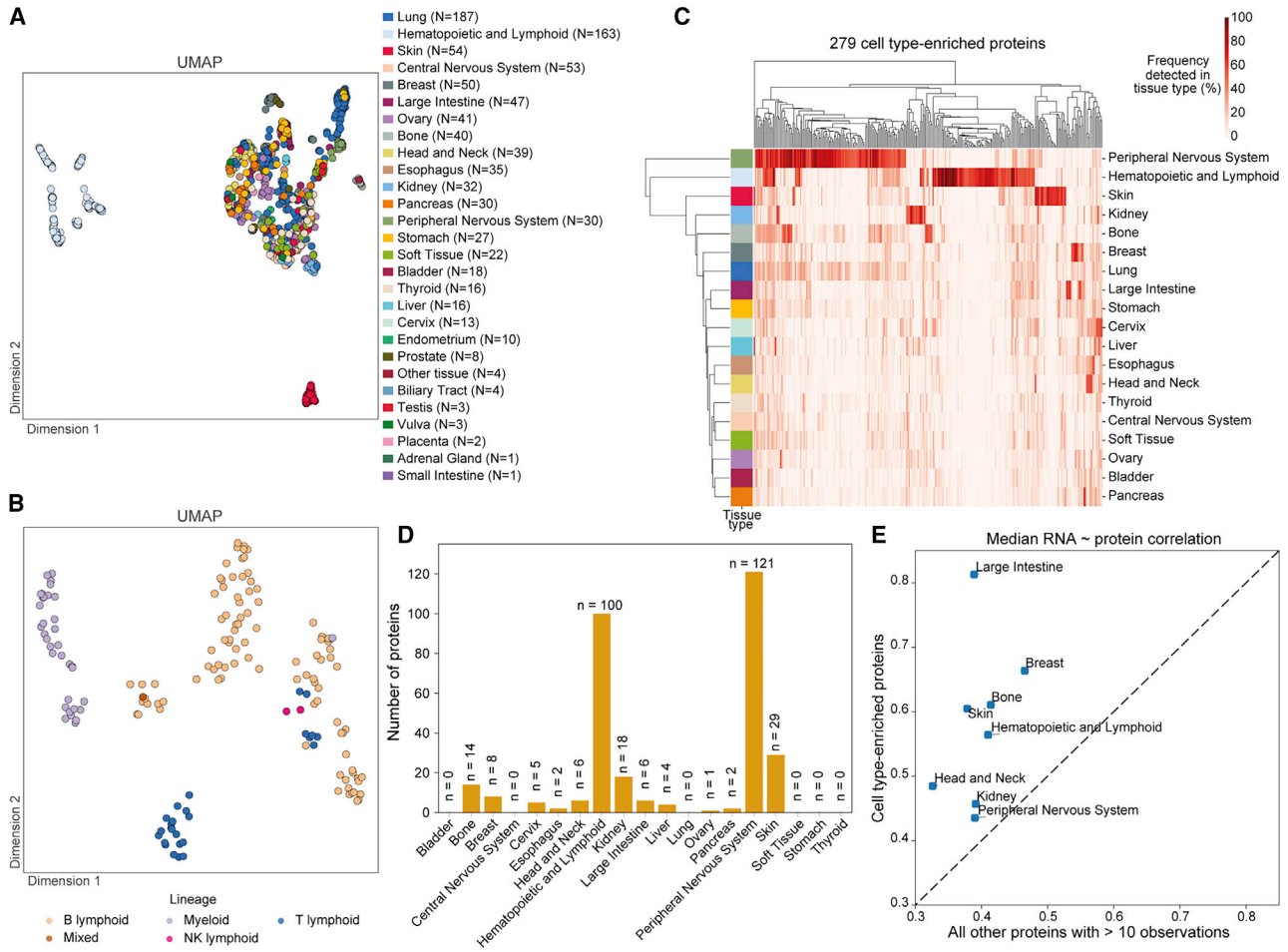


Figure 2. Distinct proteomic profiles according to cell type

(A) Proteomic data dimensionality reduction by UMAP, with cell lines colored by tissue.

(B) UMAP of hematopoietic and lymphoid cell lines colored by cell lineage.

(C) Heatmap of the frequency of cell type-enriched proteins observed within each tissue. Tissues and proteins are clustered on the vertical and horizontal axes, respectively.

(D) Number of cell type-enriched proteins identified in each tissue type represented by more than 10 cell lines.

(E) Median RNA-protein correlation of cell type-enriched proteins against all other proteins with more than 10 observations in that tissue type. Only tissues with at least five cell type-specific proteins are shown.

See also Table S3.

While EMT markers were largely concordant between transcriptomics and proteomics, more broadly, a modest and variable association was observed between protein and transcript measurements (median protein-wise Pearson's $r = 0.42$, Figure S2C). This is consistent with the expectation that proteomic data capture variability explained by post-transcriptional regulation and the proteostasis network. Focusing on the impact of genomic alterations on proteomic profiles, gene copy number was more weakly correlated with protein levels than with gene expression, indicating attenuation of copy number effects between the transcriptome and the proteome (Figures 3E and Table S4). This was particularly evident among subunits of protein complexes such as those involved in ribosomes (Figure S2D), which can co-regulate their abundance post-transcriptionally to maintain complex stability and stoichiometry (Gonçalves et al., 2017; Roumeliotis et al., 2017; Ryan et al., 2017; Sousa et al., 2019). Proteins involved in protein synthesis and degradation had some of the strongest post-transcriptional regulation, with several proteasome and ribosome subunits showing strong attenuation (Figure S2D). Together, this reflects an active proteostasis network, revealed primarily via direct measurement of the proteome (Gumeni et al., 2017).

Using somatic mutation data to stratify the full set of protein quantifications according to mutation status (Table S4) revealed 478 proteins with significant differential protein intensities between cell lines that were wild type versus those that had protein-coding mutations (false discovery rate [FDR]-adjusted p value < 0.05) (Figure 3F). When mutations were present in the gene encoding a given protein, the majority of proteins had decreased abundance (n = 354 proteins). In contrast, mutations in some genes, such as *TP53,* were associated with significantly higher protein intensities than wild-type cell lines

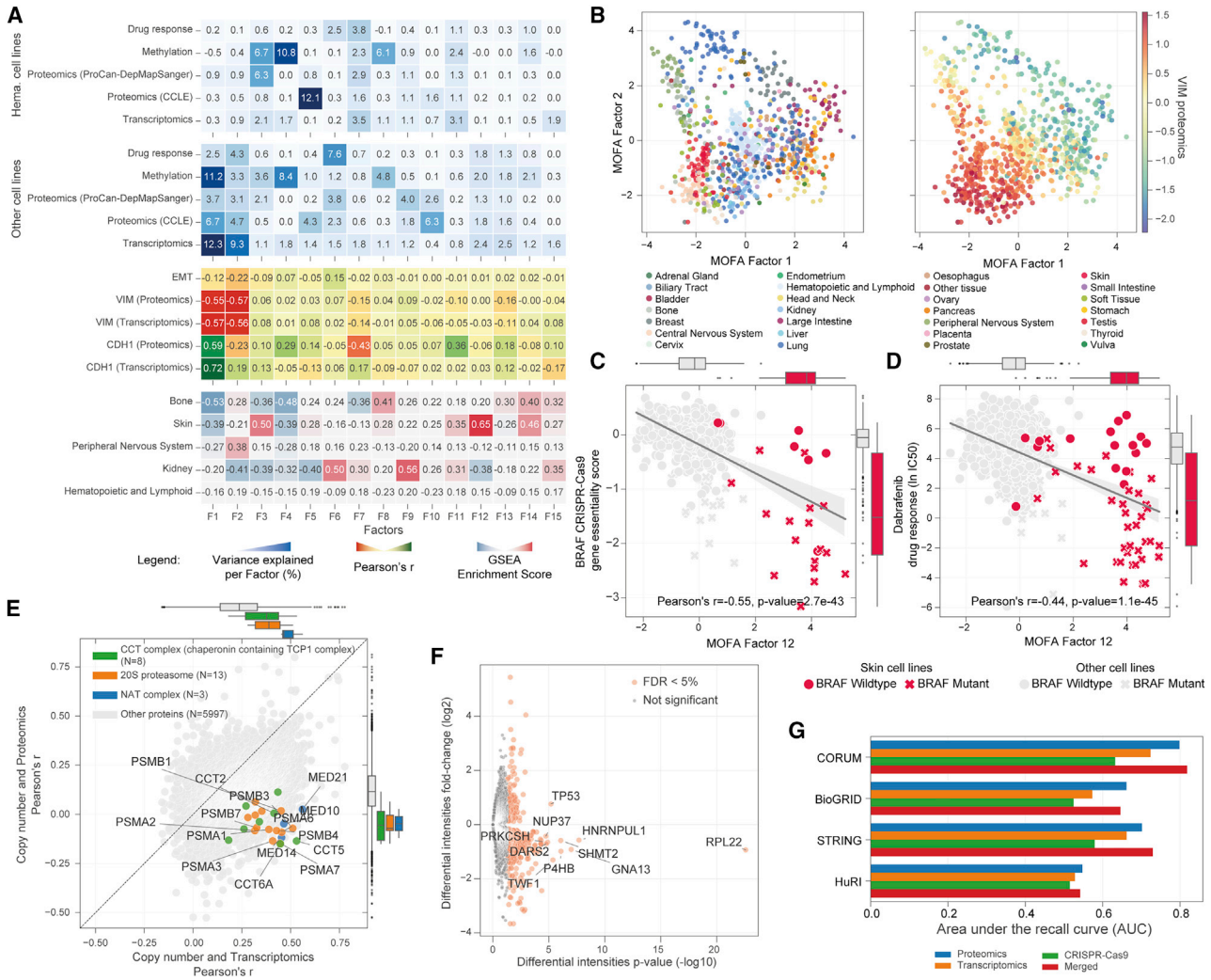


Figure 3. Post-transcriptional regulatory mechanisms of cancer cell lines

(A) Identification of shared variability (factors) from MOFA across multiple molecular and phenotypic cancer cell line datasets. Hematopoietic and lymphoid cells are grouped and trained separately from the other cell lines. The upper two heatmaps (blue) report the portion of variance explained by each factor (columns) in each dataset. The central (yellow) heatmap reports Pearson's *r* between each learned factor and various molecular characteristics of the cancer cell lines. The lower heatmap shows gene set enrichment analysis (GSEA) enrichment scores of each factor to cell type-specific proteins.

(B) Separation of cancer cell lines by MOFA factors 1 and 2, colored by tissue of origin (left) and by EMT canonical marker vimentin (VIM) protein intensities (right).

(C) Scatterplot with linear regression between MOFA factor 12 and BRAF CRISPR-Cas9 gene essentiality scores. Skin cancer cell lines are highlighted in red, and BRAF mutant cell lines are marked with a cross.

(D) Similar to (C), but instead the vertical axis indicates the dabrafenib drug response (IC_{50}) measurements.

(E) Pearson's *r* between gene absolute copy number profiles with transcriptomics (horizontal axis) and with protein intensities from the ProCan-DepMapSanger dataset (vertical axis). Representative Comprehensive Resource of Mammalian Protein Complexes (CORUM) protein complexes with the highest differences between the Pearson's *r* are shown, and the top 15 most attenuated proteins from these complexes are labeled. N indicates the number of proteins in each protein complex. Box-and-whisker plots represent the Pearson's *r* distributions of proteins involved in each highlighted gene ontology term compared with all proteins (gray).

(F) Volcano plot showing differential protein intensities between cell lines that are wild type versus mutant for each protein in the ProCan-DepMapSanger dataset that is mutated in at least 1% of the cohort. The top 10 proteins by p value are annotated. The horizontal axis shows the $-\log_{10}$ of the empirical Bayes moderated t test p value, and proteins with FDR of less than 5% are colored in red.

(G) Recall of PPIs, i.e., ability to detect known PPIs, from resources CORUM, STRING, BioGRID, and HuRI. All possible protein pairwise correlations (Pearson's p value) were ranked, using proteomics, transcriptomics, and CRISPR-Cas9 gene essentiality. The merged score was defined as the product of the p values of the different correlations.

In (C), (D), and (E), box-and-whisker plots indicate interquartile range (IQR) with a line at the median. Whiskers represent the minimum and maximum values at 1.5 × the IQRs. See also Figure S2 and Table S4.

(TP53: FDR-adjusted p value < 0.0001). This is consistent with the known increase in stability of many mutant P53 proteins, which results from a decreased rate of proteasome-mediated degradation (Vijayakumaran et al., 2015).

These results indicate that, while variability in protein expression is associated with other molecular and phenotypic layers, there is only partial correlation between transcript and protein abundance, consistent with the effects of post-transcriptional regulation. Thus, the ProCan-DepMapSanger dataset captures additional protein-specific information that can augment our understanding of the impact of genomic alterations affecting, among others, well-established cancer genes.

Co-regulatory protein networks of cancer cells

Having observed post-transcriptional co-regulation of protein complex abundance (Figure 3E), we next investigated whether the abundance of co-regulated proteins could be used to predict putative protein-protein interactions (PPIs). We assessed all possible pairwise protein correlations (n = 16,580,952) and, as a comparator, used corresponding gene expression and CRISPR-Cas9 gene essentiality profiles, where available. As expected, paralogs and protein complex subunits had some of the strongest correlations (Table S4; absolute Pearson's $r > 0.5$ and FDR adjusted p value < 0.05). We systematically assessed this enrichment using multiple resources for protein interactions: protein complex interactions from the Comprehensive Resource of Mammalian Protein Complexes (CORUM) (Ruepp et al., 2010); functional interactions from the Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) database (Szklarczyk et al., 2017); physical protein interactions from the Biological General Repository for Interaction Datasets (BioGRID) (Chatr-Aryamontri et al., 2015); and the Human Protein Interactome (HuRI) (Luck et al., 2020). For all resources, proteomic measurements had greater ability to detect known PPIs (area under the recall curve [AUC] = 0.55–0.80) than transcriptomics (AUC = 0.53–0.72) and CRISPR-Cas9 gene essentiality (AUC = 0.51–0.63) (Figure 3G), indicating that PPIs and co-regulation are best captured by proteomics.

The correlation p value score (Fisher's combined probability test) for BioGRID and HuRI databases was improved slightly by merging different datasets (AUC = 0.54–0.82), suggesting that different types of interactions are captured across multi-omic layers. Proteins with higher numbers of positive protein-protein correlations were more essential for cancer cell survival, as observed in the CRISPR-Cas9 gene essentiality dataset (Figure S2E). This is likely linked to their increased transcript and protein expression levels, and the number of pathways in which they are involved. Paralogs were an exception, as they had largely non-essential profiles independent of their gene expression (Figure S2F), consistent with their functional redundancy attenuating the impact of loss (Dandage and Landry 2019).

Considering the high correlations that were observed with known interactions, these pairwise protein correlations could be used to identify novel putative PPIs, such as between protein subunits. Consistent with this, we identified 1,182 putative PPIs with a Pearson's r greater than 0.8 that are not reported in any of the protein network resources analyzed here. For example, there were strong correlations between protein profiles for EEF2-EIF3I, RPSA-SERBP1 and CCT6A-EEF2 (Table S4). These proteins are not reported to interact directly but are also closely related in the high confidence STRING protein interaction network (Szklarczyk et al., 2017). Overall, this analysis highlights the ability for protein measurements to detect known interactions, and suggests the utility of this dataset for predicting co-regulated interactions.

Identifying biomarkers of cancer vulnerabilities

Next, we considered the application of proteomics to identify biomarkers by harnessing drug (Garnett et al., 2012; Iorio et al., 2016; Picco et al., 2019; Gonçalves et al., 2020) and CRISPR-Cas9 gene essentiality (Behan et al., 2019; Meyers et al., 2017; Pacini et al., 2021) screens. The previous Sanger pharmacological screens were expanded to include a total of 578,238 IC_{50} values (Figure 1B). These included a total of 625 unique anti-cancer drugs (48% increase in unique drugs over previously published datasets (Iorio et al., 2016; Picco et al., 2019; Gonçalves et al., 2020)) that were screened across 947 of the 949 cancer cell lines, including U.S. Food and Drug Administration-approved drugs, drugs in clinical development, and investigational compounds. To identify protein biomarkers predictive of cancer cell line response to these drugs or CRISPR-Cas9 gene essentialities (n = 17,486), we applied linear regression to test all pairwise associations between proteins, drug sensitivity, and CRISPR-Cas9 gene dependencies, while considering potentially confounding effects, such as cell line growth rate, culture media, and the average replicate correlation (see STAR Methods for more details) (Figure 4A, Table S5). Among the strongest significant associations (FDR < 5%), we observed that 57 drugs were associated with the protein abundance of their canonical target(s), including negative associations between EGFR protein abundance and its inhibitor gefitinib, and MET protein abundance and its inhibitor drug response (Figure 4A). We observed a significant negative association between ERBB2 (also known as HER2) and lapatinib, a tyrosine kinase inhibitor that targets EGFR and HER2 (Figure 4A). This association has been observed in proteomic studies using other preclinical models (breast cancer patient-derived xenografts) (Huang et al., 2017) and in human cancers, with lapatinib already an approved drug used in the treatment of HER2-positive breast cancers (Xu et al., 2017). For another 132 drugs, significant associations were identified with proteins functionally related with their targets (i.e., one step away in the STRING PPI network). The majority of the significant drug-protein target associations showed a negative effect size (Figure 4A), indicating greater sensitivity of a cell line to a drug when the target of the drug is more abundant. Last, the identification of associations with reported gene copy number alterations, such as amplification of MET and ERBB2, are also observed at the protein level (Figure 4B). Non-self interactions were also observed, such as PPA1-PPA2 paralog synthetic-lethal interaction, where cell lines with lower PPA2 abundance are more sensitive to PPA1 knockout (Figure S3A).

To identify biomarker associations that are unique to the proteome and could not be predicted by gene expression measurements alone, we developed a deep learning-based computational pipeline called Deep Proteomic Marker (DeeProM) (Figure 4C). DeeProM is powered by DeepOmicNet (see Figure S3B and STAR Methods for more details), a deep

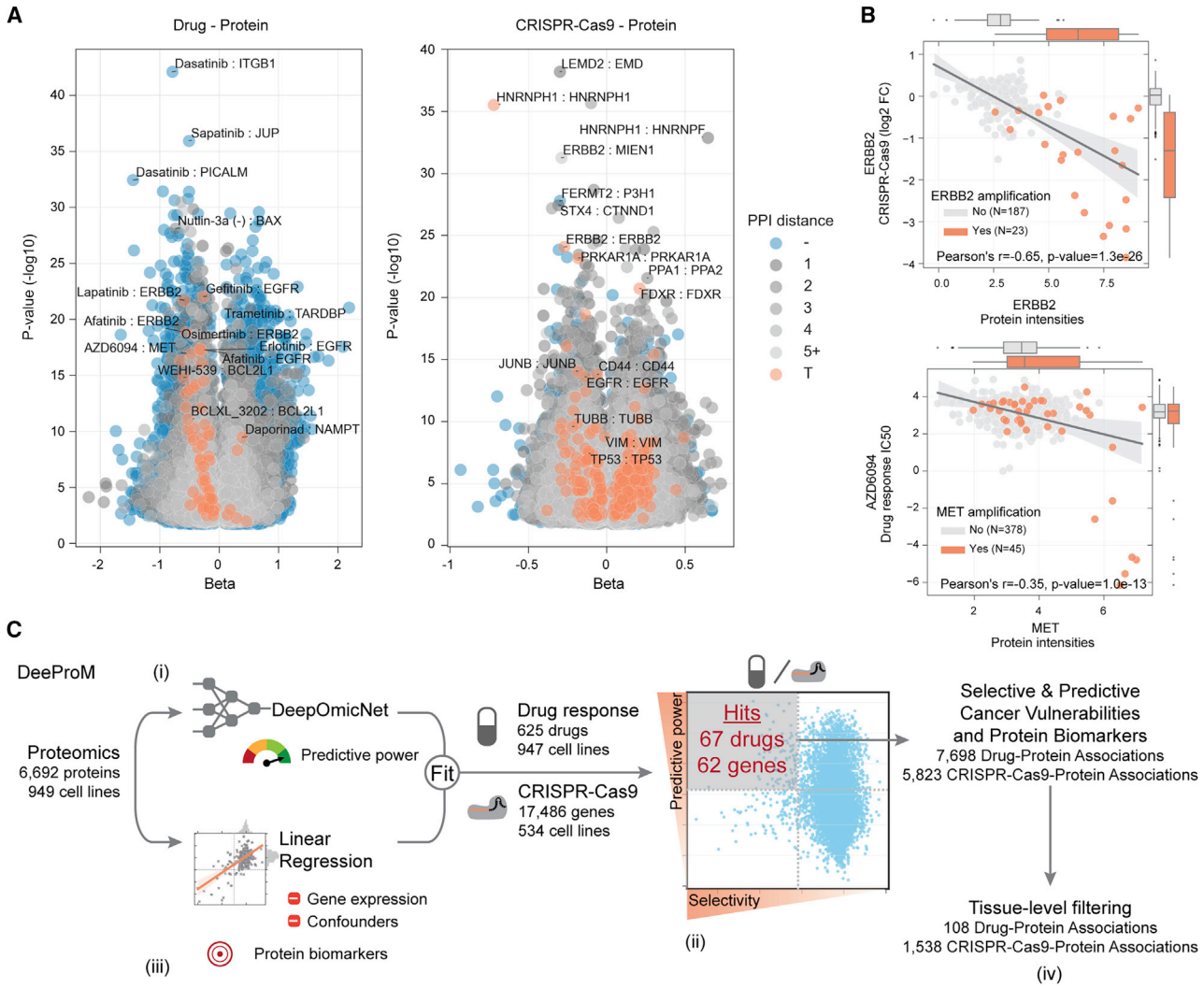


Figure 4. Biomarkers for cancer vulnerabilities

(A) Significant linear regression associations (FDR < 5%) between protein measurements and drug responses (left) and protein measurements and CRISPR-Cas9 gene essentiality scores (right). Each association is represented using the linear regression effect size (beta) and its statistical significance (log ratio test), and colored according to the distance between the target of the drug or CRISPR-Cas9 and the associated protein in a PPI network assembled from STRING. T denotes the associated protein is either a canonical target of the drug or the CRISPR-Cas9 reagents; numbers represent the minimal number of interactions separating the drug or CRISPR-Cas9 targets to the associated proteins; and the symbol '-' denotes associations for which no path was found. Representative examples are labeled.

(B) Representative top-ranked CRISPR-Cas9-protein and drug-protein associations. The top shows ERBB2 protein intensities associated with CRISPR-Cas9 gene essentiality, where cell lines with ERBB2 amplifications are highlighted in orange. The bottom shows the association between AZD6094 MET inhibitor and MET protein intensities, where MET amplified cell lines are highlighted in orange. Box-and-whisker plots indicate interquartile range (IQR) with a line at the median. Whiskers represent the minimum and maximum values at 1.5 × IQRs.

(C) Overview of the DeeProM workflow: (i) deep learning models of DeepOmicNet were trained to predict drug responses and CRISPR-Cas9 gene essentialities, prioritizing those that are best predicted by proteomic profiles; and (ii) Fisher-Pearson coefficient of skewness was calculated to identify drug responses and CRISPR-Cas9 gene essentialities that selectively occur in subsets of cancer cell lines. The selected candidates from (i) and (ii) are illustrated by the gray box. (iii) Linear regression models were fitted to identify significant associations between protein biomarkers, drug responses and CRISPR-Cas9 gene essentialities. (iv) Filtering algorithms were applied to further identify tissue-specific cancer vulnerabilities.

See also Figure S3 and Table S5.

neural network architecture designed to prioritize drug responses and CRISPR-Cas9 gene essentialities that are highly predictive and specific to subsets of cancer cell lines. In addition, DeeProM incorporates results from linear models described above to highlight biomarkers that are only evident at the proteomic level. As a benchmark, we found that the accuracy of DeepOmicNet, evaluated using Pearson's *r* between the observed and predicted IC_{50} values, consistently outperformed other machine learning approaches such as elastic net and Random Forest, across a range of multi-omic datasets used in previous studies (Iorio et al., 2016) (Figure S3C). DeeProM assessed all possible drug-protein (n = 4,218,788) and

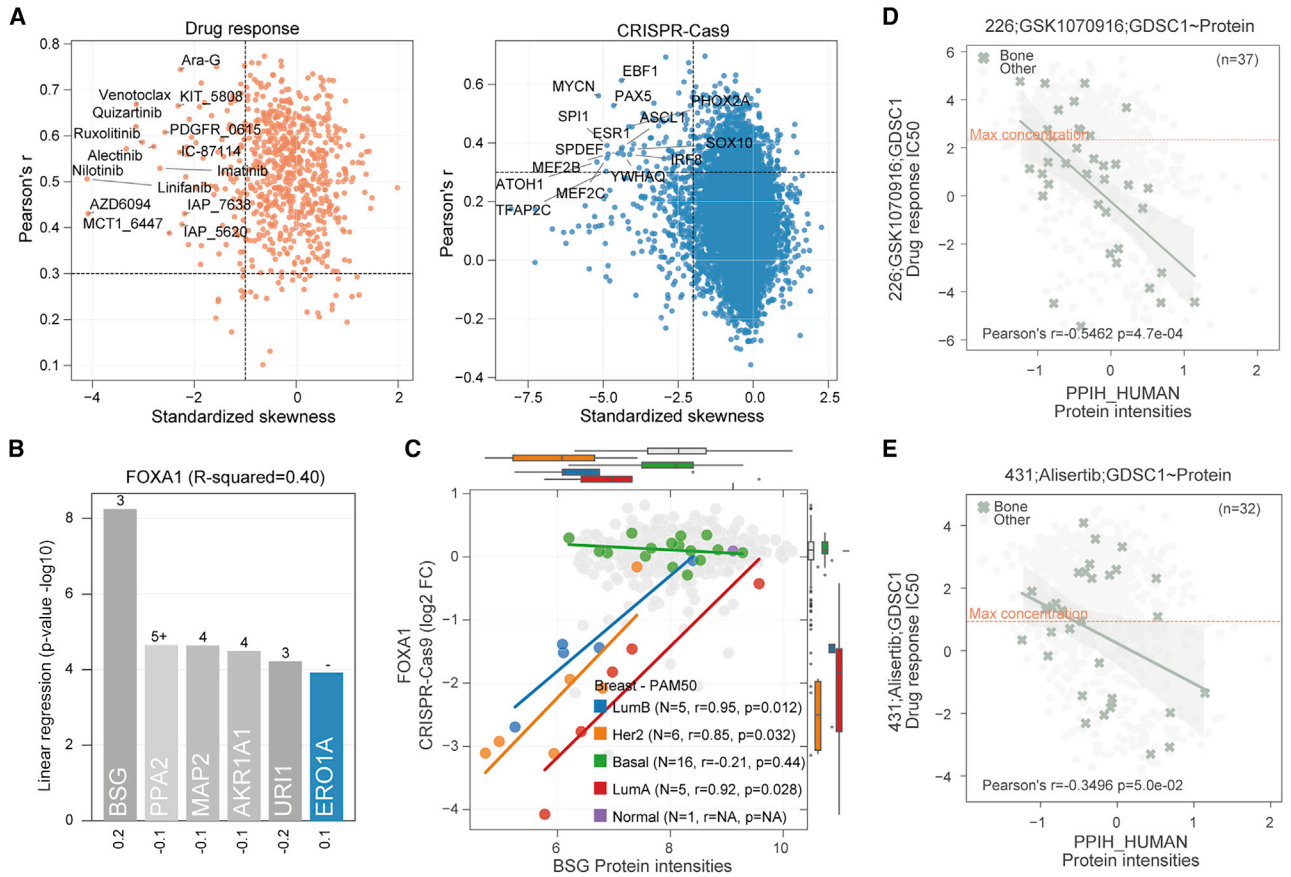


Figure 5. Protein biomarkers identified by DeeProM

(A) Predictive performance and selectivity of all drug responses (left) and CRISPR-Cas9 gene essentialities (right) across 947 and 534 cancer cell lines, respectively. Data points toward the top left corner of each plot indicate drug responses or gene essentialities that are both selective and well predicted. Top selective drugs and CRISPR-Cas9 gene essentialities are labeled.

(B) Top significant protein associations with FOXA1 CRISPR-Cas9 gene essentiality scores, each bar representing the statistical significance (log ratio test) of the linear regression, and below the effect size (beta). The minimal distance of PPIs in the STRING network between FOXA1 and each protein is annotated in each respective bar and color coded according to the description in Figure 4A.

(C) Association between FOXA1 CRISPR-Cas9 gene essentiality scores and BSG protein intensities. Breast cancer cell lines are highlighted and sub-classified using the PAM50 gene expression signature (Parker et al., 2009). Box-and-whisker plots indicate the PAM50 subtypes of breast cancers. Pearson's *r* (r), p value (p), and number of observations/cell lines (N) within each PAM50 type is provided; for "normal" subtype no correlation was performed considering N is 1. These plots indicate interquartile range (IQR) with a line at the median. Whiskers represent the minimum and maximum values at 1.5 × IQRs.

(D and E) Representative examples of tissue-specific associations between drug responses and protein markers for cell lines derived from bone (green; all other cell lines are shown in gray). The number of cell lines and Pearson's *r* from the highlighted tissue type are annotated at the top right and bottom left corners, respectively. The dashed line represents the maximum concentration used in the drug response screens. (D) The GSK1070916-PPIH association in bone supported by the ProCan-DepMapSanger proteomic dataset. (E) Similar to (D), instead showing data for the drug alisertib.

See also Figure S3 and S4; Table S5.

CRISPR-protein (n = 86,584,537) associations to identify cancer vulnerabilities that are simultaneously well predicted and selective in subsets of cell lines (Figure 5A). These two selection criteria yielded 67 drug responses and 62 gene essentialities, with a total of 7,698 drug-protein and 5,823 CRISPR-Cas9-protein associations, that had significantly improved predictions when compared to models that considered gene expression measurements alone (Figure 4C and Table S5).

Promising targeted therapeutics are often developed for specific cancer types and can display tissue-specific responses. For this reason, DeeProM was used to interrogate associations at the tissue type level by applying a filtering strategy (Figure 4C and STAR Methods). Using the DeeProM workflow, we identified 1,538 tissue type-level CRISPR-Cas9-protein associations (Table S5). Among the strongest was the dependency on FOXA1 transcription factor knockout and protein levels of basigin (BSG; also known as CD147), a plasma membrane protein expressed in breast cancer cells (Figures 5B and 5C). This association was not observed at the gene expression level (Figure S3D). FOXA1-BSG association occurred in luminal (luminal A and B) and HER2-positive (non-basal) breast cancer cell lines, in which BSG protein abundance is low relative to basal cell lines (Figure 5C). BSG has been implicated in breast cancer progression (Landras et al., 2019), and is a marker of the aggressive basal-like and triple-negative subtypes, as well as being associated with poor overall survival within these patients (Liu et al.,

2018). These data support a model where BSG protein expression is associated with basal-like breast cancer cells, whereas luminal and HER2-positive breast cancer cells with low BSG expression have an increased dependency on estrogen receptor-driven FOXA1 transcriptional activity. Further work that expands the number of samples, and confirmatory studies, would be required to validate this observation.

DeeProM also identified 108 tissue type level drug-protein associations (Table S5). Filtering by the effect size, the strongest association identified was between sensitivity to Aurora kinase B/C selective inhibitor GSK1070916 and the protein abundance of peptidyl-prolyl *cis*-trans isomerase H (PPIH) in cell lines derived from bone (Figure 5D). This association was significant at the protein level, but was not significant in the transcriptome (Figure S4A). The association was further supported by examination of the Cancer Cell Line Encyclopedia (CCLE) proteomic dataset (Figures S4B and S4C) (Nusinow et al., 2020), the PRISM drug response dataset (Figure S4D) (Corsello et al., 2020), and using an independent screening of GSK1070916 in the Sanger drug sensitivity dataset (Figures S4E and S4F), in which there is a suggestive association that does not reach statistical significance due to the smaller sample size. Furthermore, there was a strong association between PPIH protein levels and Alisertib, a second Aurora kinase inhibitor (Figures 5E, S4G and S4H). PPIH and Aurora kinase A are both regulated by the p53-p21-DREAM-CDE/CHR signaling pathway (Fischer et al., 2016), supporting the identified link between Aurora kinase inhibitor sensitivity and PPIH protein levels. Elucidating the precise mechanism underlying this association will require further research.

Taken together, these results demonstrate the added value of proteomic measurements for the discovery of cancer biomarkers. We identified both established and potential cancer related biomarkers, including protein biomarkers for selective cancer vulnerabilities that were not found using gene expression measurements alone.

Predictive power of protein sub-networks on cancer cell phenotypes

We have established the utility of proteomics to identify specific biomarkers for cancer vulnerabilities. Using an independent cell line proteomic dataset from the CCLE (Nusinow et al., 2020), we observed comparable performance to the ProCan-DepMapSanger dataset when predicting three independent drug response datasets (Figures 6A and S5A) and CRISPR-Cas9 gene essentiality profiles (Figures 6B and S5B). We next compared the predictive power of proteomic and transcriptomic data for modeling drug responses and CRISPR-Cas9 gene essentialities. The predictive power of our models was highly similar when trained using either the ProCan-DepMapSanger or the transcriptomics dataset (Figure 6C). This was recapitulated by machine learning methods such as elastic net and Random Forest (Figure 6D). Notably, the predictive performance of protein measurements and transcript measurements were highly similar, and protein measurements alone outperformed the corresponding overlapping subset of the transcriptome (p value < 0.0001, two-tailed paired Student t test) (Figure S5C). Proteomic measurements further showed overall stronger protein pairwise correlations than transcriptomics or CRISPR-Cas9 gene essentialities (Figure 6E). Taking these observations together, this demonstrates that proteomics and transcriptomics share comparable predictive power and suggests that proteomics may provide additional relevant information that is not captured by transcriptomics.

To determine how predictive power is influenced by the number of proteins used in the modeling, a random downsampling analysis was performed to predict drug responses, with a decrease of 500 proteins in each step (see STAR Methods). This showed that a randomly selected subset of 1,500 proteins was able to provide 88% of the predictive power of the full dataset (mean Pearson's $r = 0.43$ at n = 1,500 proteins versus mean Pearson's $r = 0.49$ at n = 8,498 proteins) (Figure 7A). This implies that a random fraction of quantifiable proteins is sufficient to represent fundamental elements of the proteome involved in mediating key cellular phenotypes. We propose this is in part because proteins are organized into complexes and pathways with connected and co-regulated subunits (Figure 7B).

To investigate protein networks, DeeProM analyses of drug-protein and CRISPR-Cas9-protein associations were examined in the context of protein networks (Figures 7C and 7D). The strongest overall associations were observed between the drug and CRISPR-Cas9 targets and their protein intensities (Figures 7C and 7D). Additionally, CRISPR-Cas9-protein associations showed that proteins closer in the PPI network to the targeted proteins had stronger associations than those further apart (Figure 7D). This enrichment for targets and functionally closer proteins remains even when the contribution of transcriptomic measurements is removed for the drug-protein and CRISPR-Cas9-protein associations (Figures S5D and S5E). However, many seemingly functionally distant proteins (more than two steps away from the perturbation target in PPIs) also exhibit significant drug-protein and CRISPR-Cas9-protein associations (Figures 7C and 7D).

We next explored the relationship between predictive performance and sub-networks comprising proteins of differing frequencies in the full dataset, defining categories of proteins as those found in 90% or more (category A; n = 2,944 proteins), 20%–90% (category B; n = 3,939) or less than 20% (category C; n = 1,615) of the cell lines, respectively (Figure S5F). Downsampling these protein sets at random, with a decrease of 250 proteins in each step (see STAR Methods), showed that proteins that are frequently observed in the dataset (category A), provided the highest predictive performance (e.g., mean Pearson's $r = 0.463$ for 1,500 proteins, or $r = 0.435$ for 250 proteins) when compared against less frequently observed proteins (category B, mean $r = 0.441$; and category C, mean $r = 0.432$, for 1,500 proteins) (Figure 7E). Category A proteins had a significantly higher degree of connectivity (mean of 8 degrees) than category B and C proteins (mean of 2 degrees and 1 degree, respectively) in the STRING protein interaction network (Szklarczyk et al., 2017) (Figure 7F). It is possible that category A proteins, as a consequence of being more frequently observed, are better studied and therefore have greater annotation in the STRING database. However, together these results suggest that the quantification of small subsets of commonly expressed proteins within highly interconnected networks can be used for predictive modeling of cellular phenotypes.

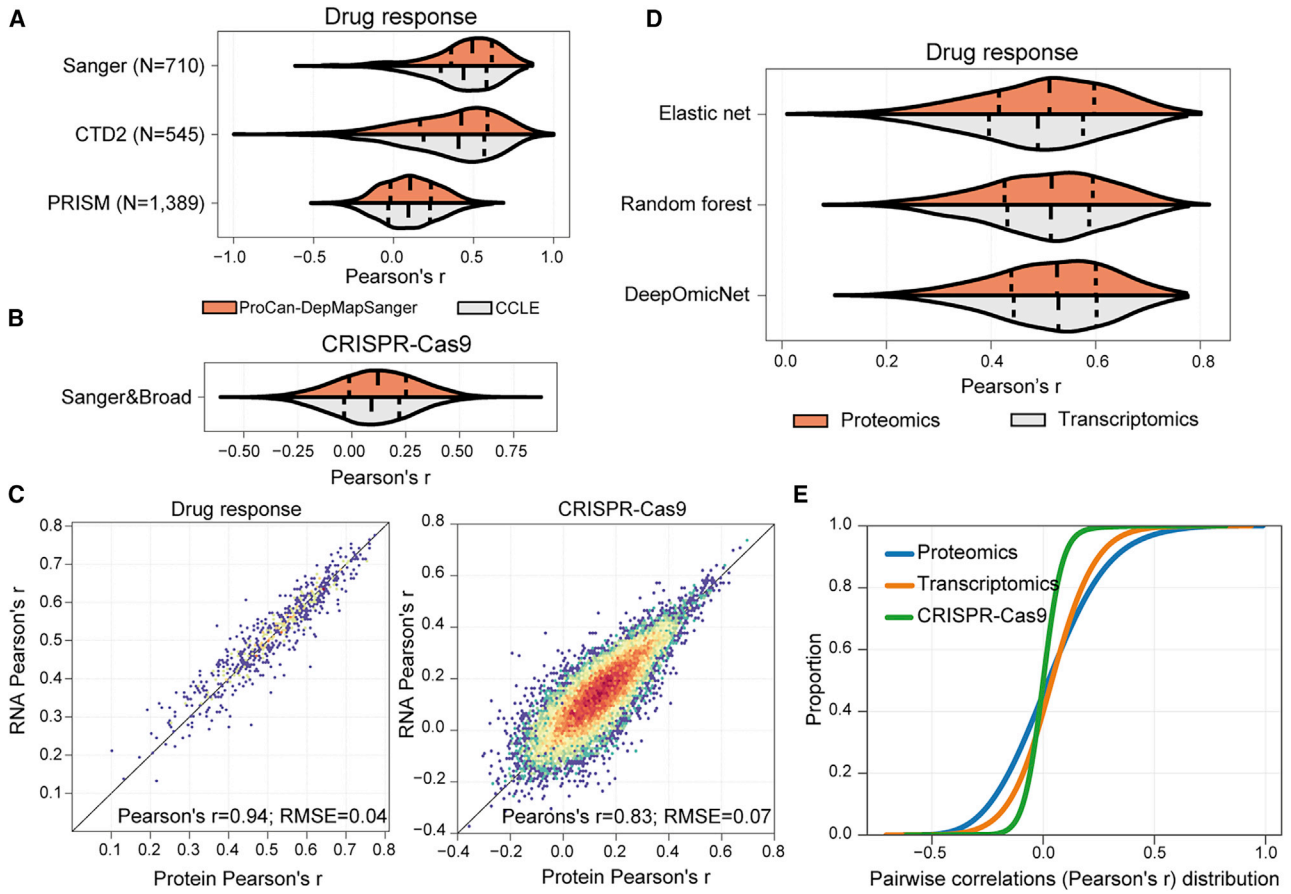


Figure 6. Evaluation of the predictive power of DeepOmicNet for multi-omic datasets

(A and B) Distribution of the predictive power (mean Pearson's *r* between predicted and observed IC_{50} values) of DeepOmicNet, comparing ProCan-DepMapSanger to an independent proteomic dataset (CCLE), using cell lines in common between the two datasets. Plots show prediction of (A) drug responses (N represents the total number of drugs tested; n = 290 cell lines) and (B) CRISPR-Cas9 gene essentialities (n = 234 cell lines).

(C) Two-dimensional density plots showing the predictive power of DeepOmicNet in predicting drug responses (left) and CRISPR-Cas9 gene essentiality profiles (right) using protein (horizontal axis) and transcript (vertical axis) measurements. Each data point denotes the Pearson's *r* between predicted and observed measurements for each drug or CRISPR-Cas9 gene essentialities.

(D) Similar to (A), distribution of the predictive power of three machine learning models using either proteomic or transcriptomic measurements to train and predict drug responses (Sanger dataset).

(E) Cumulative distribution function of the Pearson's *r* of all pairwise protein-protein correlations compared with transcriptomics and CRISPR-Cas9 gene essentiality measurements.

See also Figure S5.

DISCUSSION

The ProCan-DepMapSanger data resource is a large pan-cancer proteomic map that provides multiple insights beyond existing molecular datasets. This map quantifies 8,498 proteins across 949 human cancer cell lines, representing 28 tissues and more than 40 histologically diverse cancer types and a wide range of genotypes, significantly expanding the molecular characterization of cancer models as part of a Cancer Dependency Map (Boehm et al., 2021). All data are publicly available along with other molecular and phenotypic datasets at http://cellmodelpassports.sanger.ac.uk (van der Meer et al., 2019). This proteomic dataset is a high-quality resource for mechanistic investigation of network organization and regulatory principles of the proteome, as well as for translational discoveries.

This study demonstrated protein expression patterns reflecting cell lineage and, potentially, EMT. The data also revealed widespread protein regulatory events, such as post-transcriptional attenuation of gene copy number effects. ProCan-DepMapSanger allowed the comprehensive characterization of protein expression patterns that could not be captured by the transcriptome, exposing the benefits of directly measuring protein abundance. Furthermore, we developed a deep learning-based pipeline DeeProM, with a deep neural network architecture, which consistently outperformed other machine learning approaches. DeeProM enabled the full integration of proteomic data with drug responses and CRISPR-Cas9 gene essentiality screens to build a comprehensive map of protein-specific biomarkers of cancer vulnerabilities that are essential for cancer cell survival and growth. Notably, this demonstrates that proteomic data spanning a broad range of cancer cell types

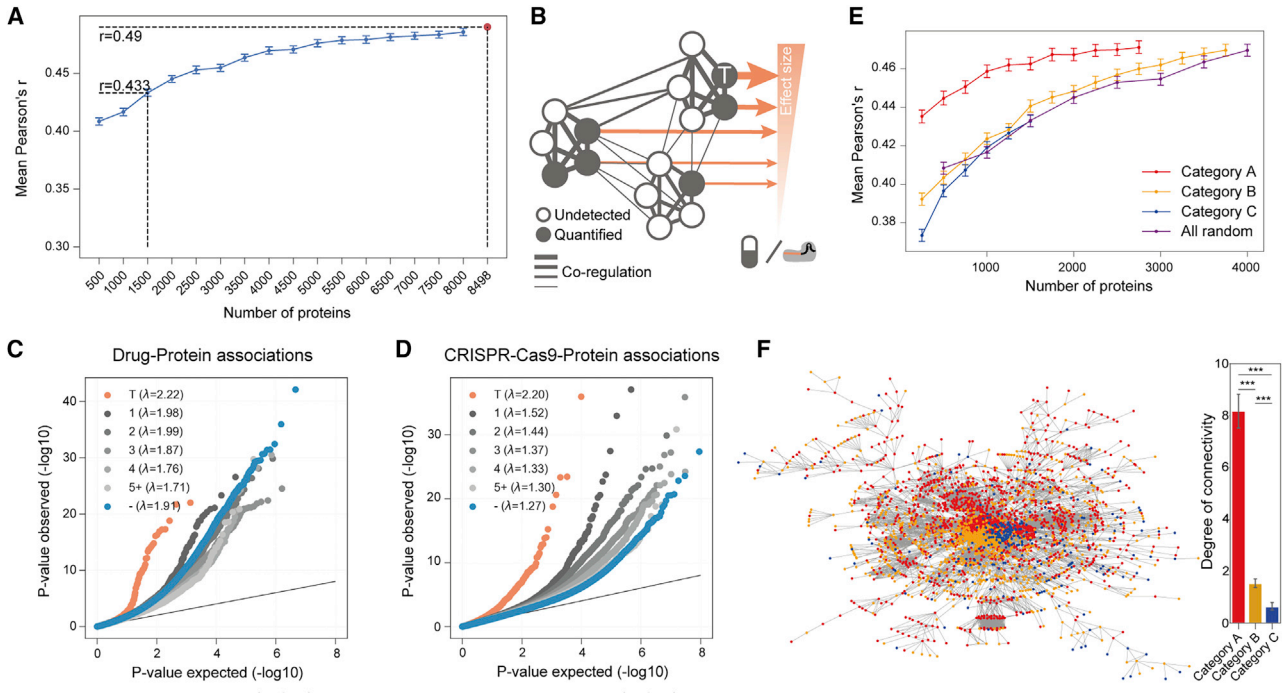


Figure 7. Proteomic support for a network pleiotropy model

(A) Comparison of the predictive power of DeepOmicNet trained with randomly downsampled sets of proteins. The dots indicate the means and vertical lines represent 95% confidence intervals derived from 10 iterations of random downsampling. The red point represents the full predictive power using all of 8,498 quantified proteins.

(B) Schematic diagram depicting protein network pleiotropy with widespread protein associations with responses to either drugs or CRISPR-Cas9, and demonstrating the strongly co-regulated nature of protein networks. Nodes represent proteins that could be either quantified or are undetected, where T represents a protein target of a drug or CRISPR-Cas9 gene essentialities. Edges showcase putative interactions, with high correlation coefficients between proteins depicted by thicker edges. Orange arrows represent the variability explained by that protein for the cancer cell line's response to a drug or CRISPR-Cas9 gene perturbation. The size of the arrow is proportional to the variance explained.

(C and D) Quantile-quantile plots of protein associations with (C) drug responses and (D) CRISPR-Cas9 gene essentiality profiles. Protein associations are grouped and colored by their distances from the drugs or CRISPR-Cas9 targets using the STRING protein interaction network, where '-' and the blue circles denote associations for which no link in the protein network could be found between the protein and the drug or CRISPR target. The p values were calculated in likelihood ratio tests on all parameters of the linear regression models. Annotation is as described in Figure 4A. For each group, the p value distribution inflation factor lambda, λ, was calculated using the median method (Aulchenko et al., 2007).

(E) Comparison of the predictive power of DeepOmicNet models trained with subsets of Category A, B and C proteins (per Figure S5F) comprising randomly downsampled sets of proteins. The dots indicate the means and vertical lines represent 95% confidence intervals derived from 10 iterations of random samplings.

(F) The STRING protein interaction network diagram (left), with proteins colored according to category. The bar chart (right) shows the network connectivity for these proteins, where degree represents the number of other proteins connected to a given protein according to the STRING PPI network. ***$p < 0.001$ by unpaired t test. Error bars represent 95% confidence intervals.

See also Figure S5.

and molecular backgrounds has significant utility for predicting cancer cell vulnerabilities.

The proteomic workflow used in this study was devised to be clinically relevant, so that our methods could be readily applied for use in human cancer tissue samples. To do so, we used shortened preparation times, low peptide loads, and a short liquid chromatography (LC)/MS run time, enabling the analysis of large numbers of very small cancer samples, with high throughput and minimal instrument downtime. This will facilitate future validation of proteomic predictions from cell line data in clinical sample cohorts for which outcome of treatment is documented and proteomic data are obtained. The CCLE proteomic dataset was generated using higher peptide loads and longer LC/MS run time to measure more proteins (12,755 proteins). Despite the different depths of protein coverage, the CCLE and ProCan-DepMapSanger proteomic datasets had equivalent power for predicting cancer dependencies. Similarly, the ProCan-DepMapSanger proteomic dataset had similar predictive power to cancer cell line transcriptomic data. Taken together, this demonstrates that a high-throughput sample workflow, as used in this study, produces data with power to inform predictions of cancer dependencies and indicates the potential of proteomics for clinical applications using small biopsies of human cancer tissue in diverse molecular contexts. Subsequent application of this proteomics sample workflow, and integration of this cell line dataset with proteomic data from cancer tissue samples, is likely to provide numerous potential clinical applications, such as the proteomic molecular identification and stratification of cancer subtypes.

Measuring even a fraction of the proteome, as small as 1,500 randomly selected proteins, provided power to predict drug

responses that were similar to the full proteome that we report. This suggests that random subsets of protein data comprising a relatively small number of proteins would be sufficient to represent many fundamental cellular processes. This is consistent with an omnigenic model (Boyle et al. 2017), whereby large numbers of genes are related to many different disease traits in an interconnected manner. This is related to the proteostasis network model of sustaining proteome balance via coordinated protein synthesis, folding, conformation and degradation (Gumeni et al., 2017). In the context of the cancer proteome, we propose that pleiotropic networks of highly connected and co-regulated proteins contribute toward establishing cellular phenotypes. This includes a small number of core protein modules that are proximal to the phenotype and have the strongest effect, and a much larger set of more distal proteins that together explain a significant portion of total variation.

In conclusion, this dataset represents a major resource for the scientific community, for biomarker discovery and for the study of fundamental aspects of protein regulation that are not evident from existing molecular datasets. This will enable the identification of targets (including cell surface proteins) and treatments for validation in cancer tissue cohorts, with applications in precision oncology.

STAR★METHODS

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

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Cell lines
- METHOD DETAILS
  - Cell lysis and digestion
  - Data independent acquisition (DIA)-MS
  - Spectral library and DIA-MS data processing
  - Assembly of multi-omics cell line datasets
  - Dimensionality reduction and visualization
  - Multi-omics factor analysis
  - Pairwise protein-protein correlations
  - DeeProM (Deep Proteomic Marker) - Overview
  - DeeProM - DeepOmicNet model
  - DeeProM - Prioritizing selective associations
  - DeeProM - Linear regression models
  - DeeProM - Tissue type level filtering
  - Comparing DeepOmicNet and other models
  - Machine learning for lethality prediction
  - Predictive power comparison with CCLE
  - Downsampling for drug response prediction
  - Figures
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.ccell.2022.06.010.

ACKNOWLEDGMENTS

We thank Ricard Argelaguet for helpful comments and discussions on the implementations of the MOFA analysis and Keith Ashman for discussions on optimizing the bank of mass spectrometers. ProCan is supported by the Australian Cancer Research Foundation, Cancer Institute New South Wales (NSW) (2017/TPG001, REG171150), NSW Ministry of Health (CMP-01), The University of Sydney, Cancer Council NSW (IG 18-01), Ian Potter Foundation, the Medical Research Future Fund (MRFF-PD), National Health and Medical Research Council (NHMRC) of Australia European Union grant (GNT1170739, a companion grant to support the European Commission's Horizon 2020 Program, H2020-SC1-DTH-2018-1, 'iPC - individualizedPaediatricCure' [ref. 826121]), and National Breast Cancer Foundation (IIRS-18-164). The work at ProCan was done under the auspices of a Memorandum of Understanding between Children's Medical Research Institute and the U.S. National Cancer Institute's International Cancer Proteogenomics Consortium (ICPC), that encourages cooperation among institutions and nations in proteogenomic cancer research in which datasets are made available to the public. R.C.P. and P.J.R. are supported by NHMRC Fellowships (GNT1138536 and GNT1137064, respectively). E.G.'s work is supported by UIDB/50021/2020 (INESC-ID multi-annual funding). This research was funded in whole, or in part, by the Wellcome Trust Grant 206194. For the purpose of Open Access, the authors have applied a CC BY public copyright license to any Author Accepted Manuscript version arising from this submission. Graphical abstract created with BioRender.com.

AUTHOR CONTRIBUTIONS

Conceptualization: P.J.R., Q.Z., M.J.G., R.R.R. Methodology: E.G., R.C.P., Z.C., P.G.H., P.J.R., Q.Z., M.J.G., R.R.R.; Software: E.G., R.C.P., Z.C.; Formal analysis: E.G., R.C.P., Z.C., Q.Z.; Investigation: E.G., R.C.P., Z.C., S.S.M., N.L., A.B., D.B.N., C.H., J.K., S.M., I.M., J.M., L.R., A.J.S., E.S., F.T., S.G.W., Y.W., D.X., K.L.M., Q.Z.; Data curation: S.B., M.D., M.H., K.L.M., B.T., H.L., R.S.; Project administration: S.V.; Writing – original draft: E.G., R.C.P., Z.C., P.J.R., Q.Z., M.J.G., R.R.R.; Writing – review and editing: All authors.

DECLARATION OF INTERESTS

M.J.G. has received research funding from AstraZeneca, GSK, Astex Therapeutics, and Open Targets, a public-private initiative involving academia and industry, and is a co-founder of Mosaic Therapeutics. All other authors declare no competing interests.

Received: December 14, 2021
Revised: March 29, 2022
Accepted: June 21, 2022
Published: July 14, 2022

REFERENCES

Argelaguet, R., Arnol, D., Bredikhin, D., Deloro, Y., Velten, B., Marioni, J.C., and Stegle, O. (2020). MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. Genome Biol. *21*, 111.

Argelaguet, R., Velten, B., Arnol, D., Dietrich, S., Zenz, T., Marioni, J.C., Buettner, F., Huber, W., and Stegle, O. (2018). Multi-omics factor Analysis-a framework for unsupervised integration of multi-omics data sets. Mol. Syst. Biol. *14*, e8124.

Aulchenko, Y.S., Ripke, S., Isaacs, A., and van Duijn, C.M. (2007). GenABEL: an R library for genome-wide association analysis. Bioinformatics *23*, 1294–1296.

Barretina, J., Giordano, C., Stransky, N., Venkatesan, K., Margolin, A.A., Kim, S., Wilson, C.J., Lehár, J., Kryukov, G.V., Sonkin, D., et al. (2012). The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature *483*, 603–607.

Behan, F.M., Iorio, F., Picco, G., Gonçalves, E., Beaver, C.M., Migliardi, G., Santos, R., Rao, Y., Sassi, F., Pinnelli, M., et al. (2019). Prioritization of cancer therapeutic targets using CRISPR-cas9 screens. Nature *568*, 511–516.

Boehm, J.S., Garnett, M.J., Adams, D.J., Francies, H.E., Golub, T.R., Hahn, W.C., Iorio, F., McFarland, J.M., Parts, L., and Vazquez, F. (2021). Cancer research needs a better map. Nature *589*, 514–516.

Boyle, E.A., Li, Y.I., and Pritchard, J.K. (2017). An expanded view of complex traits: from polygenic to omnigenic. Cell *169*, 1177–1186.

Brabletz, T., Kalluri, R., Nieto, M.A., and Weinberg, R.A. (2018). EMT in cancer. Nat. Rev. Cancer *18*, 128–134.

Chatr-Aryamontri, A., Breitkreutz, B.-J., Oughtred, R., Boucher, L., Heinicke, S., Chen, D., Stark, C., Breitkreutz, A., Kolas, N., O'Donnell, L., et al. (2015). The BioGRID interaction database: 2015 update. Nucleic Acids Res. *43*, D470–D478.

Clark, D.J., Dhanasekaran, S.M., Petralia, F., Pan, J., Song, X., Hu, Y., da Veiga Leprevost, F., Reva, B., Lih, T.S.M., Chang, H.Y., et al. (2019). Integrated proteogenomic characterization of clear cell renal cell carcinoma. Cell *179*, 964–983.e31.

Corsello, S.M., Nagari, R.T., Spangler, R.D., Jordan, R., Kocak, M., Bryan, J.G., Humeidi, R., Peck, D., Wu, X., Tang, A.A., et al. (2020). Discovering the anticancer potential of non-oncology drugs by systematic viability profiling. Nature Cancer *1*, 235–248.

Coscia, F., Watters, K.M., Curtis, M., Eckert, M.A., Chiang, C.Y., Tyanova, S., Montag, A., Lastra, R.R., Lengyel, E., and Mann, M. (2016). Integrative proteomic profiling of ovarian cancer cell lines reveals precursor cell associated proteins and functional status. Nat. Commun. *7*, 12645.

Cox, J., Hein, M.Y., Luber, C.A., Paron, I., Nagaraj, N., and Mann, M. (2014). Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. Mol. Cell. Proteomics *13*, 2513–2526.

Dandage, R., and Landry, C.R. (2019). Paralog dependency indirectly affects the robustness of human cells. Mol. Syst. Biol. *15*, e8871.

Demichev, V., Messner, C.B., Vernardis, S.I., Lilley, K.S., and Ralser, M. (2020). DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. Nat. Methods *17*, 41–44.

Edwards, N.J., Oberti, M., Thangudu, R.R., Cai, S., McGarvey, P.B., Shine, J., Madhavan, S., and Ketchum, K.K. (2015). The CPTAC data portal: a resource for cancer proteomics research. J. Proteome Res. *14*, 2707–2713.

Fischer, M., Quaas, M., Steiner, L., and Engeland, K. (2016). The p53-p21-DREAM-CDE/CHR pathway regulates G2/M cell cycle genes. Nucleic Acids Res. *44*, 164–174.

Frejno, M., Chiozzi, R.Z., Wilhelm, M., Koch, H., Zheng, R., Klaeger, S., Ruprecht, B., Meng, C., Kramer, K., Jarzab, A., et al. (2017). Pharmacoproteomic characterisation of human colon and rectal cancer. Mol. Syst. Biol. *13*, 951.

Frejno, M., Meng, C., Ruprecht, B., Oellerich, T., Scheich, S., Kleigrewe, K., Drecoll, E., Samaras, P., Hogrebe, A., Helm, D., et al. (2020). Proteome activity landscapes of tumor cell lines determine drug responses. Nat. Commun. *11*, 3639.

Garcia-Alonso, L., Iorio, F., Matchan, A., Fonseca, N., Jaaks, P., Peat, G., Pignatelli, M., Falcone, F., Benes, C.H., Dunham, I., et al. (2018). Transcription factor activities enhance markers of drug sensitivity in cancer. Cancer Res. *78*, 769–780.

Garnett, M.J., Edelman, E.J., Heidorn, S.J., Greenman, C.D., Dastur, A., Lau, K.W., Greninger, P., Thompson, I.R., Luo, X., Soares, J., et al. (2012). Systematic identification of genomic markers of drug sensitivity in cancer cells. Nature *483*, 570–575.

Ghandi, M., Huang, F.W., Jané-Valbuena, J., Kryukov, G.V., Lo, C.C., McDonald, R., 3rd, Barretina, J., Gelfand, E.T., Bielski, C.M., Li, H., et al. (2019). Next-generation characterization of the Cancer Cell Line Encyclopedia. Nature *569*, 503–508.

Gholami, A.M., Hahne, H., Wu, Z., Auer, F.J., Meng, C., Wilhelm, M., and Kuster, B. (2013). Global proteome analysis of the NCI-60 cell line panel. Cell Rep. *4*, 609–620.

Gillet, L.C., Navarro, P., Tate, S., Röst, H., Selevsek, N., Reiter, L., Bonner, R., and Aebersold, R. (2012). Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. Mol. Cell. Proteomics *11*. O111–016717.

Gonçalves, E., Fragoulis, A., Garcia-Alonso, L., Cramer, T., Saez-Rodriguez, J., and Beltrao, P. (2017). Widespread post-transcriptional attenuation of genomic copy-number variation in cancer. Cell Syst. *5*, 386–398.e4.

Gonçalves, E., Segura-Cabrera, A., Pacini, C., Picco, G., Behan, F.M., Jaaks, P., Coker, E.A., van der Meer, D., Barthorpe, A., Lightfoot, H., et al. (2020). Drug mechanism-of-action discovery through the integration of pharmacological and CRISPR screens. Mol. Syst. Biol. *16*, e9405.

Gumeni, S., Evangelakou, Z., Gorgoulis, V.G., and Trougakos, I.P. (2017). Proteome stability as a key factor of genome integrity. Int. J. Mol. Sci. *18*. https://doi.org/10.3390/ijms18102036.

Guo, T., Kouvonen, P., Koh, C.C., Gillet, L.C., Wolski, W.E., Röst, H.L., Rosenberger, G., Collins, B.C., Blum, L.C., Gillessen, S., et al. (2015). Rapid mass spectrometric conversion of tissue biopsy samples into permanent quantitative digital proteome maps. Nat. Med. *21*, 407–413.

Guo, T., Luna, A., Rajapakse, V.N., Koh, C.C., Wu, Z., Liu, W., Sun, Y., Gao, H., Menden, M.P., Xu, C., et al. (2019). Quantitative proteome landscape of the NCI-60 cancer cell lines. iScience *21*, 664–680.

Hart, T., Chandrashekhar, M., Aregger, M., Steinhart, Z., Brown, K.R., MacLeod, G., Mis, M., Zimmermann, M., Fradet-Turcotte, A., Sun, S., et al. (2015). High-resolution CRISPR screens reveal fitness genes and genotype-specific cancer liabilities. Cell *163*, 1515–1526.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In 2016 IEEE conference on computer vision and pattern recognition (CVPR) (IEEE), pp. 770–778.

Huang, K.-L., Li, S., Mertins, P., Cao, S., Gunawardena, H.P., Ruggles, K.V., Mani, D.R., Clauser, K.R., Tanioka, M., Usary, J., et al. (2017). Proteogenomic integration reveals therapeutic targets in breast cancer xenografts. Nat. Commun. *8*, 14864.

Hunter, J.D. (2007). Matplotlib: a 2D graphics environment. Comput. Sci. Eng. *9*, 90–95.

Iorio, F., Behan, F.M., Gonçalves, E., Bhosle, S.G., Chen, E., Shepherd, R., Beaver, C., Ansari, R., Pooley, R., Wilkinson, P., et al. (2018). Unsupervised correction of gene-independent cell responses to CRISPR-Cas9 targeting. BMC Genom. *19*, 604.

Iorio, F., Knijnenburg, T.A., Vis, D.J., Bignell, G.R., Menden, M.P., Schubert, M., Aben, N., Harper, S., Butler, A.P., Stronach, E.A., et al. (2016). A landscape of pharmacogenomic interactions in cancer. Cell *166*, 740–754.

Landras, A., Reger de Moura, C., Jouenne, F., Lebbe, C., Menashi, S., and Mourah, S. (2019). CD147 is a promising target of tumor progression and a prognostic biomarker. Cancers *11*, 1803.

Lawrence, R.T., Perez, E.M., Hernández, D., Miller, C.P., Haas, K.M., Irie, H.Y., Lee, S.I., Blau, C.A., and Villén, J. (2015). The proteomic landscape of triple-negative breast cancer. Cell Rep. *11*, 630–644.

Li, J., Zhao, W., Akbani, R., Liu, W., Ju, Z., Ling, S., Vellano, C.P., Roebuck, P., Yu, Q., Eterovic, A.K., et al. (2017). Characterization of human cancer cell lines by reverse-phase protein arrays. Cancer Cell *31*, 225–239.

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdottir, H., Tamayo, P., and Mesirov, J.P. (2011). Molecular signatures database (MSigDB) 3.0. Bioinformatics *27*, 1739–1740.

Liu, Y., Beyer, A., and Aebersold, R. (2016). On the dependency of cellular protein levels on mRNA abundance. Cell *165*, 535–550.

Liu, M., Tsang, J.Y.S., Lee, M., Ni, Y.-B., Chan, S.-K., Cheung, S.-Y., Hu, J., Hu, H., and Tse, G.M.K. (2018). CD147 expression is associated with poor overall survival in chemotherapy treated triple-negative breast cancer. J. Clin. Pathol. *71*, 1007–1014.

Lucas, N., Robinson, A.B., Marcker Espersen, M., Mahboob, S., Xavier, D., Xue, J., Balleine, R.L., deFazio, A., Hains, P.G., and Robinson, P.J. (2019). Accelerated barocycler lysis and extraction sample preparation for clinical proteomics by mass spectrometry. J. Proteome Res. *18*, 399–405.

Luck, K., Kim, D.-K., Lambourne, L., Spirohn, K., Begg, B.E., Bian, W., Brignall, R., Cafarelli, T., Campos-Laborie, F.J., Charloteaux, B., et al.

(2020). A reference map of the human binary protein interactome. Nature *580*, 402–408.

Ludwig, C., Gillet, L., Rosenberger, G., Amon, S., Collins, B.C., and Aebersold, R. (2018). Data-independent acquisition-based SWATH-MS for quantitative proteomics: a tutorial. Mol. Syst. Biol. *14*, e8126.

McDonald, E.R., 3rd, de Weck, A., Schlabach, M.R., Billy, E., Mavrakis, K.J., Hoffman, G.R., Belur, D., Castelletti, D., Frias, E., Gampa, K., et al. (2017). Project DRIVE: a compendium of cancer dependencies and synthetic lethal relationships uncovered by large-scale, deep RNAi screening. Cell *170*, 577–592.e10.

McInnes, L., Healy, J., Saul, N., and Großberger, L. (2018). UMAP: Uniform manifold approximation and projection. J. Open Source Soft. *3*, 861.

Mertins, P., Mani, D.R., Ruggles, K.V., Gillette, M.A., Clauser, K.R., Wang, P., Wang, X., Qiao, J.W., Cao, S., Petralia, F., et al. (2016). Proteogenomics connects somatic mutations to signalling in breast cancer. Nature *534*, 55–62.

Meyers, R.M., Bryan, J.G., McFarland, J.M., Weir, B.A., Sizemore, A.E., Han, X., Dharia, N.V., Montgomery, P.G., Cowley, G.S., Pantel, S., et al. (2017). Computational correction of copy number effect improves specificity of CRISPR-cas9 essentiality screens in cancer cells. Nat. Genet. *49*, 1779–1784.

Nusinow, D.P., Szpyt, J., Ghandi, M., Rose, C.M., McDonald, E.R., Kalocsay, M., Jané-Valbuena, J., Gelfand, E., Schweppe, D.K., Jedrychowski, M., et al. (2020). Quantitative proteomics of the Cancer Cell Line Encyclopedia. Cell *180*, 387–402.e16.

Pacini, Clare, Dempster, J.M., Boyle, I., Gonçalves, E., Najgebauer, H., Karakoc, E., van der Meer, D., Barthorpe, A., Lightfoot, H., Jaaks, P., et al. (2021). Integrated cross-study datasets of genetic dependencies in cancer. Nat. Commun. *12*, 1661.

Parker, J.S., Mullins, M., Cheang, M.C.U., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., Hu, Z., et al. (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. J. Clin. Oncol. *27*, 1160–1167.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019). Pytorch: an imperative style, high-performance deep learning library. Adv. Neural Inf. Process. Syst. *32*. https://proceedings.neurips.cc/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: machine learning in Python. J. Mach. Learn. Res. *12*, 2825–2830.

Perez-Riverol, Y., Csordas, A., Bai, J., Bernal-Llinares, M., Hewapathirana, S., Kundu, D.J., Inuganti, A., Griss, J., Mayer, G., Eisenacher, M., et al. (2019). The PRIDE database and related tools and resources in 2019: improving support for quantification data. Nucleic Acids Res. *47*, D442–D450.

Picco, G., Chen, E.D., Garcia Alonso, L., Behan, F.M., Gonçalves, E., Bignall, G., Matchan, A., Fu, B., Banerjee, R., Anderson, E., et al. (2019). Functional linkage of gene fusions to cancer cell fitness assessed by pharmacological and CRISPR-cas9 screening. Nat. Commun. *10*, 2198.

Poulos, R.C., Hains, P.G., Shah, R., Lucas, N., Xavier, D., Manda, S.S., Anees, A., Koh, J.M.S., Mahboob, S., Wittman, M., et al. (2020). Strategies to enable large-scale proteomics for reproducible research. Nat. Commun. *11*, 3793.

Pozniak, Y., Balint-Lahat, N., Rudolph, J.D., Lindskog, C., Katzir, R., Avivi, C., Pontén, F., Ruppin, E., Barshack, I., and Geiger, T. (2016). System-wide clinical proteomics of breast cancer reveals global remodeling of tissue homeostasis. Cell Syst. *2*, 172–184.

Reback, J., jbrockmendel, McKinney, W., Van den Bossche, J., Augspurger, T., Roeschke, M., Hawkins, S., Cloud, P., gfyoung, Sinhrks, et al. (2022). Pandas-Dev/pandas: Pandas 1.4.2. Zenodo. https://doi.org/10.5281/ZENODO.3509134.

Rees, M.G., Seashore-Ludlow, B., Cheah, J.H., Adams, D.J., Price, E.V., Gill, S., Javaid, S., Coletti, M.E., Jones, V.L., Bodycombe, N.E., et al. (2016). Correlating chemical sensitivity and basal gene expression reveals mechanism of action. Nat. Chem. Biol. *12*, 109–116.

Roumeliotis, T.I., Williams, S.P., Gonçalves, E., Alsinet, C., Del Castillo Velasco-Herrera, M., Aben, N., Ghavidel, F.Z., Michaut, M., Schubert, M., Price, S., et al. (2017). Genomic determinants of protein abundance variation in colorectal cancer cells. Cell Rep. *20*, 2201–2214.

Ruepp, A., Waegele, B., Lechner, M., Brauner, B., Dunger-Kaltenbach, I., Fobo, G., Frishman, G., Montrone, C., and Mewes, H.W. (2010). CORUM: the comprehensive resource of mammalian protein complexes–2009. Nucleic Acids Res. *38*, D497–D501.

Ryan, C.J., Kennedy, S., Bajrami, I., Matallanas, D., and Lord, C.J. (2017). A compendium of co-regulated protein complexes in breast cancer reveals collateral loss events. Cell Syst. *5*, 399–409.

Seashore-Ludlow, B., Rees, M.G., Cheah, J.H., Cokol, M., Price, E.V., Coletti, M.E., Jones, V., Bodycombe, N.E., Soule, C.K., Gould, J., et al. (2015). Harnessing connectivity in a large-scale small-molecule sensitivity dataset. Cancer Discov. *5*, 1210–1223.

Sousa, A., Gonçalves, E., Mirauta, B., Ochoa, D., Stegle, O., and Beltrao, P. (2019). Multi-omics characterization of interaction-mediated control of human protein abundance levels. Mol. Cell. Proteomics *18*, S114–S125.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. of Sci. USA *102*, 15545–15550.

Szklarczyk, D., Morris, J.H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., Santos, A., Doncheva, N.T., Roth, A., Bork, P., et al. (2017). The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. Nucleic Acids Res. *45*, D362–D368.

Tsherniak, A., Vazquez, F., Montgomery, P.G., Weir, B.A., Kryukov, G., Cowley, G.S., Gill, S., Harrington, W.F., Pantel, S., Krill-Burger, J.M., et al. (2017). Defining a cancer dependency map. Cell *170*, 564–576.e16.

Tully, B., Balleine, R.L., Hains, P.G., Zhong, Q., Reddel, R.R., and Robinson, P.J. (2019). Addressing the challenges of high-throughput cancer tissue proteomics for clinical application: ProCan. Proteomics *19*, e1900109.

Tully, B. (2020). Toffee - a highly efficient, lossless file format for DIA-MS. Sci. Rep. *10*, 8939.

van der Meer, D., Barthorpe, S., Yang, W., Howard, L., Hall, C., Gilbert, J., Francies, H.E., and Garnett, M.J. (2019). Cell model passports-a hub for clinical, genetic and functional datasets of preclinical cancer models. Nucleic Acids Res. *47*, D923–D929.

Vasaikar, S., Huang, C., Wang, X., Petyuk, V.A., Savage, S.R., Wen, B., Dou, Y., Zhang, Y., Shi, Z., Arshad, O.A., et al. (2019). Proteogenomic analysis of human colon cancer reveals new therapeutic opportunities. Cell *177*, 1035–1049.e19.

Vijayakumaran, R., Tan, K.H., Miranda, P.J., Haupt, S., and Haupt, Y. (2015). Regulation of mutant p53 protein expression. Front. Oncol. *5*, 284.

Vis, D.J., Bombardelli, L., Lightfoot, H., Iorio, F., Garnett, M.J., and Wessels, L.F. (2016). Multilevel models improve precision and speed of IC50 estimates. Pharmacogenomics *17*, 691–700.

Waskom, M. (2021). Seaborn: statistical data visualization. J. Open Source Soft. *6*, 3021.

Webb-Robertson, B.J.M., Wiberg, H.K., Matzke, M.M., Brown, J.N., Wang, J., McDermott, J.E., Smith, R.D., Rodland, K.D., Metz, T.O., Pounds, J.G., et al. (2015). Review, evaluation, and discussion of the challenges of missing value imputation for mass spectrometry-based label-free global proteomics. J. Proteome Res. *14*, 1993–2001.

Wei, R., Wang, J., Su, M., Jia, E., Chen, S., Chen, T., and Ni, Y. (2018). Missing value imputation approach for mass spectrometry-based metabolomics data. Sci. Rep. *8*, 663.

Xu, Z.-Q., Zhang, Y., Ning, L., Liu, P.-J., Gao, L., Gao, X., and Tie, X.-J. (2017). Efficacy and safety of lapatinib and trastuzumab for HER2-positive breast cancer: a systematic review and meta-analysis of randomised controlled trials. BMJ Open *7*, e013053.

Zhang, H., Liu, T., Zhang, Z., Payne, S.H., Zhang, B., McDermott, J.E., Zhou, J.-Y., Petyuk, V.A., Chen, L., Ray, D., et al. (2016). Integrated proteogenomic characterization of human high-grade serous ovarian cancer. Cell *166*, 755–765.

Zhang, B., Wang, J., Wang, X., Zhu, J., Qi, L., Shi, Z., Chambers, M.C., Zimmerman, L.J., Shaddox, K.F., Kim, S., et al. (2014). Proteogenomic characterization of human colon and rectal cancer. Nature *513*, 382–387.

Yang, W., Soares, J., Greninger, P., Edelman, E.J., Lightfoot, H., Forbes, S., Bindal, N., Beare, D., Smith, J.A., Thompson, I.R., et al. (2013). Genomics of drug sensitivity in cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. Nucleic Acids Res. *41*, D955–D961.

Zhang, Z., Zhao, Y., Liao, X., Shi, W., Li, K., Zou, Q., and Peng, S. (2019). Deep learning in omics: a survey and guideline. Brief. Funct. Genomics. *18*, 41–57.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Chemicals, peptides, and recombinant proteins		
Ammonium bicarbonate	Sigma-Aldrich	Cat#09830-500G CAS: 1066-33-7
Acetonitrile	VWR Chemicals	Cat#20060.32 CAS: 75-05-8
Formic Acid	Fluka	Cat#94318-250ML-F CAS: 64-18-9
Benzonase	Sigma-Aldrich	Cat#E1014-25KU
Tris(2-carboxyethyl)phosphine hydrochloride (TCEP)	Sigma-Aldrich	Cat#C4706-2G CAS: 51805-45-9
Iodoacetamide	Sigma-Aldrich	Cat#C4706-2G CAS: 51805-45-9
Sodium deoxycholate	Calbiochem	Cat#264101 CAS: 302-95-4
Peptides (MS spike-in)	JPT	Custom
Deposited data		
Cancer cell lines Proteomics	This paper	PXD030304
Cancer cell lines Drug response	This paper	cancerrxgene.org
Cancer cell lines RNA-seq transcriptomics	Garcia-Alonso et al., 2018	cellmodelpassports.sanger.ac.uk
Cancer cell lines Illumina 450k methylation	Iorio et al., 2016	cellmodelpassports.sanger.ac.uk
Cancer cell lines TMT Proteomics	Nusinow et al., 2020	depmap.org
Cancer cell lines CRISPR-Cas9	Pacini et al., 2021	score.depmap.sanger.ac.uk
Cancer cell lines Mutations and Copy Number alterations	Iorio et al., 2016	cellmodelpassports.sanger.ac.uk
Cancer cell lines CTD2 drug response	Seashore-Ludlow et al., 2015	ocg.cancer.gov/programs/ctd2/data-portal
Cancer cell lines PRISM drug response	Corsello et al., 2020	depmap.org
Experimental models: Cell lines		
949 human cancer cell lines, with details, including Source and Identifier, in Table S1	CellModelPassports	See Table S1 and cellmodelpassports.sanger.ac.uk
Software and algorithms		
Analyst TF 1.7.1	Sciex USA	https://sciex.com/support/software-support/software-downloads
DIA-NN	Demichev et al., 2020	Version 1.8
DIA-NN R package	Demichev et al., 2020	https://github.com/vdemichev/diann-rpackage
DeeProM and data analysis scripts	This paper	https://doi.org/10.5281/zenodo.6563157
scikit-learn	Pedregosa et al., 2011	scikit-learn.org
PyTorch	Paszke et al., 2019	pytorch.org
matplotlib	Hunter, 2007	matplotlib.org
pandas	Reback et al., 2022	pandas.pydata.org
seaborn	Waskom, 2021	seaborn.pydata.org
Other		
Sciex 6600 TripleTOF MS	Sciex Australia	
Eksigent HPLC 425	Sciex Australia	
Trajan analytical column	Trajan Australia	Cat#2C181-03R30KSC
Trajan trap column	Trajan Australia	Cat#2C181-03T30K
Implen Nanophotometer N80	Labgear Australia	Cat#IMPLN60-TOUCH
Barocycler 2320 EXT	Pressure Biosciences Inc	Cat#2320-EXT
Cell Model Passports	van der Meer et al., 2019	cellmodelpassports.sanger.ac.uk

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to Roger Reddel, rreddel@cmri.org.au.

Materials availability

This study did not generate new unique reagents.

Data and code availability

- The raw mass spectrometry proteomic data and accompanying files have been deposited at the ProteomeXchange Consortium via the PRIDE (Perez-Riverol et al., 2019) partner repository. The accession number is PXD030304, also listed in the key resources table. All protein measurements from the ProCan-DepMapSanger dataset are available in Table S2. Drug response measurements are available under Release v8.4 at https://www.cancerrxgene.org/. Any additional datasets (e.g. protein and peptide level measurements and drug response), together with analytical results (e.g. full list of drug-protein and CRISPR-protein associations), are available in figshare https://doi.org/10.6084/m9.figshare.19345397. Any previously published datasets used for analysis in this study are listed in the Key resources table. All data are publicly available as of the date of publication.
- All original code has been deposited at Zenodo and is publicly available as of the date of publication. The DOI is https://doi.org/10.5281/zenodo.6563157, also listed in the Key resources table.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Cell lines

Cell line authentication

The 949 human cancer cell lines used in this study have been obtained from public repositories and private collections and are cultured in DMEM/F12 or RPMI 1640 (Table S1). More detailed information about each cell line can be found at the Cell Model Passports portal (https://cellmodelpassports.sanger.ac.uk/) (van der Meer et al., 2019). All cell line stocks were tested for mycoplasma contamination prior to banking using both a polymerase chain reaction (EZ-PCR Mycoplasma Detection Kit, Biological Industries) and a biochemical test (MycoAlert, Lonza). Cultures testing positive using either method were removed from the collection.

To prevent cross-contamination or misidentification, all banked cryovials of cell lines were analyzed using a panel of 94 single nucleotide polymorphisms (SNPs) (Garnett et al., 2012) (Fluidigm, 96.96 Dynamic Array IFC). The data obtained were compared against a set of reference SNP profiles that have been authenticated by short tandem repeat (STR) back to a published reference (typically the supplying repository). Where a published reference STR profile is not available, the reference SNP profile is required to be unique within the collection/dataset. A minimum of 75% of SNPs is required to match the reference profile for a sample to be positively authenticated.

Additionally, one of the replicate cell pellets generated from each cell line for this study underwent authentication via STR profiling at CellBank Australia (Westmead, Australia). To do so, STR loci were amplified using the PowerPlex® 16HS System (Promega) and the data were analyzed using GeneMapper™ ID software (ThermoFisher). Only cell lines that passed this quality control metric were retained for analysis (n = 949).

Cell culture and harvesting

For each cell line, distinct cell pellets from a single cell culture were produced as technical replicates. Cells were cultured to semi-confluence at 37°C and 5% CO_2 in the appropriate medium and then harvested. Suspension cells were centrifuged at 200 g for 5 min at 4°C and the supernatant was removed. The cells were then washed twice by resuspension in ice-cold Dulbecco's phosphate buffered saline containing no calcium or magnesium (DPBS) and centrifugation at 200 g for 5 min at 4°C. For adherent cells, the culture medium was removed before washing with ice-cold DPBS and the cells were removed by mechanical scraping into fresh ice-cold DPBS. The harvested cells were then centrifuged as before, washed twice in ice-cold DPBS, transferred to 1.5 mL centrifuge tubes (Protein LoBind Tubes, Eppendorf) and centrifuged at 600 g for 5 min at 4°C. The DPBS was removed and the tubes containing the cell pellets were snap frozen on dry ice, then stored at −80°C.

METHOD DETAILS

Cell lysis and digestion

Three cell pellets were analyzed for each of the 949 cell lines. The cell pellets were processed using Accelerated Barocycler Lysis and Extraction (ABLE) protocol with minor modifications (Lucas et al., 2019). In brief, all cell pellets were centrifuged to remove residual DPBS, then resuspended in a volume of 1% (w/v) sodium deoxycholate (SDC) that was appropriate for the cell count (between 50 and 400 μL). To this, 1 unit of benzonase was added to digest the DNA/RNA in the samples for 5 min at 37°C and mixed with shaking at

1000 rpm. After incubation, a 50 μL aliquot was taken and further processed, with peptide digestion carried out as previously published (Lucas et al., 2019).

Data independent acquisition (DIA)-MS

We used a workflow that enables high throughput and minimal instrument downtime; 2 μg of peptide was loaded for each replicate with 90-min acquisitions. Three technical replicates of peptide preparations were generated. Each replicate was injected on two of six different SCIEX™ 6600 TripleTOF® mass spectrometers coupled to Eksigent nanoLC 425 high-performance liquid chromatography (HPLC) systems, housed in a single laboratory, ProCan in Westmead, Australia (Figure S1B). In each case, an Eksigent nanoLC 425 HPLC system (Sciex) operating in microflow mode was coupled online to a 6600 TripleTOF® system (Sciex) run in sequential windowed acquisition of all theoretical fragment ion spectra (SWATH™) mode using 100 variable isolation windows (Table S1). The parameters were set as follows: lower m/z limit 350; upper m/z limit 1250; window overlap (Da) 1.0; collision energy spread was set at 5 for the smaller windows, then 8 for larger windows; and 10 for the largest windows. MS/MS spectra were collected in the range of m/z 100 to 2000 for 30 ms in high resolution mode and the resulting total cycle time was 3.2 s.

The peptide digests (2 μg) were spiked with retention time standards and injected onto a C18 trap column (SGE TRAPCOL C18 G203 300 μm × 100 mm) and desalted for 5 min at 10 μL/min with solvent A (0.1% [v/v] formic acid). The trap column was switched in-line with a reversed-phase capillary column (SGE C18 G203 250 mm × 300 μm ID 3 μm 200 Å), maintained at a temperature of 40°C. The flow rate was 5 μL/min. The gradient started at 2% solvent B (99.9% [v/v] acetonitrile, 0.1% [v/v] formic acid) and increased to 10% over 5 min. This was followed by an increase of solvent B to 25% over 60 min, then a further increase to 40% for 5 min. The column was washed with a 4 min linear gradient to 95% solvent B held for 5 min, followed by a 9 min column equilibration step with 98% solvent A. The TripleTOF® 6600 system was equipped with a DuoSpray source and 50 μm internal diameter electrode and controlled by Analyst 1.7.1 software. The following parameters were used: 5500 V ion spray voltage; 25 nitrogen curtain gas; 100°C TEM, 20 source gas 1, 20 source gas 2.

Spectral library and DIA-MS data processing

An *in silico* spectral library was created using DIA-NN (version 1.8) (Demichev et al., 2020) for the canonical human proteome (Uniprot Release, 2021_03; 20,612 sequences), along with retention time peptides and commonly occurring microbial and viral sequences. DIA-MS data in wiff file format were collected for 6,981 MS runs (Table S1), and all of these MS runs were used to create a spectral library in DIA-NN (Demichev et al., 2020). To reduce the search space, the library was confined to precursors identified in the *in silico* library only. The following settings were used for library generation. Precursor mass ranges were set between 400 and 1250 m/z and fragment mass ranges were set between 100 and 2000 m/z. Mass accuracies of 40 ppm were set for both MS1 and MS2, with the scan window set to 9. Precursors of charges 2-4 and of length between 7 and 30 were retained. Only Carbamidomethylation at Cysteine residues was allowed as a fixed modification. Interfering precursor peaks were removed, the *robust LC* (high accuracy) quantification strategy was used, and precursors were filtered at a q-value of 0.01. Protein grouping was done at the canonical protein sequence level rather than gene level. The final spectral library contained a total of 12,487 proteins and 144,578 precursors. DIA-NN (version 1.8) (Demichev et al., 2020) was used to process the MS data using this spectral library, implemented using RT-dependent normalization. See Table S1 for the full code and parameters used to run DIA-NN. All MS runs, as well as the FASTA and spectral library files, have been deposited in the Proteomics Identification Database (PRIDE) (Perez-Riverol et al., 2019) with identifier PXD030304.

DIA-NN output data were filtered to retain only precursors from proteotypic peptides with Global.Q.Value $\leq$ 0.01. These precursors were then used for protein quantification by maxLFQ (Cox et al., 2014), implemented using the DiaNN R Package (https://github.com/vdemichev/diann-rpackage) and with default parameters. Data were then $\log_2$ - transformed. 117 files were discarded from downstream analyses (Table S1), as follows: one MS run recorded no peptides, six replicates of SW900 were removed because the cell line failed STR profiling; 32 files from the earliest pilot batch were removed, as these were repeated later in the experiment; 39 files that quantified fewer than 2,000 proteins were removed; 39 files were removed because they had a poor correlation across replicates. Cell lines with a poor replicate correlation were identified using two methods. First, the minimum correlation between replicates was calculated for each cell line. The 10% of cell lines with the lowest correlation across the cohort were then examined to identify whether any MS run had a correlation with an MS run from another cell line that was above the 75% percentile of correlations (n = 11 cell lines). MS runs were then discarded for each cell line if manual examination of replicate correlations indicated that a sample mix up could have occurred (n = 21 MS runs discarded). Second, any cell line was selected that had a minimum correlation in at least one replicate of < 0.8 or a coefficient of variation, from proteins observed in > 80% of the cohort, across replicates of > 30% (n = 9 cell lines). These MS runs were then also manually examined for each cell line and MS runs that were discordant with the remainder of replicates were removed (n = 18 MS runs discarded). The final dataset, termed ProCan-DepMapSanger, was derived from 6,864 mass spectrometry runs covering 949 cell lines (Table S1) and quantifying a total of 8,498 proteins (Table S2). A filtering was applied to identify protein quantifications derived from more than one supporting peptide (n = 6,692 human proteins; Table S2. MS runs across replicates of each cell line were combined by calculating the geometric mean. Protein quantifications and number of peptides identified per protein in each MS run are available in figshare https://doi.org/10.6084/m9.figshare.19345397.

Assembly of multi-omics cell line datasets

Drug response measurements were assembled from multiple studies (Garnett et al., 2012; Iorio et al., 2016; Picco et al., 2019; Gonçalves et al., 2020) and 204 new compounds were screened and dose-response curves fitted as previously described in detail (Iorio et al., 2016; Vis et al., 2016). A total of 625 unique drugs were included in our drug response dataset. All data and respective details can be accessed at www.cancerRxgene.org (Yang et al., 2013). Cell line growth rates were represented as the ratio between the mean of the untreated negative controls measured at day one (time of drug treatment) and the mean of the dimethyl sulfoxide (DMSO) treated negative controls at day four (72 h post drug treatment). Data acquisition and processing was performed as previously described (https://www.cancerrxgene.org/) to systematically fit drug response curves and derive half-maximal inhibitory concentration (IC_{50}) measurements for each drug across the cell lines measured (Garnett et al., 2012; Yang et al., 2013; Iorio et al., 2016; Picco et al., 2019; Gonçalves et al., 2020). The dataset comprises two different screening approaches (Yang et al., 2013), and for drugs screened with both modalities, these were kept as separate entries for the downstream analyses by constructing a unique identifier (drug_id) with the pattern of <drug_code>_<drug_name>_<GDSC_version>, resulting in 819 drug_ids. A threshold of a minimum of 300 cell lines was applied to exclude drugs that were screened without enough cell lines for DeeProM analysis, resulting in a total of 710 drug_ids. The natural log of the raw IC_{50} was used for all computations.

RNA sequencing (RNA-seq) transcriptomics and Infinium HumanMethylation450 methylation measurements for the same set of cancer cell lines were assembled from previous analyses, for which the acquisition and processing are described in detail (Garcia-Alonso et al., 2018; Iorio et al., 2016). Mutation and copy-number calls were inferred from whole-exome sequencing and Affymetrix SNP6 arrays, respectively, as described previously (Iorio et al., 2016).

Genome-wide essentiality measurements for 17,486 genes were assembled for 534 cancer cell lines that overlap with those analyzed in the ProCan-DepMapSanger dataset, using CRISPR-Cas9 screens (Pacini et al., 2021). This is an integrated CRISPR-Cas9 dataset derived from two projects (Behan et al., 2019; Meyers et al., 2017) that removes library biases and represents gene essentiality as $\log_2$ fold-changes corrected for copy number bias (Iorio et al., 2018).

Dimensionality reduction and visualization

Uniform Manifold Approximation and Projection (UMAP) (McInnes et al., 2018) was calculated using Python package umap-learn (v.0.4.2) with the default setting of 15 nearest neighbors and the first 50 principal components derived from the protein matrix. Missing values were replaced with a value representing the first percentile of the input data matrix to calculate the principal components with the Python package scikit-learn (v.0.22.1) (Pedregosa et al., 2011). The first two dimensions were used for visualization.

Multi-omics factor analysis

Multi-omics decomposition by factor analysis was performed using the mofapy2 Python module (v0.5.6) (Argelaguet et al., 2018, 2020). Datasets with continuous measurements were selected for this analysis, i.e., drug responses, methylation, proteomics, and transcriptomics. For proteomic measurements, we used both the ProCan-DepMapSanger dataset and an independently acquired dataset (Nusinow et al., 2020) measuring an overlapping set of 290 cancer cell lines. Considering the strong separation of hematopoietic and lymphoid cell lines from the rest of the cell lines (Figure 2A), these were treated as a separate group in the analysis. Different numbers of factors were tested and n = 15 was chosen as it represented a trade-off between the total variance explained and the correlation between factors. Higher numbers of factors increased the correlation between factors and only marginally increased the variance explained, indicating that some factors were unnecessary. For the ProCan-DepMapSanger dataset, the mean sample intensity was regressed out prior to the factor analysis, thereby avoiding it being captured by any factor and artifactually increasing the total variance explained. Mofapy2 was run with convergence mode set to 'slow'. Scale views and groups were set to 'True' to have a unit variance.

Pairwise protein-protein correlations

We considered proteins with corresponding data also measured in the transcriptomics and CRISPR-Cas9 datasets (n = 6,347). For all pairwise protein combinations, we calculated Pearson's *r* correlations between their protein, gene expression and essentiality measurements. A minimum of 15 complete observations was required to calculate the correlation, yielding 16,580,952 pairwise combinations. Protein-protein correlations were annotated using multiple sources of protein interactions: Comprehensive Resource of Mammalian Protein Complexes (CORUM) (Ruepp et al., 2010); Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) (Szklarczyk et al., 2017); Biological General Repository for Interaction Datasets (BioGRID) (Chatr-Aryamontri et al., 2015); and Human Protein Interactome (HuRI) (Luck et al., 2020). For BioGRID, only physical interactions between proteins from humans were considered. For STRING, the most stringent threshold of the confidence score was chosen, and only interactions with a score $\geq$ 900 were considered. The average path length of the STRING PPI networks was 3.9. Protein-protein Pearson's correlations were then used to estimate the capacity to recover interactions from the different resources by ranking in ascending order all correlations according to their p value (x axis) and drawing the cumulative distribution curve of the interactions found in the resource (y axis). The area under the recall curve (AUC) was estimated using the corresponding function from the Python package scikit-learn (v0.24.2) (Pedregosa et al., 2011).

DeeProM (Deep Proteomic Marker) - Overview

We developed a multistep computational workflow, Deep Proteomic Marker (DeeProM), to identify protein biomarkers of cancer vulnerabilities. The analysis steps in DeeProM are fourfold. First, it prioritizes drug responses and CRISPR-Cas9 gene essentialities that can be confidently predicted using proteomic profiles. Second, it prioritizes strong drug responses and gene essentialities that are specific to small subsets of cancer cell lines. Third, it prioritizes protein biomarkers that show significant associations with drug responses or gene essentialities. Fourth, it prioritizes protein biomarkers that are present in specific tissues. The source code is provided in the GitHub repository (see Data and code availability).

DeeProM - DeepOmicNet model

DeeProM is powered by a deep neural network architecture, DeepOmicNet, to predict drug responses and CRISPR-Cas9 gene essentialities. DeepOmicNet ranks drugs and gene essentialities based on predicted cellular responses using the ProCan-DepMapSanger dataset as the input. Both the proteomic and the drug response datasets contain missing values, while the gene essentiality dataset provides a complete data matrix. DeepOmicNet models these missing values accordingly. Multilayer perceptron (MLP) is a classic neural network architecture that has been used by default for deep learning in numerous biomedical studies (Z. Zhang et al., 2019). To enhance the predictive performance of MLP with proteomic data, we modified its network architecture and developed DeepOmicNet with the following three major improvements:

Grouped bottleneck

DeepOmicNet uses grouped bottlenecks to avoid fully connected layers, which involves a large number of parameters being optimized. Compared with a fully-connected layer, breaking the connections into smaller groups allows the network to be more memory efficient, thus enabling wider or deeper layers. A weight matrix $W \in R^{(k \times k)}$ containing k^2 parameters with grouped bottlenecks can not only reduce the number of parameters, but also provide better predictive performance. Instead of connecting all pairs of neurons, neurons can be broken into groups, and only neurons within the same group are connected between layers (Figure S3B). The group size g can be set as any number that is divisible by the hidden layer width k. When $g=k$, all neurons are treated as one group, which reduces to a normal fully-connected layer. Multiple configurations were tested and the optimal group size g was set to 2. The number of parameters for one layer with grouped bottlenecks is significantly reduced from k^2 to $k/2 \times 2^2=2k$. The number of parameters with grouped bottlenecks is calculated as $k/g \times g^2=g \times k$, thus the run-time complexity is decreased from quadratic to linear.

Skip connections

The complete network architecture is visualized in Figure S3B. Neurons between every two consecutive layers are connected in MLP, which is computationally intensive and suboptimal for model training. To mitigate this problem, DeepOmicNet utilizes skip connections (He et al., 2016) to connect alternate layers. Let $x \in R^k$ be the vector of the i^{th} hidden layer of a real coordinate space of dimension k (corresponds to the number of neurons in a layer, also known as the layer width), the value of x_i is calculated with skip connections as:

$$x_i = f(W_{i-1}\, x_{i-1} + b_{i-1} + x_{i-2}) \qquad \text{(Equation 1)}$$

where f is the activation function, $W_{i-1} \in R^{(kxk)}$ is the weight matrix, $b_{i-1} \in R^k$ is the bias vector and $x_{i-2} \in R^k$ is the hidden layer ahead of the hidden layer x_{i-1}. The value of the hidden layer *i-2* is fed into the hidden layer *i* by skipping the hidden layer *i-1,* resulting in skip connections (Figure S3B). Each hidden layer is set with the same width k, which is a hyperparameter for model tuning, and usually is chosen to be slightly smaller than the input feature dimension. In DeepOmicNet, a sigmoid function was chosen to be the activation function f, because it outperformed the rectified linear activation function and the hyperbolic tangent function.

Loss function

DeepOmicNet is trained with mini-batches using a customized mean squared error (MSE) as the loss function. DeepOmicNet is applied to predict both drug responses and CRISPR-Cas9 gene essentialities. For the cell line m, the loss for the target variable n is defined as:

$$L_{m,n} = \begin{cases} (y_{m,n} - \hat{y}_{m,n})^2 \text{ if ground truth is present for } n \\ 0 \text{ otherwise} \end{cases} \qquad \text{(Equation 2)}$$

where $y_{m,n}$ is the ground truth label of the target n (either drug response or gene essentiality) and $\hat{y}_{m,n}$ is the predicted value of the target n. The label $y_{m,n}$ is missing if a particular cell line m was not screened with drug n.

In addition to the three major improvements, other characteristics of DeepOmicNet include the following:

Missing values

One specific challenge of DIA-MS based proteomics is the missing values in the data matrix (Webb-Robertson et al., 2015; Poulos et al., 2020). Imputation is widely used but often leads to distortion to some extent (Wei et al., 2018). For DeepOmicNet, missing values were replaced by zeros, thus allowing the neural network to ignore the weight update for these inputs.

Hyperparameter tuning

Hyperparameters including model width, depth, learning rate and batch size were tuned to achieve the highest predictive performance. Pearson's r between true and predicted values is used as the evaluation metric. Hyperparameters that resulted in the highest performance in five-fold cross-validation of the 80% training data were chosen for the final evaluation on the 20% independent test set. The chosen hyperparameters can be found in the configuration files in the source code (see Data and code availability).

Thresholding

The thresholds are set to Pearson's $r > 0.4$ and Pearson's $r > 0.3$ to prioritize highly predictive drug responses and CRISPR-Cas9 gene essentialities, respectively. This yielded 67 drug responses and 62 gene essentialities.

DeeProM - Prioritizing selective associations

DeeProM prioritizes drug responses and CRISPR-Cas9 gene essentialities that are likely to be non-toxic to normal cells. Since the cell lines used in this study were derived from cancer (the majority) or viral transformation, we approximated this task by finding drug responses and gene essentialities that were selective for only a small fraction of the cell lines, which indicates that the drug response or gene essentiality is less likely to be toxic to normal cells. Ranking of strongly selective drug responses and gene essentialities was performed using Python package scipy (v1.5.2) skew function, which calculates the Fisher-Pearson's coefficient of skewness. Skewness values of -1 and -2 were used for drug responses and gene essentialities, respectively.

DeeProM - Linear regression models

Associations between protein and phenotypic measurements, drug responses and gene essentialities were performed using linear regression models (sklearn v0.24.2 class LinearRegression). Several technical and biological covariates were added to the model to remove potentially spurious associations. First, we built the following technical covariates into the model: (i) the growth rate of the cell lines; (ii) cell culture medium, D/F12 (DMEM/F12: 10% FBS, 1% PenStrep) or R (RPMI 1640: 10% FBS, 1% PenStrep, 4.5 mg/mL Glucose, 1 mM Sodium Pyruvate); (iii) cell line growth properties (i.e., adherent, semi-adherent or suspension), ploidy, and if they are hematopoietic and lymphoid cell lines; (iv) sample mean protein replicates Pearson's correlation; (v) for the CRISPR-Cas9 gene essentiality only, we considered the institute of origin of the CRISPR-Cas9 screen, i.e., Wellcome Sanger or Broad Institute; and (vi) for the drug response models only, we considered the cell line mean IC_{50} across all drugs. Discrete covariates were represented as dummy binary variables. Second, to identify associations that were exclusively found at the protein level, we added the following gene expression covariates to the model: (i) the first ten gene expression principal components using the Python package scikit-learn (v0.24.2) (Pedregosa et al., 2011); and (ii) the corresponding transcript level of the protein being tested. Formally, we fit the following linear regression model for each drug response/gene essentiality–protein pair:

$$d = M\beta_0 + E\,\beta_1 + e\,\beta_2 + p\,\beta_3 + \varepsilon \qquad \text{(Equation 3)}$$

where, d represents an n x 1 vector of the drug response IC_{50} (n = 710 drugs) for 947 cell lines or CRISPR-Cas9 gene essentiality $\log_2$ fold changes (n = 17,486 genes) for 534 cell lines; M is the n x k matrix of covariates (k = 11 covariates); E is an n x m matrix containing the first (m = 10) principal components of the gene expression dataset; e is a vector of size n x 1 containing the transcriptomics measurements of the corresponding protein p; p is a vector of size n x 1 containing the protein measurements; and ε is the error vector of size n x 1. For each protein, cell lines with missing values were dropped from the modeling. For drug response, missing values were replaced by the drug mean IC_{50}. The number of cell lines with complete information per fit was provided. The model was fitted by minimizing the residual sum of squares to estimate the parameters β_n of each variable. In total, there were 710 × 6,692 = 4,751,320 drug-protein pairs and 17,486 × 6,692 = 117,016,312 possible CRISPR-Cas9 gene essentiality-protein pairs, however, both drug response and CRISPR-Cas9 data covered various subsets of cell lines. We required a minimum number of 60 cell lines to test the association. As a result, a total of 4,218,788 drug-protein and 86,584,537 CRISPR-Cas9-protein tests were performed.

Statistical assessment of the improvement of adding protein measurements to the linear regression was performed using a likelihood ratio test between the full model [Equation 3] and the null model, which excludes the protein measurement and its parameter β_2. Likelihood ratio test's p value was estimated using a chi-square distribution with one degree of freedom. Adjustment for multiple testing was performed per drug or CRISPR-Cas9 gene essentiality using the Benjamini-Hochberg procedure to control the false discovery rate (FDR). Associations with FDR < 10% for models with covariates, or FDR < 0.1% for models that do not use covariates, were considered as significant associations.

DeeProM - Tissue type level filtering

To investigate drug-protein and CRISPR-Cas9-protein associations for a given tissue type, we used metrics derived from DeepOmicNet, Fisher Pearson's coefficient of skewness and linear regression to prioritize drug responses and CRISPR-Cas9 gene essentialities according to the thresholds described above. The overlap of these three methods was used as the final result, which included 7,698 drug-protein and 5,823 CRISPR-Cas9-protein associations. Finally, we applied additional filters to further prioritize associations that are unique to certain tissue types, yielding the final list of 108 drug-protein associations for 18 drugs and 1,538 CRISPR-Cas9-protein associations for 38 genes (Figure 4C). The filtering steps are described below:

Step 1

Tissue types with < 20 protein measurements were filtered out.

Step 2

For each significant association identified from 7,698 drug-protein and 5,823 CRISPR-Cas9-protein associations, Pearson's r was calculated. Using protein data, the significance level was set to 0.1 and Pearson's r was defined as $r_{(target,\ protein)}^{tissue}$ for a target-protein association in a given tissue type. Here, protein indicates a protein of interest, target indicates either a drug response or CRISPR-Cas9 gene essentiality and tissue refers to the tissue type under investigation. We then calculated the Pearson's r for the gene that

encodes each protein, and we defined this value as $r_{(target,\ RNA)}^{tissue}$. To prioritize associations that were uniquely identified at the protein level, the difference d for a given tissue type between target-protein and target-RNA associations was set to be larger than 0.15, where $d = |r_{(target,\ protein)}^{tissue}| - |r_{(target,\ RNA)}^{tissue}|$. This prioritizes associations that have either strong positive or strong negative correlations at the protein level but have weak correlations around zero at the RNA level. Rare cases where $r_{(target,\ protein)}^{tissue}$ and $r_{(target,\ RNA)}^{tissue}$ have opposite signs and d is close to 0 were not considered.

Step 3

For drug-protein associations, a large value of d alone is insufficient to select candidate associations, because a drug may be entirely ineffective for all the cell lines in a particular tissue type. Therefore, we applied an additional filter to ensure that a drug is effective on the cell lines for which the protein abundance is high. That is, for a drug-protein association to be included for a given tissue type, the median IC_{50} of the 20% of cell lines with the highest corresponding protein abundance must be lower than the maximal concentration for that drug.

Step 4

The remaining associations were ranked in descending order according to d.

Comparing DeepOmicNet and other models

DeepOmicNet was compared against traditional machine learning models, including elastic net and Random Forest. A total of 947 cell lines were randomly separated into a training set comprising 80% of the cell lines, and a test set with the remaining cell lines for unbiased evaluation. Grid search was used to find the best hyperparameters for elastic net and Random Forest in the training set by cross-validation. Hyperparameter tuning for DeepOmicNet was performed manually due to the limit of graphics processing unit (GPU) memory. For each model, missing values were imputed using the method that gave the best prediction based on cross-validation. Specifically, four imputation methods were considered, including imputation by minimum, first percentile of the whole input matrix, mean and zero. Based on the predictive performance of models in cross-validation, imputation by one percentile of the whole matrix was chosen for the elastic net and Random Forest, and imputation by zero was used for DeepOmicNet. This strategy yielded the best prediction accuracy in comparison with other imputation and normalization methods, such as imputation with k-nearest-neighbor, mean values of proteins and zeros. A cut-off was set at a minimum of 300 screened cell lines for the drug response dataset to filter out drugs without sufficient data. A simplified version of DeepOmicNet without grouped bottlenecks was used for omics data other than proteomic data due to the large input dimension. The Python package scikit-learn (v.0.22.1) (Pedregosa et al., 2011) was used to train elastic net and Random Forest models. DeepOmicNet was implemented and trained using PyTorch (v.1.4.0) (Paszke et al., 2019).

Machine learning for lethality prediction

Due to the limit of GPU memory, elastic net, Random Forest and DeepOmicNet were applied only to transcriptomic and proteomic data to predict CRISPR-Cas9 gene essentialities. The same computational strategy for drug response prediction was used to predict CRISPR-Cas9 gene essentialities.

Predictive power comparison with CCLE

The predictive power of machine learning models for two proteomic datasets (ProCan-DepMapSanger and the Cancer Cell Line Encyclopedia (CCLE) (Nusinow et al., 2020)) were compared independently on three drug response datasets (Sanger, CTD2 and PRISM) and the CRISPR-Cas9 gene essentiality dataset (Behan et al., 2019; Meyers et al., 2017; Pacini et al., 2021). The analysis was performed using the 290 overlapping cell lines to ensure a fair comparison. The proteomic (Nusinow et al., 2020) and drug response (CTD2 and PRISM) (Corsello et al., 2020; Seashore-Ludlow et al., 2015; Rees et al., 2016) datasets were retrieved from the DepMap portal (https://depmap.org/portal/). DeepOmicNet was used to compare the predictive power of models for CRISPR-Cas9 gene essentialities, and Random Forest was used for drug response prediction due to the limited number of cell lines for certain drugs. AUC instead of IC_{50} was used for the drug response (PRISM) dataset due to a large proportion of drugs having no IC_{50} values provided.

Downsampling for drug response prediction

For downsampling analysis, the full set of 8,498 proteins were randomly downsampled using a step decrease of 500 proteins (Figure 7A). Each step was repeated ten times and for each iteration, results from five-fold cross-validations and an unbiased test were included in evaluating predictive power. Therefore, each downsampling step used ten different random subsets of proteins for six distinct experiments (the five-fold cross-validation and one unbiased test). The predictive power of each DeepOmicNet model was evaluated for each protein set and each iteration, with confidence intervals summarizing the results across the ten iterations. This random downsampling procedure was also performed with step sizes of 250 proteins for proteins in Categories A, B and C (Figure 7E).

Figures

Figures were generated using Pandas (Reback et al., 2022), Seaborn (Waskom 2021) and Matplotlib (Hunter 2007) Python packages, as detailed in the Key resources table.

QUANTIFICATION AND STATISTICAL ANALYSIS

DIA-NN (version 1.8) (Demichev et al., 2020) was used to build the peptide spectral library and process raw MS data. MaxLFQ (Cox et al., 2014) was used to quantify relative protein intensities. Sigmoid drug response curves were fitted to estimate IC50s (Vis et al., 2016). Associations between pairs of continuous variables were tested by Peason's correlation coefficient *r*. Statistical tests were adjusted for multiple hypotheses correction using the Benjamini-Hochberg False Discovery Rate (FDR), and statistical significance was considered when FDR < 5%, except when otherwise specified (such as when multiple thresholds were compared). Quantification methods and statistical analyses for the proteomics, drug response and multi-omics datasets are described in the respective sections of the STAR Methods. Unless otherwise stated, relevant statistical parameters are reported in the legend of each figure.