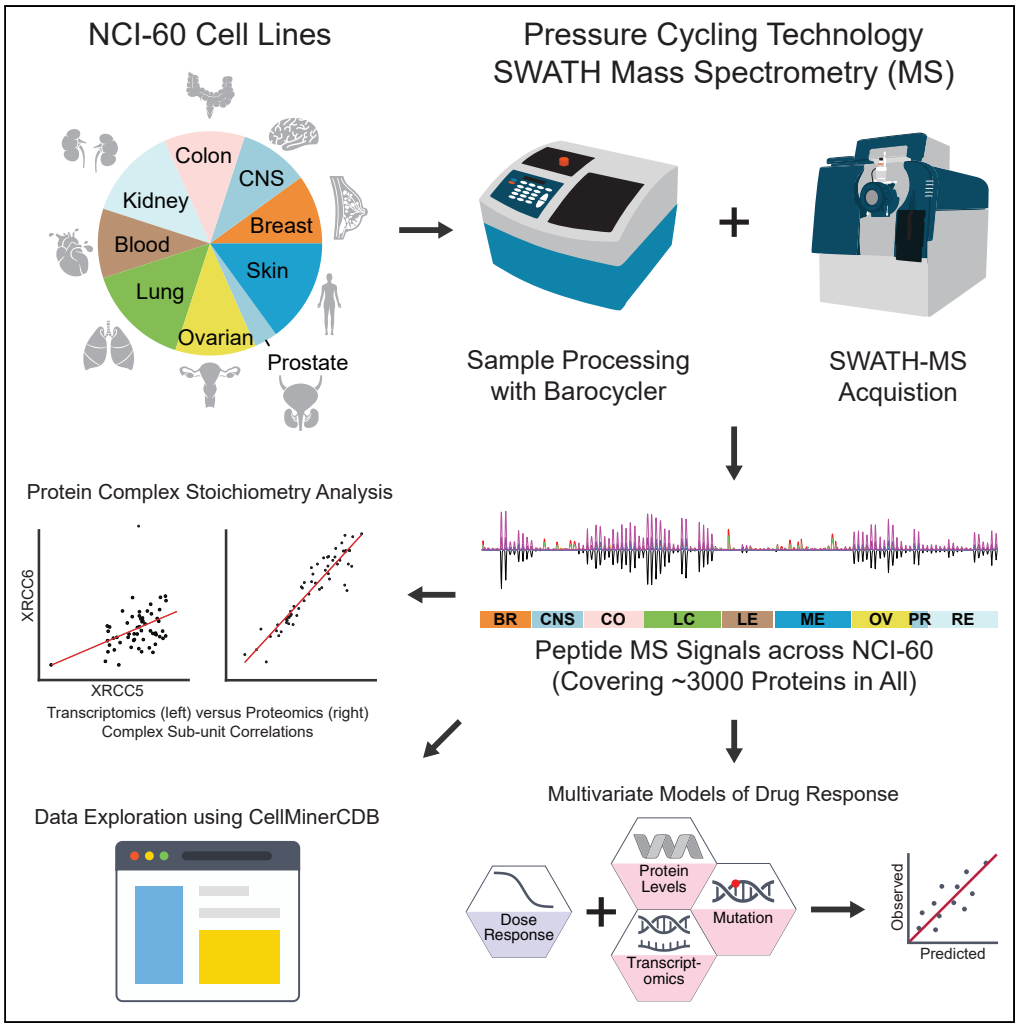
iScience

CellPress

Article

Quantitative Proteome Landscape of the NCI-60 Cancer Cell Lines



Tiannan Guo, Augustin Luna, Vinodh N. Rajapakse, ..., Chris Sander, Yves Pommier, Ruedi Aebersold

guotiannan@westlake.edu.cn (T.G.)
yves.pommier@nih.gov (Y.P.)
aebersold@imsb.biol.ethz.ch (R.A.)

HIGHLIGHTS

High-quality NCI-60 proteotypes created using pressure cycling technology and SWATH-MS

Proteotypes improve drug response prediction in multi-omics regression analysis

~3000 measured proteins allow investigation into protein complex stoichiometry

CellMinerCDB (discover.nci.nih.gov/cellminercdb) portal allows dataset exploration

DATA AND CODE AVAILABILITY

PXD003539

Guo et al., iScience 21, 664–680
November 22, 2019
https://doi.org/10.1016/j.isci.2019.10.059

iScience

CellPress

Article

Quantitative Proteome Landscape of the NCI-60 Cancer Cell Lines

Tiannan Guo,[1,2,3,29,30,*] Augustin Luna,[4,5,29] Vinodh N. Rajapakse,[6,29] Ching Chiek Koh,[3,27,29] Zhicheng Wu,[1,2] Wei Liu,[1,2,25] Yaoting Sun,[1,2] Huanhuan Gao,[1,2] Michael P. Menden,[7,24] Chao Xu,[8] Laurence Calzone,[9] Loredana Martignetti,[9] Chiara Auwerx,[3] Marija Buljan,[3] Amir Banaei-Esfahani,[3,10] Alessandro Ori,[11] Murat Iskar,[12,28] Ludovic Gillet,[3] Ran Bi,[25] Jiangnan Zhang,[25] Huanhuan Zhang,[26] Chenhuan Yu,[26] Qing Zhong,[13,23] Sudhir Varma,[14] Uwe Schmitt,[15] Peng Qiu,[16] Qiushi Zhang,[1,2] Yi Zhu,[1,2,3] Peter J. Wild,[13] Mathew J. Garnett,[17] Peer Bork,[12,18,19,20] Martin Beck,[12,21] Kexin Liu,[25] Julio Saez-Rodriguez,[7] Fathi Elloumi,[6] William C. Reinhold,[6] Chris Sander,[4,5] Yves Pommier,[6,*] and Ruedi Aebersold[3,22,*]

SUMMARY

Here we describe a proteomic data resource for the NCI-60 cell lines generated by pressure cycling technology and SWATH mass spectrometry. We developed the DIA-expert software to curate and visualize the SWATH data, leading to reproducible detection of over 3,100 SwissProt proteotypic proteins and systematic quantification of pathway activities. Stoichiometric relationships of interacting proteins for DNA replication, repair, the chromatin remodeling NuRD complex, β-catenin, RNA metabolism, and prefoldins are more evident than that at the mRNA level. The data are available in CellMiner (discover.nci.nih.gov/cellminercdb and discover.nci.nih.gov/cellminer), allowing casual users to test hypotheses and perform integrative, cross-database analyses of multi-omic drug response correlations for over 20,000 drugs. We demonstrate the value of proteome data in predicting drug response for over 240 clinically relevant chemotherapeutic and targeted therapies. In summary, we present a novel proteome resource for the NCI-60, together with relevant software tools, and demonstrate the benefit of proteome analyses.

INTRODUCTION

To date, forays into the molecular landscape of diseases, in particular cancers, have primarily focused on genomics and transcriptomics (Barretina et al., 2012; Cancer Genome Atlas Research et al., 2013; Garnett et al., 2012) due to the maturity and availability of high-throughput DNA- and RNA-based techniques. Protein-level measurements, although important for providing the granularity and detail necessary for personalized therapeutic decisions, are underutilized due to technical hurdles. Advances in data-dependent acquisition (DDA) mass spectrometry (MS) have permitted quantitative proteomic profiling of hundreds of tumor samples using multi-dimensional fractionated MS analyses of each sample (Mertins et al., 2016; Zhang et al., 2014, 2016), demonstrating the added value of protein measurement in classifying tumors. Nevertheless, such DDA workflows suffer from relatively lower sample-throughput, higher sample consumption, and increased technical complexity relative to genomic analyses. These factors have precluded their routine use in clinically relevant applications (e.g. tumor classification and drug response prediction) at the speed and scale achieved by genomic and transcriptomic approaches (Barretina et al., 2012; Garnett et al., 2012; Rajapakse et al., 2018; Reinhold et al., 2019).

The NCI-60 human cancer cell line panel contains 60 lines from nine different tissue types. The NCI-60 have been molecularly and pharmacologically characterized with unparalleled depth and coverage, offering a prime *in vitro* model to further our understanding of cancer biology and cellular responses to anti-cancer agents (Monks et al., 2018; Reinhold et al., 2012, 2019; Shoemaker, 2006). Discoveries enabled by the NCI-60 in recent years include the development of the FDA-approved drugs, such as oxaliplatin for the treatment of colon cancers (Fojo et al., 2005), eribulin for metastatic breast cancers (Shoemaker, 2006), bortezomib for the treatment of multiple myeloma (Holbeck et al., 2010), and romidepsin for cutaneous T cell lymphomas (Bates et al., 2015), and development of the indenoisoquinoline class of non-camptothecin topoisomerase I inhibitors (Burton et al., 2018). The sensitivity of the NCI-60 to over 100,000 synthetic or natural compounds derived from a wide range of academic and industrial sources has been measured,

[1]Key Laboratory of Structural Biology of Zhejiang Province, School of Life Sciences, Westlake University, Hangzhou, Zhejiang, P. R. China

[2]Guomics Laboratory of Proteomic Big Data, Institute of Basic Medical Sciences, Westlake Institute for Advanced Study, 18 Shilongshan Road, Hangzhou 310024, Zhejiang Province, China

[3]Department of Biology, Institute of Molecular Systems Biology, ETH Zurich, Zurich, Switzerland

[4]cBio Center, Division of Biostatistics, Department of Data Sciences, Dana-Farber Cancer Institute, Boston, MA 02115, USA

[5]Department of Cell Biology, Harvard Medical School, Boston, MA 02115, USA

[6]Developmental Therapeutics Branch, Center for Cancer Research, National Cancer Institute, National Institutes of Health, Bethesda, MD 20892, USA

[7]RWTH Aachen University, Faculty of Medicine, Joint Research Centre for Computational Biomedicine (JRC-COMBINE), Aachen, Germany

[8]Faculty of Software, Fujian Normal University, Fuzhou, China

[9]Institut Curie, PSL Research University, INSERM, U900, Mines Paris Tech 75005, Paris, France

[10]PhD Program in Systems Biology, Life Science Zurich

Continued

constructing the most comprehensive open resource for cancer pharmacology. The NCI-60 remains actively used by many academic laboratories and drug companies to assess overall toxicity and drug response selectivity. In addition, many of the NCI-60 cell lines are widely used for cell biology and pharmacology (MCF-7, MDA-MB231, HCT116, HCT15, HT29, HL60, CCR-CEM, K562, etc.), and 55 and 44 of the NCI-60 cell lines overlap within larger cancer cell line databases GDSC and CCLE, respectively (Rajapakse et al., 2018), providing a unique and highly valuable resource for cross-comparisons.

The proteome of the NCI-60 cells has been analyzed previously by data-dependent analysis with a commonly used discovery MS technique (Gholami et al., 2013). This proteome dataset was obtained using a sophisticated two-dimensional peptide fractionation strategy. However, peptides and proteins were quantified without technical replicates (Gholami et al., 2013), making it difficult to evaluate quantitative accuracy. To achieve reproducible and high-throughput proteomic profiling while developing new technologies, we have developed a workflow (Guo et al., 2015; Shao et al., 2015) integrating pressure cycling technology (PCT) with SWATH-MS. PCT is an emerging sample preparation method that accelerates and standardizes sample preparation for proteomic profiling (Powell et al., 2012). SWATH-MS is an MS-based proteomic technique that consists of data-independent acquisition (DIA) and a targeted data analysis strategy with unique advantages over other MS-based proteomic methods (Gillet et al., 2012; Rost et al., 2014). With this technique, all MS-measurable peptides of a sample are fragmented and recorded recursively, and the resulting digital proteome maps can be used to reproducibly detect and quantify proteins across large numbers of samples without the need for isotope labeling. The integrated PCT-SWATH workflow thus significantly increases the sample throughput and data reproducibility, providing quantitative accuracy, while also reducing sample consumption to ca. 1 microgram of total peptide mass per sample (Guo et al., 2015; Shao et al., 2015).

Here, we describe the acquisition of proteome maps of the NCI-60 in duplicate by PCT-SWATH and make them available via the CellMiner portals (discover.nci.nih.gov/cellminercdb/ and discover.nci.nih.gov/cellminer), enabling interactive exploration and data download (Rajapakse et al., 2018). The techniques described in this report allowed the efficient acquisition of 120 proteome maps (within about 30 working days from sample preparation to SWATH data acquisition on a single instrument) with minimal sample requirement (ca. 1 microgram of total peptide mass). We focused on 3,171 SwissProt proteotypic proteins that were identified across all cell lines, generating a data matrix (120 proteomes vs. 3,171 proteins). Raw signals of each peptide and protein in each sample were curated and visualized with an expert system. The proteomic data expand the existing NCI-60 molecular landscapes (Holbeck et al., 2010; Monks et al., 2018; Rajapakse et al., 2018; Reinhold et al., 2012, 2019) and their integration with the larger databases from the Broad-MIT (CCLE, CTRP) and MGH-Sanger (GDSC) (Rajapakse et al., 2018), allowing systematic investigation of the complementarity among genomics, transcriptomics, and proteomics.

RESULTS

Acquisition of the NCI-60 Proteome

We applied the PCT-SWATH workflow (Guo et al., 2015) to generate quantitative proteome maps of the NCI-60 cell lines in technical replicates, resulting in 120 SWATH maps with high reproducibility at the raw data level (Figure S1). Approximately 1 microgram peptide mass per sample was sufficient for analyses. The PCT-assisted sample preparation took about 18 working days and the SWATH-MS data acquisition about 12 working days. Thus, the entire process, from sample preparation to data acquisition, could be accomplished within 30 working days. This results from the elimination of multidimensional fractionation and the consequent processing of each sample using one barocycler per mass spectrometer, from which a single file per sample was acquired (Figure S1, Table S1). We have matched our cell line IDs with a previous publication from the Kuster group (Gholami et al., 2013) and corrected a few known errors in the cell line identifiers (Table S1). These cell lines were shuffled randomly to avoid bias from tissue types and minimize batch effects from PCT-assisted sample preparation. The two sets of technical replicates were acquired using SWATH-MS in different time periods to allow the evaluation of batch effects from the MS analysis. This approach constitutes an advance in sample-throughput compared with other cancer proteomic workflows of similar scale (Gholami et al., 2013; Mertins et al., 2016; Zhang et al., 2014, 2016).

SWATH proteome maps contain fragment ion chromatograms from all MS-measurable peptides, albeit in a highly convoluted form. To interpret the SWATH maps, we built a human cancer cell line spectral library containing 86,209 proteotypic peptides, i.e. peptides that uniquely identify a specific protein from 8,056

Graduate School, University of Zurich and ETH Zurich, Zurich, Switzerland

[11]Leibniz Institute on Aging, Fritz Lipmann Institute (FLI), Beutenbergstrasse 11, 07745 Jena, Germany

[12]Structural and Computational Biology Unit, European Molecular Biology Laboratory, 69117 Heidelberg, Germany

[13]Institute of Surgical Pathology, University Hospital Zurich, Zurich, Switzerland

[14]HiThru Analytics, Laurel, MD 20707, USA

[15]Scientific IT Services, ETH Zurich, Zurich, Switzerland

[16]Department of Biomedical Engineering, Georgia Institute of Technology and Emory University, 313 Ferst Dr., Atlanta, GA 30332, USA

[17]Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK

[18]Molecular Medicine Partnership Unit, University of Heidelberg and European Molecular Biology Laboratory, 69120 Heidelberg, Germany

[19]Max Delbrück Centre for Molecular Medicine, 13125 Berlin, Germany

[20]Department of Bioinformatics, Biocenter, University of Würzburg, 97074 Würzburg, Germany

[21]Cell Biology and Biophysics Unit, European Molecular Biology Laboratory, 69117 Heidelberg, Germany

[22]Faculty of Science, University of Zurich, Zurich, Switzerland

[23]Cancer Data Science Group, Children's Medical Research Institute, University of Sydney, Sydney, NSW, Australia

[24]Bioscience, Oncology, IMED Biotech Unit, AstraZeneca, Cambridge, UK

[25]Department of Clinical Pharmacology, College of Pharmacy, Dalian Medical University, Dalian, Liaoning, China

[26]Key Laboratory of Experimental Animal and Safety Evaluation, Zhejiang Academy of Medical Sciences, Hangzhou, Zhejiang, China

[27]Present address: Sanger Institute, Wellcome Trust

SwissProt proteins (Table S1). Using this library and the OpenSWATH software (Rost et al., 2014), we identified 6,556 protein groups, covering 81% of the library (Figure S2). To avoid ambiguity of peptide/protein quantification, we limited our analyses to canonical and proteotypic peptides and proteins by excluding protein isoforms, un-reviewed protein sequences, peptide/protein sequence variants, and protein groups that could not be deconvoluted.

Development of DIA-expert for SWATH/DIA Data Curation

We evaluated the technical variation of each measurement through manual inspection of the OpenSWATH results based on the replicated measurement for each cell line. Observed missing values and technical variation were attributed to cell-type-specific interfering signals leading to invalid SWATH assays and the presence of irregular liquid chromatography (LC) and MS behavior of certain peptides. These phenomena have been observed previously in selected reaction monitoring (SRM)-based targeted proteomics studies (Picotti and Aebersold, 2012). To obtain high accuracy quantitative data, we developed an expert system, i.e. DIA-expert to refine the peptide identification and quantification (Figure S3).

The DIA-expert reads SWATH search results containing a q-value for each peptide identified in a sample and then selects the sample in which a peptide precursor is identified with the highest confidence among all samples (Figure S3). The selected sample then becomes the reference against which identification of the particular peptide in the other samples is evaluated. This step is iterated for each peptide precursor analyzed. Then, DIA-expert selects from the SWATH assay library the peptides identified for the specific sample set and proceeds to build a new library containing all the transitions for each peptide precursor. Extracted chromatograms for each precursor and its fragments are obtained. This initial transition set is used for subsequent transition refinement. We next applied empirical expert rules (Keller et al., 2002; Shao et al., 2015), including peak detection expert, reference sample expert, and peak group pairing expert. The software outputs a data matrix of quantities of each peptide in all samples and graphically presents the peak groups of curated peptide fragments used for generating the reported results. In contrast to typical SRM or SWATH/DIA analysis strategies, which apply the same few selected peptide fragments as indicators of peptide abundance in all cohort samples, the DIA-expert examines the sample-specific suitability of all peptide fragments and builds peptide abundance values based on *ad hoc* curated peptide fragments.

DIA-Expert-Curated Results of the NCI-60 Proteome

Excluding proteins/peptides that were not technically reproducible resulted in 22,554 proteotypic peptides from 3,171 proteins, with 8% missing values at the peptide level and 0.1% missing values at the protein level across all MS runs (Table S1). On average, seven peptide precursors and six unique peptide sequences were identified per protein. Several proteins were identified with more than 200 peptides (Figure 1B). The proteins excluded by DIA-expert may not be incorrect identifications but rather irreproducible quantifications due to either technical (for instance the signal-to-noise ratio) or biological issues (such as post-translational modifications or splicing variants). Improved computational methods may recover more information from this dataset.

Most peptides for the 3,171 proteins were quantified in all cell lines at both MS1 and MS2 levels. Although the replicates show consistent quantification, different cell lines expressed variable levels of proteins. Two representative peptides are shown in Figure 1A. The coefficients of determination (R^2) between technical replicates for the overall expression of peptides (Figure 1C) and proteins (Figure 1D) were 0.974 and 0.978, respectively, with a dynamic range over five orders of magnitude (Figure 1E). The DIA-expert provides the raw MS signals for each quantitative value, allowing visual inspection of the MS signal for every peptide in each sample. Increasing the minimal number of peptides identified per protein to 2, 3, or 4 resulted in fewer proteins quantified (2,200; 1,741; and 1,428 proteins, respectively). However, this did not substantially improve quantitative accuracy (Figure S4).

Characterization of the NCI-60 Quantitative Proteomes

The landscape of the 120 proteotypes is displayed in Figure 2A. Technical replicates of the quantified proteotypes were clustered using an unsupervised approach, confirming high quantitative accuracy. In most cases, the proteotypes are not strikingly different across different cancer cell lines, in sharp contrast with the distinct proteomes of tumor versus non-tumor kidney tissues (Guo et al., 2015). The median protein intensity coefficient of variation (CV) of the different cell lines was 48%. The CV demonstrated a low

Genome Campus, Hinxton, Cambridge CB10 1SA, UK

[28]Present address: Division of Molecular Genetics, German Cancer Research Center (DKFZ), Im Neuenheimer Feld 280, 69120 Heidelberg, Germany

[29]These authors contributed equally

[30]Lead Contact

*Correspondence: guotiannan@westlake.edu.cn (T.G.), yves.pommier@nih.gov (Y.P.), aebersold@imsb.biol.ethz.ch (R.A.)

https://doi.org/10.1016/j.isci.2019.10.059

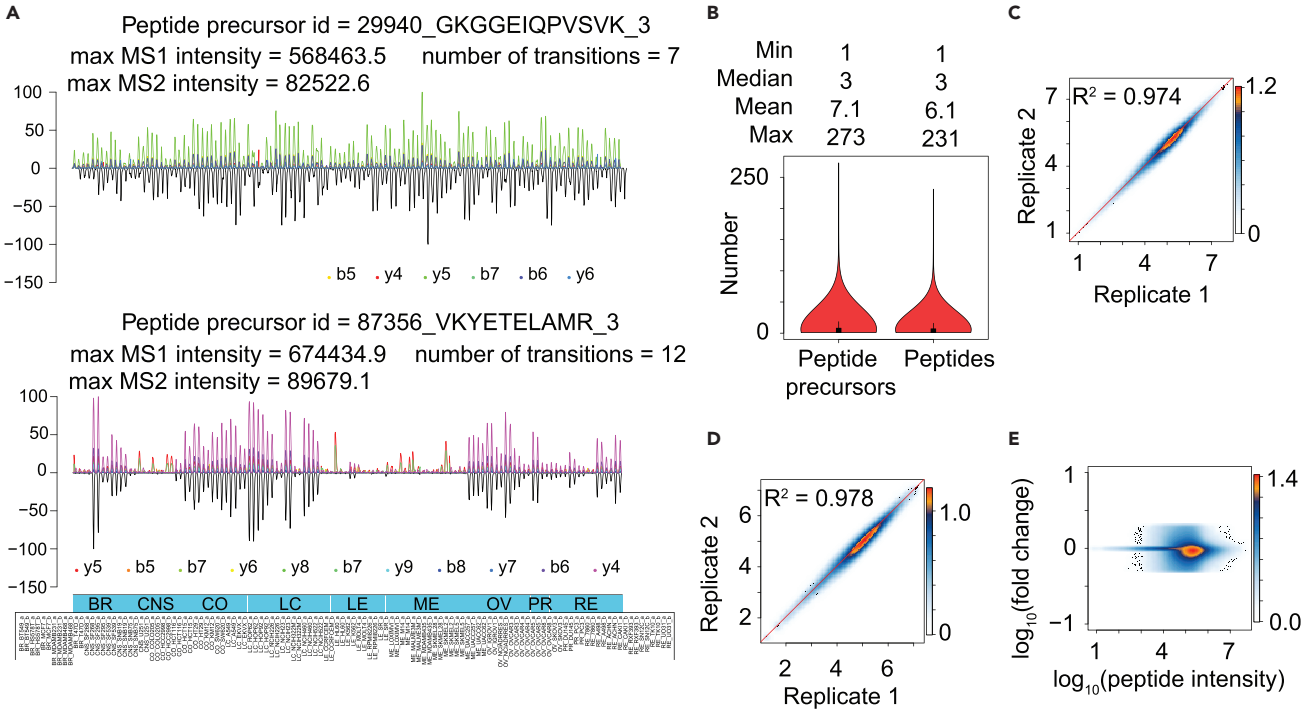


Figure 1. Acquisition of NCI-60 Proteotype

(A) Representative peptide signals as curated and visualized by the DIA-expert software.

(B–D) (B) Distribution of peptide precursors and peptides per protein. Overall coefficient of determination between technical replicates at the peptide level (C) and the protein level (D). Heatmap of the log10 transformed intensity of each peptide/protein in each cell line technical replicate.

(E) Dynamic range of the MS signals for 22,968 proteotypic peptides.

dependence on protein abundance, as evident from the distribution of its values for different expression level quantile groups of the measured proteins (Figure 2B).

We then compared the data with the previously reported DDA-MS proteomic data for the NCI-60 (Gholami et al., 2013). Whereas the DDA data reported a comparable number of IPI protein groups per cell line as the SwissProt proteotypic protein number from this SWATH dataset (Table S2), the SWATH data exhibited a much higher degree of consistency (Figure S5) and better quantitative accuracy (Figures S6–S32).

Accessibility of the NCI-60 Proteotypes

To enable easy data access, visualization, and comparison with other NCI-60 datasets, we have incorporated the SWATH data into the CellMiner databases and web application (Rajapakse et al., 2018; Reinhold et al., 2012; Shankavaram et al., 2009). This allows direct downloads of the data, as well as direct comparative and integrative analyses with other molecular and pharmacological data, (e.g. sensitivity of each cell line to over 20,000 compounds) and the inspection of specific genes, up to 150 per query. The detailed instructions for using this resource are provided in Figure S33 and at the project websites (discover.nci.nih.gov/cellminer and discover.nci.nih.gov/cellminercdb). Figure 2E shows snapshots of data queries for KU70 versus KU80 protein and transcript expression levels (XRCC6 and XRCC5, respectively). Raw and processed data matrices of the NCI-60 proteotype have also been deposited in public databases, including PRIDE (Jones et al., 2006) and ExpressionArray (Brazma et al., 2003).

Insights from a Quantitative Comparison of Protein versus Transcript Expression

Because of the extensive prior characterization of the NCI-60 transcriptome (Monks et al., 2018; Rajapakse et al., 2018; Reinhold et al., 2012), we were able to correlate protein and gene expression for each of the 3,171 proteins quantified across the NCI-60. Table S3 shows that some proteins exhibit a high correlation with their transcript, indicating the transcripts are the main drivers of protein expression. Correlations and

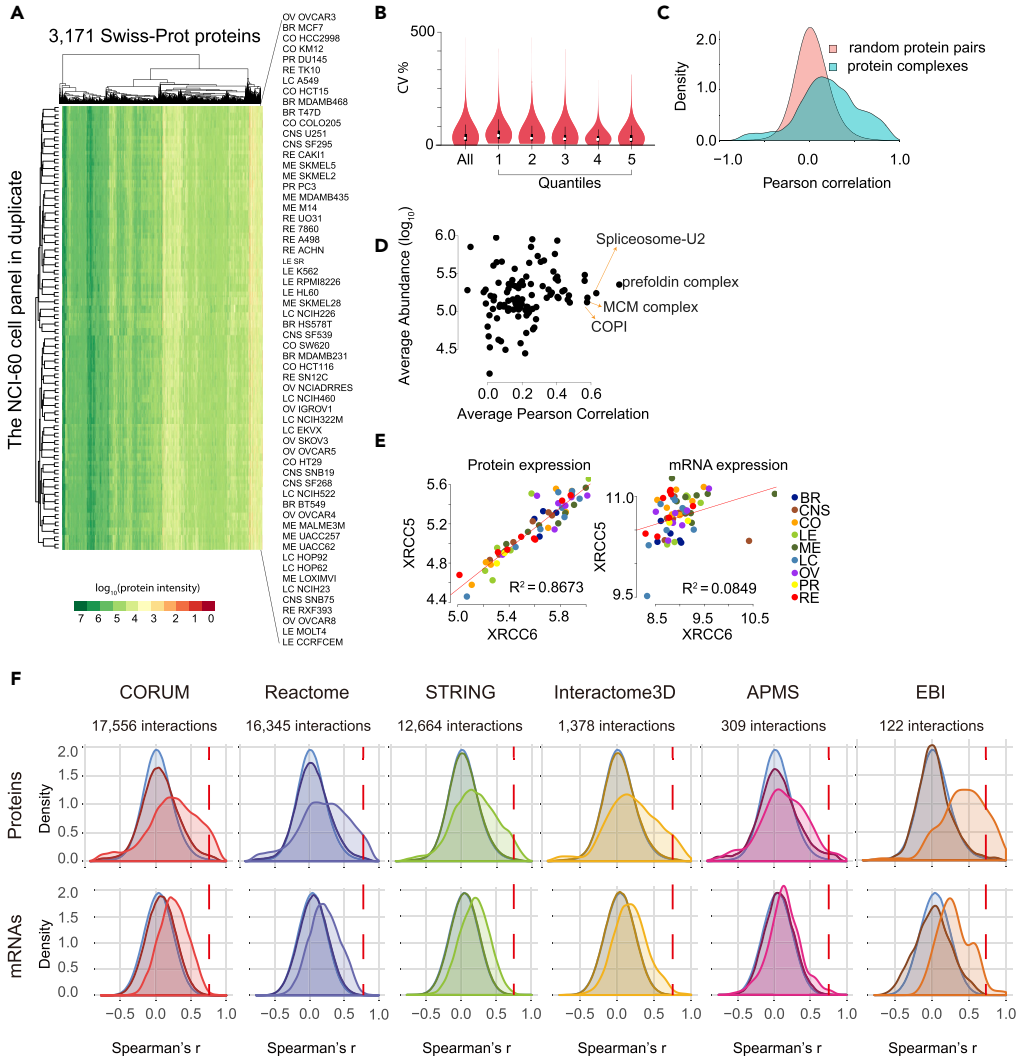


Figure 2. Characterizing the NCI-60 Quantitative Proteomes

(A) Heatmap overview of NCI-60 proteotype data matrix. Quantification of 3,171 Swiss-Prot proteins in 120 SWATH runs.

(B) Variation of protein expression for all proteins (All) and proteins in each abundance quantile group (from low abundance to high abundance).

(C) Density plot of correlations between pairs of random proteins versus pairs of proteins within a complex.

(D) Expression ratio variation of protein complexes in the NCI-60. The x axis shows the average Pearson correlation of each protein complex across the NCI-60. The y axis shows the average abundance of proteins in a complex.

(E) Snapshot image obtained with CellMinerCDB (Rajapakse et al., 2018) protein (left image) and mRNA (right image) expression of XRCC6/KU70 and XRCC5/KU80.

(F) The correlation densities for protein pairs derived from the same complex are significantly shifted relative to those from random pairs across different resources. In each plot, the light blue density is for correlation values from random protein pairs. Densities associated with specific resources for protein complexes or stable protein interactions are indicated with different colors. The relatively lighter resource-colored density plot shows the distribution of correlation values for true protein interactors, whereas the darker one is derived from random pairs present in that database. The upper panel shows correlation values for the measured protein quantities, whereas the lower one corresponds to the mRNA levels of the same proteins. The vertical dashed red line indicates a value of the Spearman's coefficient of correlation of 0.75.

cell line identification can be readily checked with CellMinerCDB (see Figures S34–S46). The most highly correlated proteins include MARCKSL1 (myristoylated alanine-rich C kinase; r = 0.93), LGALS3 (galectin 3; r = 0.90), and ITGB1 (integrin-β1; r = 0.88) (Table S3 and Figure S34). PARP1 (poly(ADP-ribose)polymerase 1; r = 0.77) and CDK2 (cyclin-dependent kinase 2; r = 0.58) also showed high correlation. High correlations

would be expected for structural proteins and proteins with short half-lives. By contrast, some protein levels are not correlated with their transcripts (Table S3). These include TP53 (r = 0.14), TOP1 (r = 0.13), TOP2B (r = 0.09), and DHX9 (RNA helicase A; r = - 0.1; see Figure S35C). Such proteins are likely primarily regulated by post-transcriptional modifications and protein turnover.

From a translational and omic viewpoint, these results indicate that the proteins exhibiting high correlation with transcripts could be indirectly assessed by transcriptome analyses, including RNA-Seq, whereas transcriptome analyses are insufficient for the proteins that are not consistently correlated with their transcripts. In these cases, proteomic analyses, including those enabled by the SWATH-proteome, are most useful to phenotype samples. Our analyses and the CellMinerCDB tools provide insight into identifying such proteins (Table S3).

Protein Complex Predictions Based on Stoichiometry at the Protein Levels across the NCI-60

A unique benefit of proteomic data, compared with genomic and transcriptomic data, is its capacity to reveal the abundance of protein complexes and their stoichiometry (Ori et al., 2016). Our measurements included 101 predicted protein complexes comprising 1,045 proteins (Table S4) from a curated resource (Ori et al., 2016). Significantly high Pearson correlation coefficients for pairs of proteins that are part of a complex further supported the quantitative accuracy of our data matrix (Figure 2C). This was also reflected by the conserved stoichiometry of stable protein complexes, such as prefoldins (PFN1, PFN6, PFN4, and PFN5), transcription complexes (FUS, EWSR1, and DHX9), DNA repair complexes (KU70 and KU80; Figure 2E), replication and chromatin complexes, as well as membrane protein complexes (catenins and EPCAM) (Figure 2D, with additional examples in the next section).

We further investigated whether this trend was present when we used different public resources to assign protein complexes. We compared interacting protein pairs or proteins assigned to complexes according to (1) a curated CORUM database of mammalian protein complexes, (2) annotations for stable interactions in the Reactome database, (3) high-quality interaction partners in the STRING database, (4) known and modeled interactions in the Interactome3D based on available protein structures, (5) interaction pairs observed in at least three affinity purification-mass spectrometry experiments (APMS, see Methods), or (6) a small set of curated and annotated protein complexes available from the EMBL-EBI Complex Portal (Figure 2F, top panel). Compared with random protein pairs, the correlation of the measured protein quantities for the annotated interaction partners was strongly shifted to higher values (p value $<1.1 \times 10^{-10}$, Wilcoxon test). Of interest, the shift also reflected the confidence of protein complex assignments with the mean values of 0.43, 0.24, 0.21, 0.21, 0.19, and 0.13 for protein pairs from the EMBL-EBI complex portal, CORUM, Interactome3D, Reactome, STRING databases, and APMS studies, respectively. In addition, for almost all resources, a shoulder with overrepresented negative correlations ($r \leq -0.5$) was visible.

Next, we performed the same analysis, substituting transcriptomics data for protein quantities for the same protein pairs (Figure 2F, bottom panel). The co-expression of the interacting protein pairs was verified by the positive shift of correlation values for mRNA quantities. However, this shift was smaller compared with the proteomic comparisons (Figure 2F), and the right-skewed shoulder reflecting overrepresentation of highly correlating protein interactors was absent using mRNA levels (demarcated by the red vertical dashed line in Figure 2F). Moreover, the left shoulder corresponding to negatively correlated protein pairs ($r < -0.5$) at the protein level disappeared when using mRNA levels (Figure 2F). Correlations of expression values for protein interaction partners often reflect a preserved stoichiometry of protein complexes. Our comparison of mRNA and protein quantities across the NCI60 demonstrates the benefits of proteomics data for detecting protein-protein interactions.

Examples of Stoichiometric Protein Complexes

KU70 and KU80 (XRCC6 and XRCC5, respectively) form a heterodimer critical for DNA recombination, immune system maturation, DNA repair, and resistance to radiotherapy and chemotherapy. Figure 2E shows the high correlation between KU70 and KU80 protein levels across the NCI-60. Remarkably, this correlation is not seen using mRNA measurements (Figures 2E and S36), indicating that the expression of Ku is tightly regulated by post-transcriptional mechanisms independent of cancer types. Indeed, KU80 is degraded when not bound to KU70 (Chang et al., 2016; Kanungo, 2010).

Another example of a small protein complex stoichiometrically regulated across the NCI-60 is the heterotrimeric RP1/2/3 complex, which is critically important for coating single-stranded DNA during replication and repair. Using the CellMinerCDB "Compare Patterns" tool with RPA3 as the "identifier" and selecting "swa" as "Data Type" yields RPA2 as top correlate ($p = 0.75$; $r = 6.58 \times 10^{-12}$), followed by the two subunits of the MCM replicative helicase MCM5 and MCM7 ($p = 0.63$; $r = 1 \times 10^{-7}$) (Figure S37). Such highly significant correlations are not observed for the corresponding transcripts (Figure S38). We also tested proteins co-expressed with PCNA, the essential cofactor for replicative DNA polymerase processivity. Because PCNA is also included in the small number of proteins determined by reverse phase proteomic array (RPPA) (Nishizuka et al., 2003), we were able to establish the reproducibility of the SWATH measurements by plotting PCNA protein expression with SWATH vs. RPPA using CellMinerCDB ($r = 0.63$; $p = 1 \times 10^{-7}$; Figure S39). Repeating the CellMinerCDB "Compare Patterns" with PCNA, we found protein co-expression of PCNA with MCM3 and notably with FEN1, the replicative nuclease for the maturation of Okazaki fragments (Figure S38), both of which are biologically logical. To our knowledge, the stoichiometric relationship of PCNA with FEN1 has not been reported previously.

Other hypotheses and correlations can be readily found by users with the CellMinerCDB Compare Patterns tool. For instance, the large subunit of ribonucleotide reductase (RRM1) is highly correlated with the purine metabolic enzyme PAICS by SWATH (0.76, p value <1.1e-10). Notable instances detailed below also include the two RNA binding proteins involved in mRNA splicing, DHX9 and FUS, and the nucleosome remodeling complex NuRF and β-catenin (CTNNB1).

Both DHX9 (RNase A) and FUS (Fused in sarcoma and associated with liposarcoma and amyotrophic lateral sclerosis [AML]) are stoichiometrically coregulated across the NCI-60 with highly significant correlations at the protein levels ($r = 0.81$; $p = 4.7 \times 10^{-15}$) even more than at the transcript levels ($r = 0.42$; $p = 0.001$) (Figure S35). Looking further at the cells co-expressing FUS and DHX9 transcripts across the larger MGH-Sanger (GDSC) database (Garnett et al., 2012) using CellMinerCDB (Rajapakse et al., 2018) confirmed the coregulation of these two RNA binding genes across 986 cell lines ($r = 0.52$; $p = 1.8e\text{-}70$), with highest expression in leukemia, lymphomas, and small cell lung cancer cell lines (Figure S40).

For large protein complexes, the nucleosome remodeling deacetylase (NuRD) complex (Basta and Rauchman, 2015) provides a notable example of protein complex stoichiometry. NuRD consists of at least 11 proteins (Figures S41 and S47B), including the two retinoblastoma binding proteins RBB7 and RBB4, the two metastasis-associated proteins MTA3 and MTA1, the two histone deacetylases HDAC2 and HDAC1, the three methyl-CpG-binding proteins GATAD2B, GATAD2A, and MBD3, and the chromodomain helicase CHD4. All of them show a high stoichiometric correlation across the NCI-60 at the protein levels, as determined by SWATH (Figures S41 and S42). Similarly, we found stoichiometric correlation across the NCI-60 for β-catenin (CTNNB1) and its membrane-associated family members CTNND1, CTNNA1, CTNNA2, as well as EPCAM, all of which are involved in cell-cell interactions (Figures S43 and S44). Together, these examples illustrate the potential value of SWATH analyses to explore and predict stoichiometric protein complexes.

Google-Map-Based Visualization of Cancer Signaling Pathways

Our NCI-60 proteotypes cover 648 proteins in the Atlas of Cancer Signaling Networks (ACSN), a manually curated pathway database presenting published biochemical reactions involved in cancer using a Google-Maps-style visualization (Figure S48) (Kuperstein et al., 2015). When mapping the mean protein expression per cancer type, we found that, in different cell types, multiple pathways, including apoptosis, cell survival, motility, and DNA repair, displayed a similar pattern, consistent with the fact that immortal cancer cells retain cancer hallmarks in tissue culture (Hanahan and Weinberg, 2011). An example of a proteotypic pattern is the delta isoenzyme of protein kinase C, i.e. PRKCD, involved in cancer progression and a drug target (Mackay and Twelves, 2007). In agreement with PRKCD downregulation in renal clear cell carcinoma lines (Engers et al., 2000), PRKCD stood out in our visualization, with significantly lower protein expression in the NCI-60 renal carcinoma cells, relative to the average expression across the NCI-60 panel.

We also tested cellular pathways using ROMA (Representation and quantification Of Module Activities) (Martignetti et al., 2016) (Figure S48), a gene-set-based quantification algorithm. This approach revealed substantial diversity of pathway activity between different proteotypes as evidenced by two-tailed *t*-tests of activity scores (p value < 0.05). When mapping activity scores onto ACSN, some tissue specificities were

revealed, with particular cell line proteotypes displaying distinct patterns. For instance, the activity of apoptosis (with both caspases and apoptosis genes modules) was found significantly higher in ovarian cell lines (Table S5). Although there are only two prostate cancer cell lines in the panel, our analysis was able to highlight modules including "AKT-mTOR" and "Apoptosis," whose differential activity can be attributed to HSP90AA1 and PRDX. The latter protein has been independently reported to be overexpressed in prostate tumors (Ummanni et al., 2012).

Drug Response Predictions

The SWATH proteotypes covered 105 established protein targets for FDA-approved anti-cancer compounds, 661 protein targets annotated in DrugBank (Law et al., 2014) (including 68 drug-metabolizing enzymes, 5 drug carriers, and 15 drug transporters), 694 proteins linked with human diseases (Law et al., 2014; Uhlen et al., 2015), 58 protein kinases, 2 topoisomerases (TOP1 and TOP2β), and 9 tubulins. Some kinases were found to be broadly expressed with high abundance across cell lines, including MST4 and WNK1 (Figure S49), consistent with previous reports regarding their abundance (Huang et al., 2007; Lin et al., 2001). Other kinases were highly expressed in specific cell lines, for example, EGFR in the breast cancer cell line MDA-MB468, ERBB2 in the SKOV3 ovarian cell line, and CDK6 in leukemia MOLT4 cells, in agreement with previous studies using antibody-based methods (Uhlen et al., 2015; Xu et al., 2005). TOP1, TOP2β, and tubulins tended to be expressed across cell lines, consistent with their ubiquitous functions.

To assess drug response prediction, we used two main methods, an automated regression-based pipeline and a complementary interactive analysis for developing regression models (a functionality easily accessible to readers using CellMinerCDB, discover.nci.nih.gov/cellminercdb). In both cases, drug response was predicted as a weighted sum of selected feature values, where the signs and magnitudes of the feature weights indicate the direction and strength of feature influence, respectively. For the automated pipeline, we used the elastic net regression algorithm to select model features. The interactive approach involved examination of individual features that correlate with drug response (i.e. univariate models), along with features selected using the LASSO algorithm. These were integrated with experimentally established features and then assessed using the univariate and multivariate (regression models) analysis tools of the CellMinerCDB website.

Complementarity of the SWATH Proteotype with Genomic Measurement for Drug Response Predictions

First, we investigated the utility of the SWATH-based proteotype with existing genomic and transcriptomic features for 158 FDA-approved or investigational compounds in CellMiner (Luna et al., 2015; Rajapakse et al., 2018; Reinhold et al., 2012; Shankavaram et al., 2009) (Figure 3A, Table S6). A number of these compounds have been screened multiple times as they were submitted independently to the NCI-DTP and/or served as positive controls. Drugs are given unique NSC (National Service Center) identifiers for each submission to the DTP NCI-60 screen (Reinhold et al., 2012), and 47 of the compounds are represented by multiple replicates. For instance, doxorubicin is represented by two independent NSC numbers, 123127 and 759155, with Pearson correlation of 0.962 ($p = 6 \times 10^{-34}$), validating data reproducibility. Each of the compounds is categorized by mechanism-of-action annotation (Figure 3A). The largest group of drugs is DNA damaging agents (including DNA alkylating and cross-linking agents, DNA synthesis and topoisomerase inhibitors).

Using the elastic net algorithm (Barretina et al., 2012; Garnett et al., 2012; Rajapakse et al., 2018), we developed multivariate linear models to predict the NCI-60 response for each compound based on selected genomic, transcriptomic, and SWATH proteomic features. The Pearson's correlation between observed drug response values and leave-one-out cross-validation-predicted response values was applied to evaluate the performance of each predictive dataset. As different numbers of features were measured for each omics dataset, two strategies were adopted in the modeling analyses. In one case, we used all omics features, with and without the SWATH proteotype, as inputs to evaluate their general performance (see Methods). In the second, we selected 1,566 features that were available for all three molecular data types (denoted as common features). In both cases, we obtained models for 224 (93%) of the drugs (with adequate numbers of responsive lines). The predictive power achieved with all features was slightly higher than that obtained using the common features for all three data types (Figure 3B); a likely reason for this is that the latter excluded some genomic and transcriptomic features not detected at the protein level. Here, we focus on the analysis derived from all available molecular features (Figures 3H and S50, Table S7).

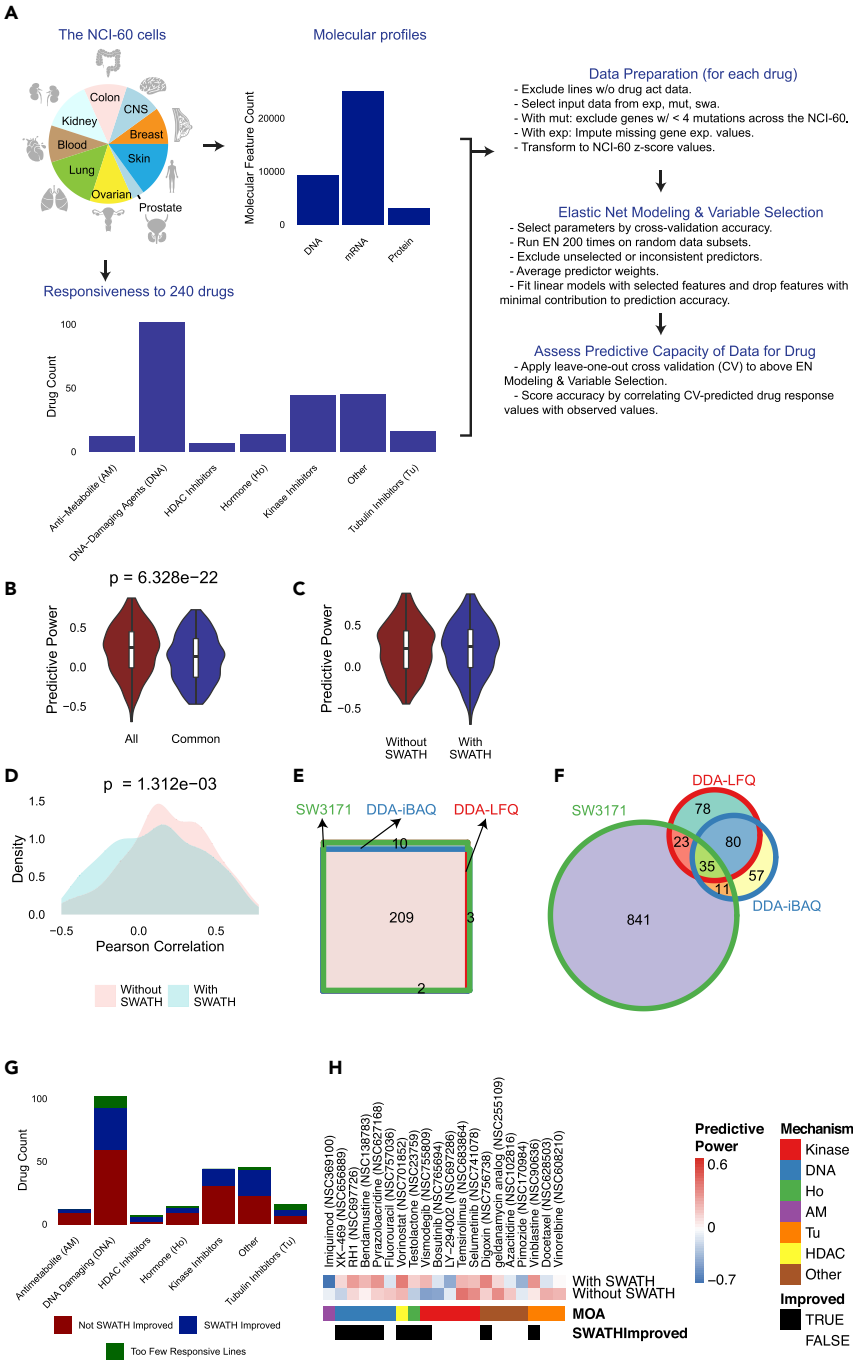


Figure 3. Prediction of Drug Responsiveness Using the NCI-60 Proteome

(A) Workflow for drug responsiveness prediction. Drug mechanism categories are shown.

(B) Distribution of predictive power (Pearson's correlation of cross-validation predicted vs. observed response) for 240 compounds using all molecular features (All) versus common features (Common) available for all molecular data types.

(C) Distribution of predictive power for molecular features (i.e. gene expression and mutation profiles) with and without the SWATH proteotype.

(D) Pearson correlation coefficient distribution of drug responsiveness predicted with and without SWATH data. p value for Kolmogorov–Smirnov test (two-sided, two-sample) was computed.

(E) Venn diagram of drugs successfully modeled using elastic net with the SWATH data containing 3,171 proteins (SW3171) and the DDA data based on iBAQ (DDA-iBAQ) and LFQ (DDA-LFQ).

(F) Venn diagram of protein predictors using the SWATH and DDA datasets.

Figure 3. *Continued*

(G) Distribution of predictive improvement using the SWATH proteotype within each mechanism class.

(H) Predictive power of different omics data combinations for the activity of 20 FDA-approved compounds based on elastic net modeling of the drug response. Each row indicates compound-specific results using gene expression and mutation input data alone or in combination with proteomic abundances; each column represents a compound. The color indicates the predictive power, measured by Pearson correlation of cross-validation predicted and observed drug response values. The top and bottom 10 drugs by difference of the absolute value of predictive power are shown. Columns specifying compound-specific response prediction accuracies are sorted by mechanism of action and whether the inclusion of the SWATH data improved the overall model.

We identified several validated predictors for drug response. For instance, mRNA expression of SLFN11 was the most dominant indicator of sensitivity to a number of DNA-targeted compounds (including FDA-approved drugs spanning platinum drugs, topoisomerase inhibitors, alkylating agents, PARP inhibitors, and DNA synthesis inhibitors), in agreement with recent reports (Barretina et al., 2012; Rajapakse et al., 2018; Zoppoli et al., 2012) (Table S7). This pipeline generated models for 224 compounds (Figure S50, Table S7). The results of these models, summarized in Figures 3G and 3H, show that predictive improvement using SWATH data was achieved across the mechanisms of actions analyzed. Given the relatively small sample size, it was not surprising that accurate predictive models could not be found for every drug (Figure 3G), particularly those with limited numbers of responsive cell lines amid a diversity of cancer types. The SWATH-MS-derived proteotypes displayed a higher percentage of predictive features than mutations and transcripts. Twelve percent of SWATH features were selected in one or more predictive models, whereas the corresponding proportions were 2% for mutation features and 6% for transcript expression features (Table S7). The responsiveness of 49 screened drugs (22%) was best predicted with SWATH data, and 83 compounds (37%) were best predicted by combining SWATH data with transcripts and mutational data.

Through an examination of predictive gain by mechanism-of-action (Figure 3G), we made a few notable observations. Out of six HDAC inhibitors, four showed improvement using the SWATH feature set making it the most improved mechanism category; this observation should be taken with caution given the limited number of compounds. Kinase inhibitors were one of the least improved categories, but this observation should be revisited using future phospho-proteomic datasets, which would include additional markers of kinase regulation (Ardito et al., 2017; Johnson, 2009). Lastly, the diverse set of compounds in the "Other" category also showed predictive improvement with the SWATH features and merits further future study. The remaining categories show a largely even distribution in predictive improvement. Next, we compared the distributions of predictive power as measured by Pearson correlation coefficients for models derived with and without the inclusion of SWATH features. The SWATH-included distribution is wider, but spans a comparable range of (higher-accuracy-associated) positive correlations relative to the strictly genomic feature-based distribution (Figures 3C and 3D). Based on these analyses, we conclude the complementarity value of the SWATH proteotype with genomic features.

Exploration of the SWATH data can also reveal secondary predictors of drug response. The protein kinase inhibitor vemurafenib (VEM, NSC 761431) yielded a multivariate model where the most prominent SWATH feature was the expression level of LAMTOR3. Although the BRAF V600E mutation is a highly significant predictor of vemurafenib activity in the NCI-60 (Abaan et al., 2013), we speculate that LAMTOR3 may be a secondary drug response predictor, although further studies outside of the scope of the current work are necessary for validation. LAMTOR3 (MP1) is part of an endosomal scaffolding complex interacting with components of the RAF/MEK/ERK mitogenic signaling pathway. LAMTOR3 binds MEK1 and ERK1, facilitating activation of the latter protein (Schaeffer et al., 1998). Elevated LAMTOR3 protein expression was correlated with vemurafenib resistance ($r = 0.44$), consistent with the hypothesis that LAMTOR3 has the capacity to enhance RAF/MEK/ERK pathway signaling downstream of RAF. Increased protein expression of LAMTOR3 was observed in two BRAF mutant cell lines, SK-MEL-5 and LOXIMVI, which are relatively resistant to vemurafenib (Abaan et al., 2013). By contrast, the two cell lines with the lowest LAMTOR3 protein expression (MALME-3M and HT29) were notably among the most sensitive to vemurafenib.

We also compared the predictive power of the DDA data from the literature (Gholami et al., 2013) with the SWATH data. Although the DDA data were able to generate multivariate models for a comparable number of drugs (Figure 3E), the number of selected protein expression predictors was lower than in the SWATH data, with some overlap (Figure 3F). The DDA datasets (Gholami et al., 2013) analyzed using iBAQ and LFQ algorithms displayed moderate overlap with each other in terms of selected model predictors (Figure 3F).

Examples of Novel SWATH Predictors for NCI-60 Drug Responses Using CellMinerCDB

Our automated analysis along with our interactive exploration with CellMinerCDB (Rajapakse et al., 2018) produced multiple predictors with plausible drug response associations. For instance, ABCC4 was a SWATH predictor for resistance to alkylating agents, including chlorambucil (NSC 3088), uracil mustard (NSC 34462), and nitrogen mustard (NSC 762) in highly ranked models (by predictive power), consistent with its established role as a drug efflux pump (Borst and Elferink, 2002). Across molecular features (mutation, transcript, or protein expression), 14 ATP-binding cassette family transporters were predictive of sensitivity to 51 compounds. P-glycoprotein (encoded by *ABCB1*), which mediates resistance to a broad range of anticancer agents (Robey et al., 2018), predicted resistance to widely used chemotherapeutic anticancer drugs including doxorubicin ($r = 0.38$; $p = 0.003$; Figure S45B) and taxol ($r = 0.43$; $p = 0.00086$; Figure S45D), as well as to the HDAC inhibitor romidepsin (depsipeptide; $r = 0.41$; $p = 0.0015$; Figure S45C) and the HSP90 inhibitor alvespimycin. ABCB1 protein expression across the NCI-60 was also positively correlated with its transcripts ($r = 0.44$; $p = 0.00044$; Figure S45A). Of note, the BCR-ABL inhibitor nilotinib was correlated with the ABC transporter protein (ABCF1) (Table S7). These results confirm the importance of measuring ABC transporters to optimize the use of anticancer agents and warrant further investigation.

Another negatively weighted SWATH predictor is CTNND1 for several compounds targeting DNA, including daunorubicin (NSC 756717), valrubicin (NSC 246131), and carmustine (NSC 409962). CTNND1 encodes δ-catenin, which promotes cell survival through activation of the WNT signaling pathway (Tang et al., 2016). Inhibition of apoptosis (Chen et al., 2001) plausibly confers drug resistance in cells with high CTNND1 protein. In addition, as discussed previously and shown in Figure S43, CTNND1 is stoichiometrically correlated with β-catenin (CTNNB1), α-catenins (CTNNA1 and CTNAA2), and EPCAM, the epithelial cell adhesion molecule, indicating a potential role of plasma membrane signaling in cellular response to DNA-targeted agents. Consistent with this hypothesis and the potential predictive value of β-catenin, analysis performed in CellMinerCDB showed significant negative correlation with etoposide ($r = -0.518$, $p = 0.000032$), topotecan ($r = -0.3$, $p = 0.02$; Figure 4), melphalan ($r = 0.534$, $p = 1.34 \times 10^{-6}$), chlorambucil ($r = -0.526$, $p = 1.85 \times 10^{-6}$), and cisplatin ($r = -0.366$, $p = 0.00254$; Figure S48). EPCAM expression was also significantly predictive of cisplatin resistance ($r = -0.44$, $p = 0.00047$; Figure S44) as was EPCAM promoter methylation ($r = -0.52$, $p = 2 \times 10^{-5}$; Figure S44).

As noted above, CellMinerCDB (discover.nci.nih.gov/CellMinerCDB) allows biologically plausible drug response correlates to be integrated within exploratory multivariate regression models. Figures 4 and S44 provide two examples using SWATH measurements. In the first, cisplatin is used as the response variable. The top response determinants are β-catenin and EPCAM, as discussed above (Figure S44). In the second example, the top predictive response determinants for the clinical topoisomerase I inhibitor topotecan are POLD1, RNASEH2B, and BAX (Figure 4A). POLD1 is the large subunit of the replicative polymerase δ and RNaseH2B one of the subunits of RNaseH2, which removes ribonucleotides misincorporated during DNA synthesis. The higher sensitivity of cells with high POLD1 and high RNaseH2 is likely reflective of hyperactive replication, which determines response to TOP1 inhibitors such as topotecan (Pommier et al., 2016). The value of β-catenin (CTNNB1) as a negative (resistance) predictor can be related to its anti-cell death activity (see prior section). Conversely, the identification of BAX, the mitochondrial pro-apoptotic effector, as a positive predictive (sensitivity) determinant at the protein level (Figure 4A) is consistent with apoptotic propensity shaping the response to DNA damaging drugs. Together, Figures 4B and 4C show this to be a predictive multivariate model at the protein, but not the RNA level, incorporating established features associated with replication stress and apoptosis.

DISCUSSION

Complementarity of protein and transcript data (Liu et al., 2016; Mertins et al., 2016; Zhang et al., 2014, 2016) can be expected to reveal new biological insights that are not apparent from the commonly used mutation and transcriptome profiles and which could be applied to enhance precision medicine. However, due to technical limitations, the acquisition of proteomic cohort datasets has been challenging. Here, using the NCI-60 cell line compendium, we demonstrate the ability of the PCT-SWATH proteomic technique to consistently quantify over 3,000 proteins across each of the NCI-60 cell lines measured in duplicate with a realistic turnaround. The data were acquired in approximately 30 working days on a single mass spectrometer, and for each sample measurement ca. 1 microgram of total peptide mass was consumed. This has been enabled by the pressure cycling technology, which minimizes sample consumption and the data-independent MS data acquisition using SWATH-MS (Guo et al., 2015).

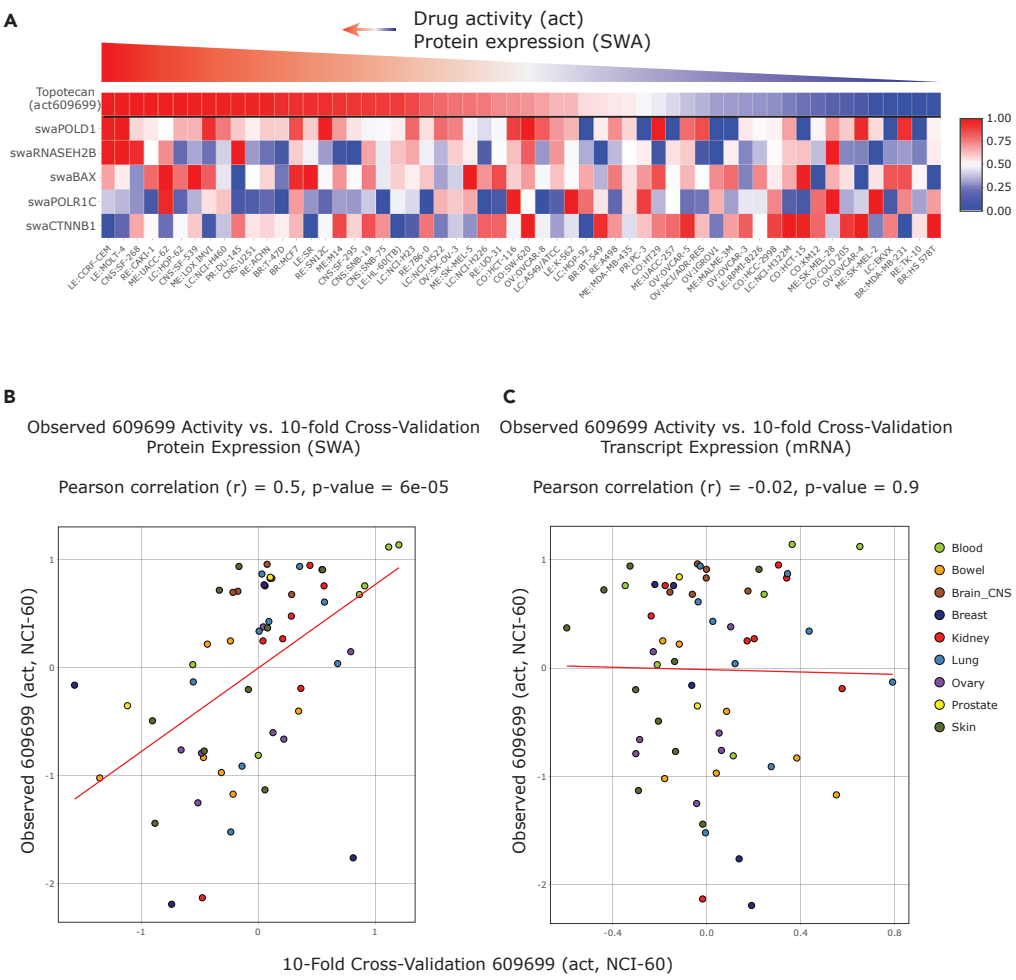


Figure 4. Pedictive Protein Biomarkers for Topotecan (NSC609699) Activity

(A) The snapshots from discover.nci.nih.gov/CellMinerCDB show results obtained with the "Multivariate Analyses" tool of CellMinerCDB using topotecan as "Response Identifier" for the query.

(B and C) Plots of the observed response values for topotecan (NSC 609699) (y axis) versus the 10-fold cross-validation predicted response values (x axis), using SWATH measurements for predictors (B) and gene expression-based predictors (C).

Because of their extensive omic annotation and drug databases including over 21,000 publicly accessible compounds, the NCI-60 are a unique platform for testing new technologies and exploring determinants and predictors of drug response (Abaan et al., 2013; Monks et al., 2018; Rajapakse et al., 2018; Reinhold et al., 2012, 2019; Zoppoli et al., 2012). In addition, most of the NCI-60 cell lines are widely used for cell biology and pharmacology (MCF-7, MDA-MB231, HCT116, HCT15, HT29, HL60, CCR-CEM, and K562 to name a few). Hence, providing MS data for over 3,000 proteins and making the data available and easy to mine through CellMinerCDB will provide a unique resource for the scientific community worldwide. This is especially true because of the overlap between the cell lines in the NCI-60 and the Broad-MIT (CCLE/CTRP) and the MGH-Sanger (GDSC) (55 and 44 of the 60 cell lines, respectively) (Rajapakse et al., 2018). The proteome of the NCI-60 has been previously measured by sample fractionation and DDA-MS analysis of over 1,000 fractionated samples (Gholami et al., 2013). However, in this early study with basic MS technology, data acquisition for each cell line required an average of about 29 h MS instrument time. By contrast, our SWATH analyses required about 140 min MS instrument time, demonstrating the feasibility of extensive human proteotypes with a throughput almost comparable with genomic and transcriptomic analyses and its potential translation to tumor material in the future. Our integration of the SWATH data into CellMinerCDB (discover.nci.nih.gov/cellminercdb) also enables the user to readily check the SWATH data for concordance with a small number of proteins measured by RPPA (Nishizuka et al., 2003) (see the example for PCNA in Figure S38).

Two aspects of our workflow ensure robust and quantitatively accurate protein expression measurements. First, we obtained technical duplicates for the entire set of NCI-60 proteotypes. This was feasible due to the unparalleled high sample-throughput of the PCT-SWATH methodology, which is gaining popularity in proteomic profiling of clinical specimens. Additionally, we developed an expert system software to further curate peptide and protein identification and quantification. Applying stringent criteria, 3,171 proteins were included for further analyses and are made available through CellMiner (discover.nci.nih.gov/cellminer and discover.nci.nih.gov/cellminercdb). The raw MS signal for each of the quantified proteins, in each cell line, was inspected by the expert system, simulating manual inspection, and is available for visual inspection in the Supplemental Information. Because the NCI-60 cell lines are widely used in cell biology, we anticipate the broad utility of these highly curated proteomic data. Additionally, our rapid proteotype acquisition pipeline using PCT-SWATH requires little biological material, making it suitable for clinical settings and in precision medicine efforts (Guo et al., 2015; Shao et al., 2015, 2016).

Compared with other omics data, proteotypes offer unique insights into the coordinated expression of protein complexes (Dudley et al., 2005; Fraser and Plotkin, 2007; Ori et al., 2016; Wang et al., 2012), which are dysregulated in many diseases, especially cancer (Le, 2015). Our high-quality proteomic data allowed a systematic investigation of the composition of 101 protein complexes in 60 cell lines. We expect that this represents a proof-of-principle for a generic, high-throughput approach, applicable to larger cancer cell line databases and clinical specimens (Guo et al., 2015), for exploring biological networks and the association of defective protein complexes with diseases and drug responses. In addition, using CellMinerCDB (discover.nci.nih.gov/cellminercdb) casual users can test the stoichiometric expression of proteins belonging to small complexes. Several examples are provided in the manuscript including dimeric complexes (XRCC6/KU70 and XRCC5/KU80; Figures 2E and S36), trimeric complexes (RPA1, RPA2 and RPA3), larger complexes such as the chromatin remodeling NuRD complex (Figures S37, S41, and S47), the β-catenin plasma membrane complexes (Figure S43), and replication complexes including MCM helicases, RPAs, PCNA, RFC4, RFC2, and FEN1 (Figures S37–S39), as well as stoichiometrically coordinated RNA binding protein complexes, such FUS and the RNase A DHX9 (Figures S35 and S40).

The NCI-60 panel has enabled many landmark discoveries, and often emerging technologies are first tested on this panel due to its diversity and depth of surrounding knowledge (Abaan et al., 2013; Barretina et al., 2012; Garnett et al., 2012; Reinhold et al., 2019; Shoemaker, 2006; Weinstein, 2012). Each cancer cell line in the NCI-60 has been tested against tens of thousands of compounds, including the FDA-approved and investigational drugs featured in our analyses. With the addition of the SWATH proteomic data, the NCI-60 remains positioned as one of the most comprehensive models for cancer research and drug discovery. The NCI-60 uniquely enabled our thorough, integrative analysis of different molecular profiles (genomic, transcriptomic, and proteomic) in predicting drug responsiveness. Our findings strengthen the body of work, highlighting the importance of integrative omics approaches in understanding drug mechanisms, and establish the benefit of large-scale proteomic measurements. Therefore, we expect our study to become a seminal work in the area of pharmacoproteomics, the benefit of which will grow with anticipated expansion of sample size, proteomic coverage including extension to phosphoproteomic expression, as well as extension to mouse models (Gao et al., 2015) and human specimens (Guo et al., 2015).

The existing SWATH data specifically enabled the use of advanced analysis techniques to produce multivariate models of drug response. Examples of multivariate analyses using CellMinerCDB are provided for topotecan and cisplatin (Figures 4 and S44) with predictive protein biomarkers including the sensitivity signature of the proapoptotic protein BAX and a resistance signature centered around β-catenin. Yet, it has been challenging to identify such signatures, and the combination of proteomic, transcriptomic, and mutation data will likely be necessary to generate predictive signatures for precision medicine. Likely, a limitation at the current technical level is the number of proteins identified, and their skewing toward the higher expressed proteins. As technology improves, and a broader group of proteins is identified; it can be anticipated that the predictive utility of the protein data will increase rapidly.

Effort was put into making our work publicly available and easily accessible through data submission to the NCI-60 CellMiner database and an accompanying R package, rcellminer (Luna et al., 2015). We expect that the analyses developed, including those based on the widely used LASSO and elastic net methods, will continue to evolve and enable future studies on additional datasets and phenotypes. Although the strengths of these methods over other related methods have been previously described (Jang et al.,

2014; Papillon-Cavanagh et al., 2013), the resulting models still require careful scrutiny by individual researchers. The interpretation of the models developed here and by others using our pipeline, should be guided by understanding of the biological activities of the associated predictors in the context of the mechanisms of action for the input drugs.

Limitations of the Study

This study and the resulting dataset invite further investigation of this unique proteomic data resource. The analysis of protein complexes in the current study highlights the value of mining beyond transcriptomic data, at the functionally critical proteomic level. The proteomic data of this study were acquired by the SWATH-MS technology, a massively parallel targeting method, at an early stage in its development. Over the last years the technology has rapidly advanced in terms of the proteomic depth that it can achieve. Our observation of the common lack of correlation between mRNA and protein expression has been similarly made in tumor samples across multiple tissue types (Kosti et al., 2016). Understanding these differences should help drive future studies in the development of mathematical and experimental models that leverage these -omics datasets effectively. Future investigations considering different approaches to handling the data, providing increased penetrance in the number of genes assessed, extending the functional implications of the data, and assessing perturbed proteomes in these cells will push back the current limitations of the field. Additionally, we have strived to make our work publicly available and easily accessible through data submission to the NCI-60 CellMiner databases (http://discover.nci.nih.gov/and http://discover.nci.nih.gov/cellminercdb) and an accompanying R package, rcellminer (Luna et al., 2015).Although the strengths of these methods over other related methods have been previously described (Jang et al., 2014; Papillon-Cavanagh et al., 2013), the resulting models still require scrutiny. The interpretation of the models developed here, and by others using our pipeline, should be guided by an understanding of the biological activities of the associated predictors in the context of the mechanisms of action and survival assays used for the input drugs.

METHODS

All methods can be found in the accompanying Transparent Methods supplemental file.

DATA AND CODE AVAILABILITY

The NCI-60 SWATH datasets and SWATH assay library has been deposited in PRIDE. Project Name: NCI60 proteome by PCT-SWATH; Project accession: PXD003539.

Reviewer account details:

Username: reviewer15254@ebi.ac.uk

Password: dWdyptzf

The protein data matrix has also been deposited in ArrayExpress. Project accession: E-PROT-2. Project title: Proteomic profiling of NCI60 cell lines from Cancer Cell Line Encyclopedia.

Reviewer account details:

Username: Reviewer_E-PROT-2

Password: gdgywGco

The protein data matrix is also accessible in CellMiner website (Reinhold et al., 2012) and R package rcellminer (Luna et al., 2015).

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.isci.2019.10.059.

ACKNOWLEDGMENTS

We thank Margot Sunshine who developed the early versions of CellMiner and Jeffrey Wang within the CCR CellMiner team for developing the current versions of CellMiner and CellMinerCDB and the NCI-DTP team (Dr. Jerry Collins and Dr. James H. Doroshow) for sharing the drug data and supporting for data acquisition from the DTP to the CellMiner team. We also thank Emanuel Gonçalves for comments on the manuscript.

The work was supported by the SystemsX.ch project PhosphoNetX PPM (to R.A.), the Swiss National Science Foundation (grant no. 3100A0-688 107679 to R.A.), the European Research Council (grants no. ERC-2008-AdG 233226 and ERC-20140AdG 670821 to R.A.), the Ruth L. Kirschstein National Research Service Award (grant no. F32 CA192901 to A.L.), the National Resource for Network Biology (NRNB) from the National Institute of General Medical Sciences (NIGMS) (grant no. P41 GM103504 to C.S.), the Center for Cancer Research, Intramural Program of the National Cancer Institute (grant no. Z01 BC006150 to Y.P), the Wellcome Trust Award (102696 to M.J.G.), the Westlake Startup Grant (to T.G.), Zhejiang Provincial Natural Science Foundation of China (grant No. LR19C050001 to T.G.), Hangzhou Agriculture and Society Advancement Program (grant No. 20190101A04 to T.G.), National Natural Science Foundation of China (General Program) (grant No. 81972492 to T.G.), and National Science Fund for Young Scholars (grant No. 21904107 to Y.Z.). A.B.E was supported by the SystemsX.ch project TbX and the National Institutes of Health project Omics4TB Disease Progression (U19 AI106761). We thank An Guo for helping in preparing the graphics.

AUTHOR CONTRIBUTIONS

T.G., R.A., and Y.P. designed and coordinated the project. C.C.K. processed the samples. L.G., C.C.K., and T.G. acquired the SWATH data. T.G. performed the SWATH data interpretation and benchmarking with help from C.C.K., and the expert system analysis with help from C.X., U.S. A.L., V.N.R., Z.W., and Y.P. performed the drug response prediction analysis and developed the reproducible research infrastructure, with critical inputs from M.P.M., J.S.R., M.J.G., S.V., W.C.R., C.S., and Y.P.. L.C. and L.M. performed the pathway analysis. A.L., V.N.R., W.C.R., S.V., and Y.P. integrated the SWATH data into rcellminer and CellMiner. A.O., M.I., R.C., C.A., M.B., and A. B.E. performed the protein complex analysis, with help from A.L., Z.W., Y.C., V.N.R., C.S., Y.S., Y.Z., Y.P.. P.Q. and Q.Z. contributed to the data analysis. W.L., H.G., R.B., J.Z., H.Z., and Y.S. performed the validation and PRM experiments. T.G., A.L., V.N.R., and Y.P. wrote the manuscript with inputs from all co-authors. R.A., Y.P., and T.G. supervised the project.

DECLARATION OF INTERESTS

R.A. holds shares of Biognosys AG, which operates in the field covered by the article. The research groups of R.A. and T.G. are supported by SCIEX, which provides access to prototype instrumentation, and Pressure Biosciences Inc., which provides access to advanced sample preparation instrumentation. All remaining authors declare no competing interests.

Received: October 11, 2019
Revised: October 21, 2019
Accepted: October 28, 2019
Published: November 22, 2019

REFERENCES

Abaan, O.D., Polley, E.C., Davis, S.R., Zhu, Y.J., Bilke, S., Walker, R.L., Pineda, M., Gindin, Y., Jiang, Y., Reinhold, W.C., et al. (2013). The exomes of the NCI-60 panel: a genomic resource for cancer biology and systems pharmacology. Cancer Res. *73*, 4372–4382.

Ardito, F., Giuliani, M., Perrone, D., Troiano, G., and Lo Muzio, L. (2017). The crucial role of protein phosphorylation in cell signaling and its use as targeted therapy (Review). Int. J. Mol. Med. *40*, 271–280.

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A.A., Kim, S., Wilson, C.J., Lehar, 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.

Basta, J., and Rauchman, M. (2015). The nucleosome remodeling and deacetylase complex in development and disease. Transl. Res. *165*, 36–47.

Bates, S.E., Eisch, R., Ling, A., Rosing, D., Turner, M., Pittaluga, S., Prince, H.M., Kirschbaum, M.H., Allen, S.L., Zain, J., et al. (2015). Romidepsin in peripheral and cutaneous T-cell lymphoma: mechanistic implications from clinical and correlative data. Br. J. Haematol. *170*, 96–109.

Borst, P., and Elferink, R.O. (2002). Mammalian ABC transporters in health and disease. Annu. Rev. Biochem. *71*, 537–592.

Brazma, A., Parkinson, H., Sarkans, U., Shojatalab, M., Vilo, J., Abeygunawardena, N., Holloway, E., Kapushesky, M., Kemmeren, P., Lara, G.G., et al. (2003). ArrayExpress–a public repository for microarray gene expression data at the EBI. Nucleic Acids Res. *31*, 68–71.

Burton, J.H., Mazcko, C.N., LeBlanc, A.K., Covey, J.M., Ji, J.J., Kinders, R.J., Parchment, R.E., Khanna, C., Paoloni, M., Lana, S.E., et al. (2018). NCI Comparative Oncology Program testing of non-camptothecin indenoisoquinoline topoisomerase i inhibitors in naturally occurring canine lymphoma. Clin. Cancer Res. *24*, 5830–5840.

Cancer Genome Atlas Research Network, Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J.M. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. Nat. Genet. *45*, 1113–1120.

Chang, H.W., Nam, H.Y., Kim, H.J., Moon, S.Y., Kim, M.R., Lee, M., Kim, G.C., Kim, S.W., and Kim, S.Y. (2016). Effect of beta-catenin silencing in overcoming radioresistance of head and neck cancer cells by antagonizing the effects of AMPK on Ku70/Ku80. Head Neck *38* (*Suppl 1*), E1909–E1917.

Chen, S., Guttridge, D.C., You, Z., Zhang, Z., Fribley, A., Mayo, M.W., Kitajewski, J., and Wang, C.Y. (2001). Wnt-1 signaling inhibits apoptosis by activating beta-catenin/T cell factor-mediated transcription. J. Cell Biol. *152*, 87–96.

Dudley, A.M., Janse, D.M., Tanay, A., Shamir, R., and Church, G.M. (2005). A global view of pleiotropy and phenotypically derived gene function in yeast. Mol. Syst. Biol. *1*, 2005 0001.

Engers, R., Mrzyk, S., Springer, E., Fabbro, D., Weissgerber, G., Gernharz, C.D., and Gabbert, H.E. (2000). Protein kinase C in human renal cell carcinomas: role in invasion and differential isoenzyme expression. Br. J. Cancer *82*, 1063–1069.

Fojo, T., Farrell, N., Ortuzar, W., Tanimura, H., Weinstein, J., and Myers, T.G. (2005). Identification of non-cross-resistant platinum compounds with novel cytotoxicity profiles using the NCI anticancer drug screen and clustered image map visualizations. Crit. Rev. Oncol. Hematol. *53*, 25–34.

Fraser, H.B., and Plotkin, J.B. (2007). Using protein complexes to predict phenotypic effects of gene mutation. Genome Biol. *8*, R252.

Gao, H., Korn, J.M., Ferretti, S., Monahan, J.E., Wang, Y., Singh, M., Zhang, C., Schnell, C., Yang, G., Zhang, Y., et al. (2015). High-throughput screening using patient-derived tumor xenografts to predict clinical trial drug response. Nat. Med. *21*, 1318–1325.

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.

Gholami, A.M., Hahne, H., Wu, Z.X., 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., Rost, 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.

Guo, T., Kouvonen, P., Koh, C.C., Gillet, L.C., Wolski, W.E., Rost, 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.

Hanahan, D., and Weinberg, R.A. (2011). Hallmarks of cancer: the next generation. Cell *144*, 646–674.

Holbeck, S.L., Collins, J.M., and Doroshow, J.H. (2010). Analysis of Food and Drug Administration-approved anticancer agents in the NCI60 panel of human tumor cell lines. Mol. Cancer Ther. *9*, 1451–1460.

Huang, C.L., Cha, S.K., Wang, H.R., Xie, J., and Cobb, M.H. (2007). WNKs: protein kinases with a unique kinase domain. Exp. Mol. Med. *39*, 565–573.

Jang, I.S., Neto, E.C., Guinney, J., Friend, S.H., and Margolin, A.A. (2014). Systematic assessment of analytical methods for drug sensitivity prediction from cancer cell line data. Pac. Symp. Biocomput. *19*, 63–74.

Johnson, L.N. (2009). The regulation of protein phosphorylation. Biochem. Soc. Trans. *37*, 627–641.

Jones, P., Cote, R.G., Martens, L., Quinn, A.F., Taylor, C.F., Derache, W., Hermjakob, H., and Apweiler, R. (2006). PRIDE: a public repository of protein and peptide identifications for the proteomics community. Nucleic Acids Res. *34*, D659–D663.

Kanungo, J. (2010). Exogenously expressed human Ku70 stabilizes Ku80 in Xenopus oocytes and induces heterologous DNA-PK catalytic activity. Mol. Cell Biochem. *338*, 291–298.

Keller, A., Nesvizhskii, A.I., Kolker, E., and Aebersold, R. (2002). Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Anal. Chem. *74*, 5383–5392.

Kosti, I., Jain, N., Aran, D., Butte, A.J., and Sirota, M. (2016). Cross-tissue analysis of gene and protein expression in normal and cancer tissues. Sci. Rep. *6*, 24799.

Kuperstein, I., Bonnet, E., Nguyen, H.A., Cohen, D., Viara, E., Grieco, L., Fourquet, S., Calzone, L., Russo, C., Kondratova, M., et al. (2015). Atlas of cancer signalling network: a systems biology resource for integrative analysis of cancer data with google maps. Oncogenesis *4*, e160.

Law, V., Knox, C., Djoumbou, Y., Jewison, T., Guo, A.C., Liu, Y., Maciejewski, A., Arndt, D., Wilson, M., Neveu, V., et al. (2014). DrugBank 4.0: shedding new light on drug metabolism. Nucleic Acids Res. *42*, D1091–D1097.

Le, D.H. (2015). A novel method for identifying disease associated protein complexes based on functional similarity protein complex networks. Algorithms Mol. Biol. *10*, 14.

Lin, J.L., Chen, H.C., Fang, H.I., Robinson, D., Kung, H.J., and Shih, H.M. (2001). MST4, a new Ste20-related kinase that mediates cell growth and transformation via modulating ERK pathway. Oncogene *20*, 6559–6569.

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

Luna, A., Rajapakse, V.N., Sousa, F.G., Gao, J., Schultz, N., Varma, S., Reinhold, W., Sander, C., and Pommier, Y. (2015). rcellminer: exploring molecular profiles and drug response of the NCI-60 cell lines in R. Bioinformatics *32*, 1272–1274.

Mackay, H.J., and Twelves, C.J. (2007). Targeting the protein kinase C family: are we there yet? Nat. Rev. Cancer *7*, 554–562.

Martignetti, L., Calzone, L., Bonnet, E., Barillot, E., and Zinovyev, A. (2016). ROMA: representation and quantification of module activity from target expression data. Front. Genet. *7*, 18.

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.

Monks, A., Zhao, Y., Hose, C., Hamed, H., Krushkal, J., Fang, J., Sonkin, D., Palmisano, A., Polley, E.C., Fogli, L.K., et al. (2018). The NCI transcriptional pharmacodynamics workbench: a tool to examine dynamic expression profiling of therapeutic response in the NCI-60 cell line panel. Cancer Res. *78*, 6807–6817.

Nishizuka, S., Charboneau, L., Young, L., Major, S., Reinhold, W.C., Waltham, M., Kouros-Mehr, H., Bussey, K.J., Lee, J.K., Espina, V., et al. (2003). Proteomic profiling of the NCI-60 cancer cell lines using new high-density reverse-phase lysate microarrays. Proc. Natl. Acad. Sci. U S A *100*, 14229–14234.

Ori, A., Iskar, M., Buczak, K., Kastritis, P., Parca, L., Andres-Pons, A., Singer, S., Bork, P., and Beck, M. (2016). Spatiotemporal variation of mammalian protein complex stoichiometries. Genome Biol. *17*, 47.

Papillon-Cavanagh, S., De Jay, N., Hachem, N., Olsen, C., Bontempi, G., Aerts, H.J., Quackenbush, J., and Haibe-Kains, B. (2013). Comparison and validation of genomic predictors for anticancer drug sensitivity. J. Am. Med. Inform. Assoc. *20*, 597–602.

Picotti, P., and Aebersold, R. (2012). Selected reaction monitoring-based proteomics: workflows, potential, pitfalls and future directions. Nat. Methods *9*, 555–566.

Pommier, Y., Sun, Y., Huang, S.N., and Nitiss, J.L. (2016). Roles of eukaryotic topoisomerases in transcription, replication and genomic stability. Nat. Rev. Mol. Cell Biol. *17*, 703–721.

Powell, B.S., Lazarev, A.V., Carlson, G., Ivanov, A.R., and Rozak, D.A. (2012). Pressure cycling technology in systems biology. Methods Mol. Biol. *881*, 27–62.

Rajapakse, V.N., Luna, A., Yamade, M., Loman, L., Varma, S., Sunshine, M., Iorio, F., Sousa, F.G., Elloumi, F., Aladjem, M.I., et al. (2018). CellMinerCDB for integrative cross-database genomics and pharmacogenomics analyses of cancer cell lines. iScience *10*, 247–264.

Reinhold, W.C., Sunshine, M., Liu, H.F., Varma, S., Kohn, K.W., Morris, J., Doroshow, J., and Pommier, Y. (2012). CellMiner: a web-based suite of genomic and pharmacologic tools to explore transcript and drug patterns in the NCI-60 cell line set. Cancer Res. *72*, 3499–3511.

Reinhold, W.C., Varma, S., Sunshine, M., Elloumi, F., Ofori-Atta, K., Lee, S., Trepel, J.B., Meltzer, P.S., Doroshow, J.H., and Pommier, Y. (2019). RNA sequencing of the NCI-60: integration into CellMiner and CellMiner CDB. Cancer Res. *79*, 3514–3524.

Robey, R.W., Pluchino, K.M., Hall, M.D., Fojo, A.T., Bates, S.E., and Gottesman, M.M. (2018). Revisiting the role of ABC transporters in multidrug-resistant cancer. Nat. Rev. Cancer *18*, 452–464.

Rost, H.L., Rosenberger, G., Navarro, P., Gillet, L., Miladinovic, S.M., Schubert, O.T., Wolskit, W., Collins, B.C., Malmstrom, J., Malmstrom, L., et al. (2014). OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. Nat. Biotechnol. *32*, 219–223.

Schaeffer, H.J., Catling, A.D., Eblen, S.T., Collier, L.S., Krauss, A., and Weber, M.J. (1998). MP1: a MEK binding partner that enhances enzymatic

activation of the MAP kinase cascade. Science *281*, 1668–1671.

Shankavaram, U.T., Varma, S., Kane, D., Sunshine, M., Chary, K.K., Reinhold, W.C., Pommier, Y., and Weinstein, J.N. (2009). CellMiner: a relational database and query tool for the NCI-60 cancer cell lines. BMC Genomics *10*, 277.

Shao, S., Guo, T., Gross, V., Lazarev, A., Koh, C.C., Gillessen, S., Joerger, M., Jochum, W., and Aebersold, R. (2016). Reproducible tissue homogenization and protein extraction for quantitative proteomics using micropestle-assisted pressure-cycling technology. J. Proteome Res. *15*, 1821–1829.

Shao, S., Guo, T., Koh, C.C., Gillessen, S., Joerger, M., Jochum, W., and Aebersold, R. (2015). Minimal sample requirement for highly multiplexed protein quantification in cell lines and tissues by PCT-SWATH mass spectrometry. Proteomics *15*, 3711–3721.

Shoemaker, R.H. (2006). The NCI60 human tumour cell line anticancer drug screen. Nat. Rev. Cancer *6*, 813–823.

Tang, B., Tang, F., Wang, Z., Qi, G., Liang, X., Li, B., Yuan, S., Liu, J., Yu, S., and He, S. (2016). Overexpression of CTNND1 in hepatocellular carcinoma promotes carcinous characters through activation of Wnt/beta-catenin signaling. J. Exp. Clin. Cancer Res. *35*, 82.

Uhlen, M., Fagerberg, L., Hallstrom, B.M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, A., Kampf, C., Sjostedt, E., Asplund, A., et al. (2015). Proteomics. Tissue-based map of the human proteome. Science *347*, 1260419.

Ummanni, R., Barreto, F., Venz, S., Scharf, C., Barett, C., Mannsperger, H.A., Brase, J.C., Kuner, R., Schlomm, T., Sauter, G., et al. (2012). Peroxiredoxins 3 and 4 are overexpressed in prostate cancer tissue and affect the proliferation of prostate cancer cells in vitro. J. Proteome Res. *11*, 2452–2466.

Wang, Q., Liu, W., Ning, S., Ye, J., Huang, T., Li, Y., Wang, P., Shi, H., and Li, X. (2012). Community of protein complexes impacts disease association. Eur. J. Hum. Genet. *20*, 1162–1167.

Weinstein, J.N. (2012). Drug discovery: cell lines battle cancer. Nature *483*, 544–545.

Xu, H., Yu, Y., Marciniak, D., Rishi, A.K., Sarkar, F.H., Kucuk, O., and Majumdar, A.P. (2005). Epidermal growth factor receptor (EGFR)-related protein inhibits multiple members of the EGFR family in colon and breast cancer cells. Mol. Cancer Ther. *4*, 435–442.

Zhang, B., Wang, J., Wang, X., Zhu, J., Liu, Q., 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.

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.

Zoppoli, G., Regairaz, M., Leo, E., Reinhold, W.C., Varma, S., Ballestrero, A., Doroshow, J.H., and Pommier, Y. (2012). Putative DNA/RNA helicase Schlafen-11 (SLFN11) sensitizes cancer cells to DNA-damaging agents. Proc. Natl. Acad. Sci. U S A *109*, 15030–15035.

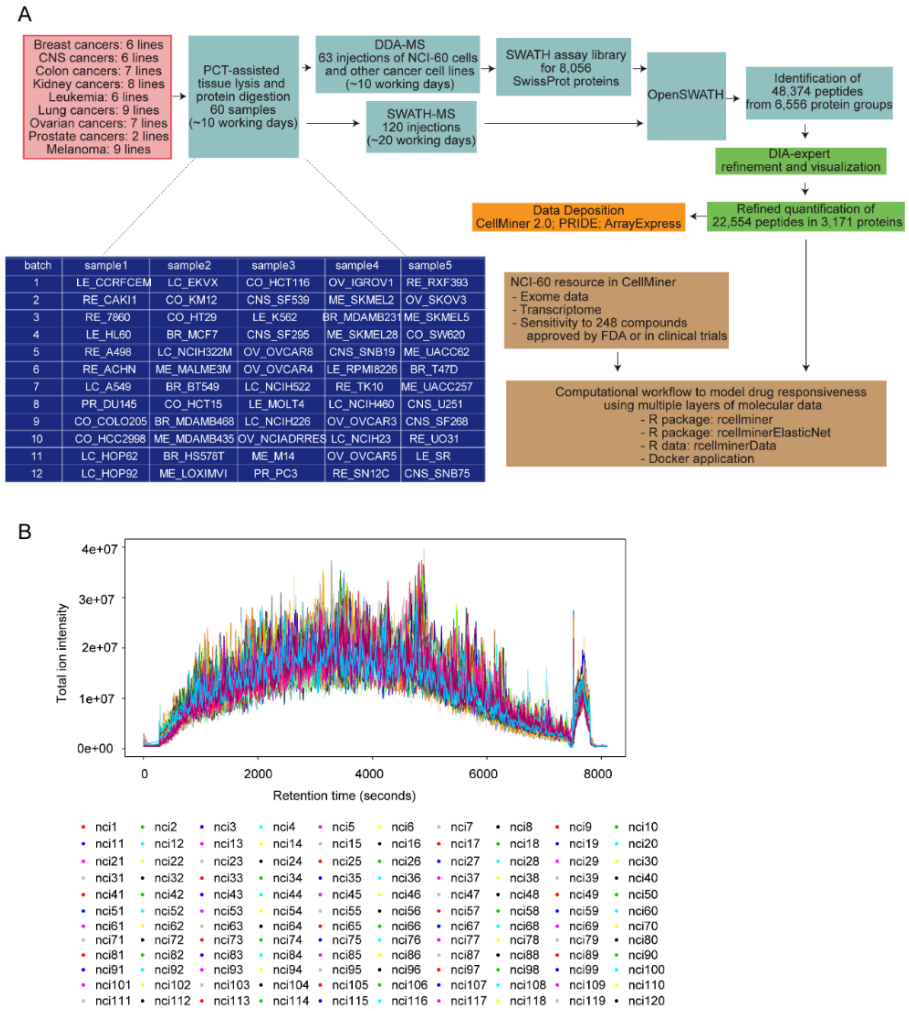
ISCI, Volume *21*

Supplemental Information

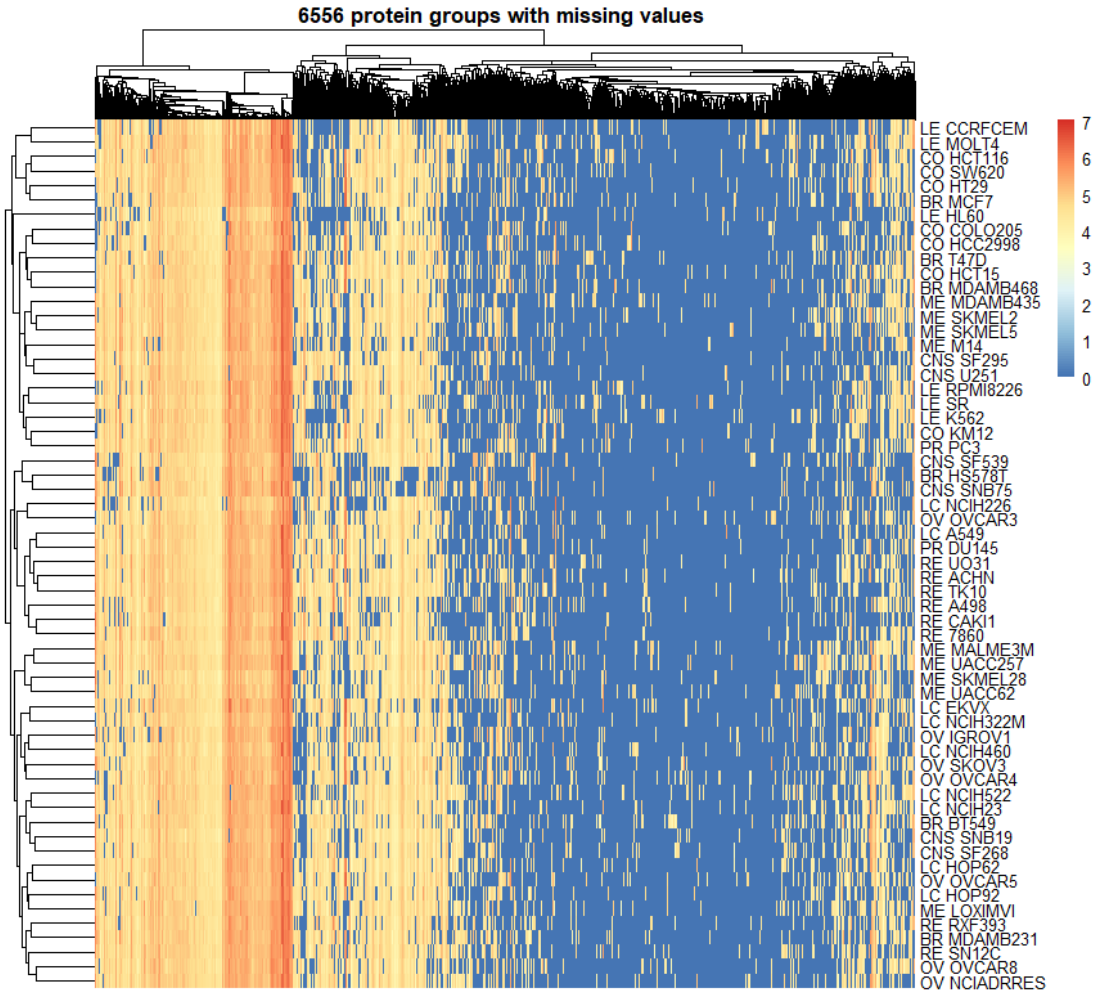
Quantitative Proteome Landscape of the NCI-60 Cancer Cell Lines

Tiannan Guo, Augustin Luna, Vinodh N. Rajapakse, Ching Chiek Koh, Zhicheng Wu, Wei Liu, Yaoting Sun, Huanhuan Gao, Michael P. Menden, Chao Xu, Laurence Calzone, Loredana Martignetti, Chiara Auwerx, Marija Buljan, Amir Banaei-Esfahani, Alessandro Ori, Murat Iskar, Ludovic Gillet, Ran Bi, Jiangnan Zhang, Huanhuan Zhang, Chenhuan Yu, Qing Zhong, Sudhir Varma, Uwe Schmitt, Peng Qiu, Qiushi Zhang, Yi Zhu, Peter J. Wild, Mathew J. Garnett, Peer Bork, Martin Beck, Kexin Liu, Julio Saez-Rodriguez, Fathi Elloumi, William C. Reinhold, Chris Sander, Yves Pommier, and Ruedi Aebersold

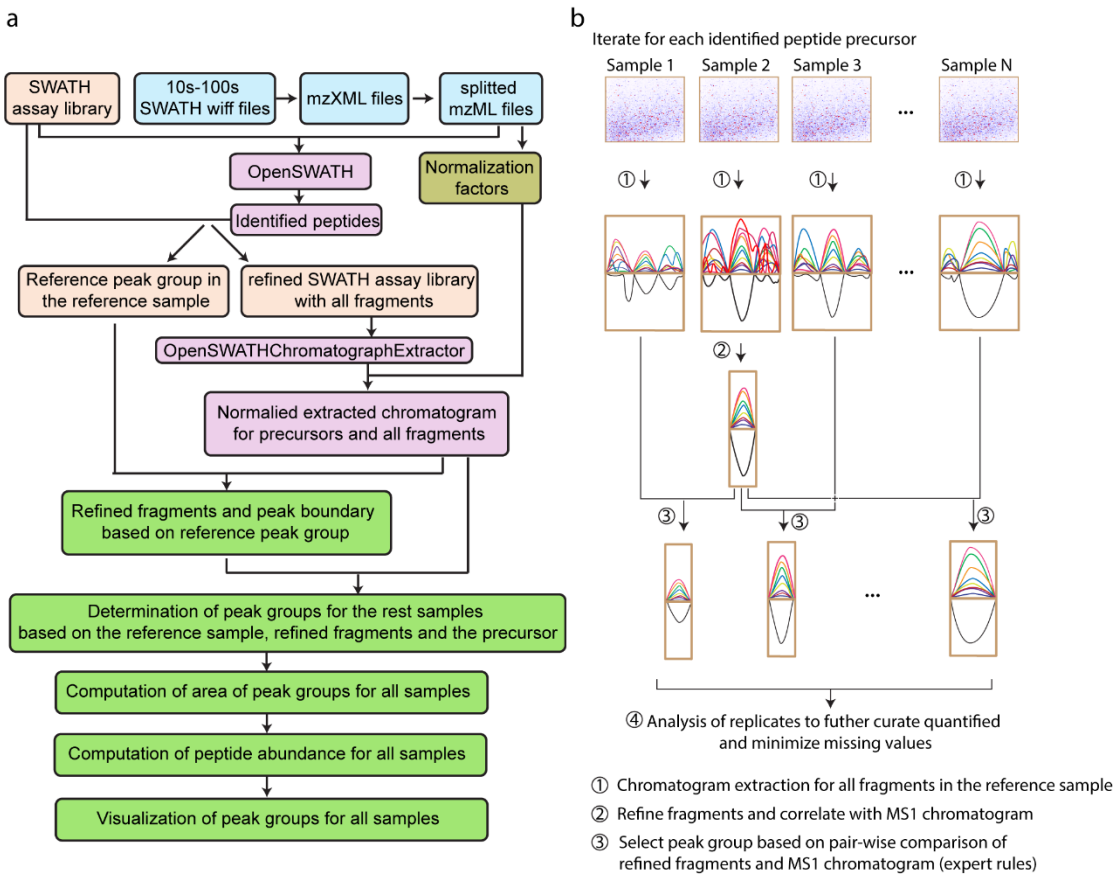
Supplementary Figures



Supplementary Figure 1, Related to Figure 1. Workflow for generating the NCI-60 proteome maps and predicting phenotypes. (A) Flowchart of experimental design. The NCI-60 cell pellets were divided into 12 batches, lysed and digested using the PCT method. The peptides were first analyzed in DDA mode to build a SWATH assay library. We also included DDA files from U2OS and HeLa cell digests. In total we performed 63 DDA injections either from whole cell lysate or fractionated samples. Each sample was analyzed in SWATH mode twice. The SWATH data were processed using software tools including OpenSWATH and DIA-expert in sequence. Our data were deposited in several public databases including CellMiner 2.0. Subsequently, we developed a computational workflow to model drug responsiveness using multiple layers of molecular data. The generation of a spectral library specifically for the NCI-60 cells consumed ca. 10 working days. For studies of this type this step is optional because similar results can be obtained from the use of publicly accessible, extensive human spectral libraries such as the pan-human library (Rosenberger et al., 2014). (**B**) raw mass spectrometric signal for the 120 SWATH runs. Total ion chromatography graphs are shown. The index of the 120 NCI SWATH files is explained in **Supplementary Table 1**.



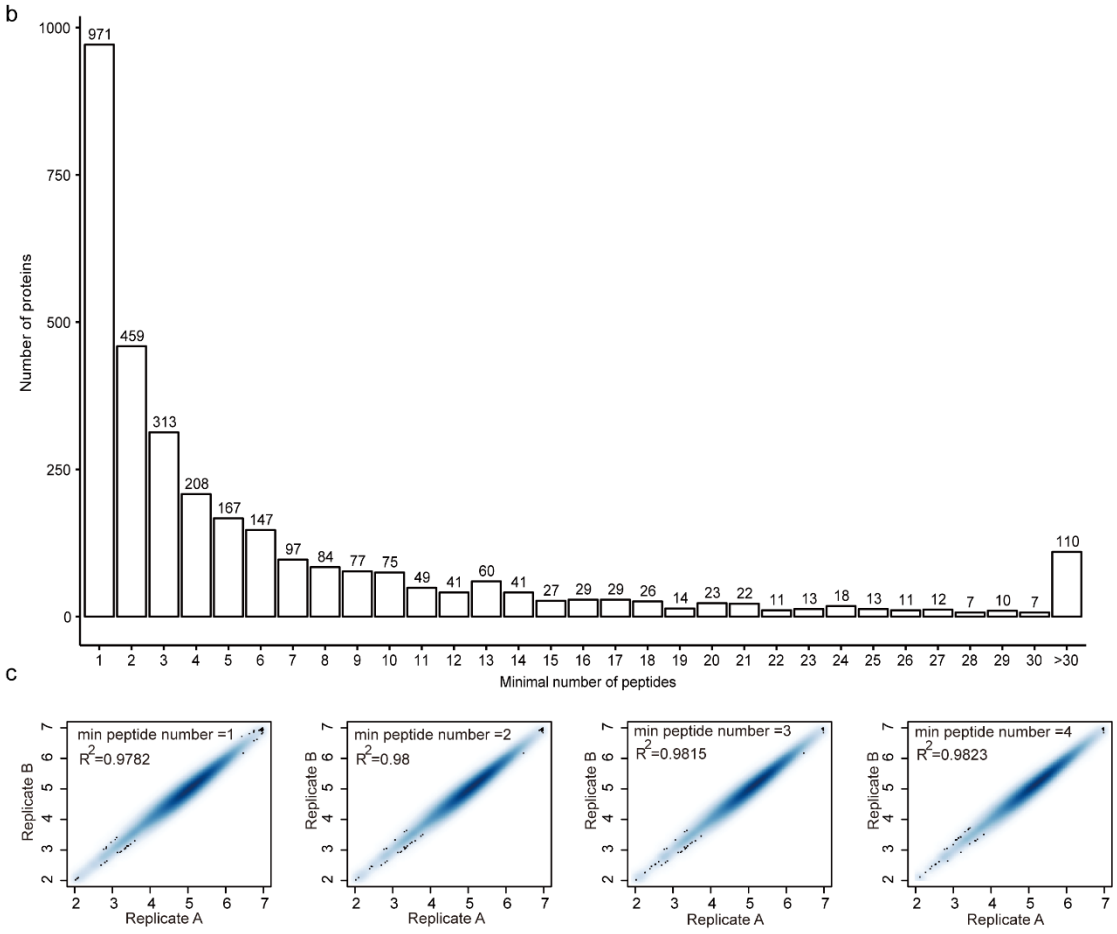
Supplementary Figure 2, Related to Figure 1. Unsupervised clustering of 6556 protein groups identified and quantified in the NCI-60 cells. Using the SWATH library containing 8056 protein groups, we displayed the identified and quantified protein groups after unsupervised clustering of both cells and proteins based on their log10 transformed intensity values.



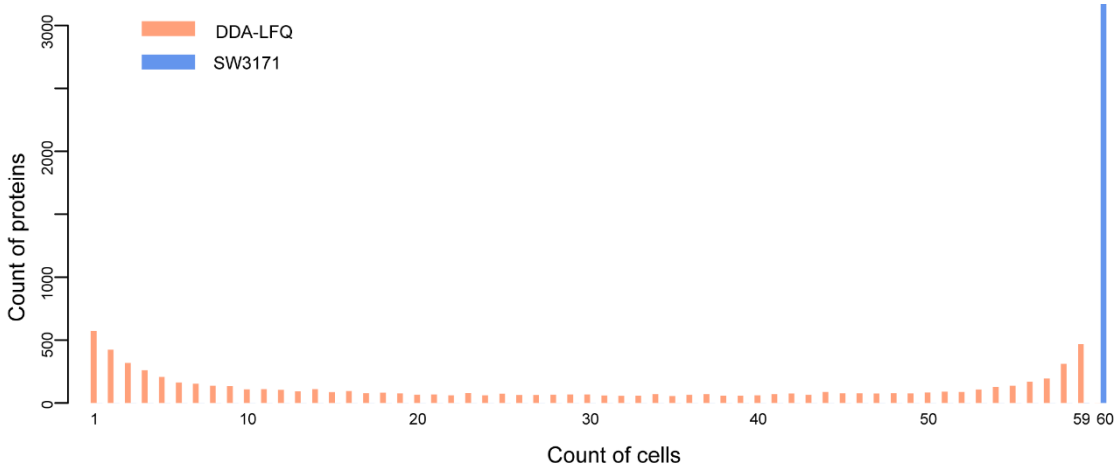
Supplementary Figure 3, Related to Figure 1. Design and implementation of DIA-expert, Related to Figure 1. (**a**) DIA-experts reads output data from OpenSWATH analysis of SWATH/DIA maps and then curates and visualizes quantitative ion chromatogram signals. (**b**) DIA-expert analyses each identified peptide precursor in a sample set. In Step ① it extracts ion chromatography signals for any number of fragments and the precursor ion chromatogram for all samples. In Step ②, it selects reference sample(s) from the sample set and refines non-contaminated chromatographic signals by learning the signal characteristics of the reference sample(s). In step 3 the system performs pair-wise comparisons of the reference sample(s) and a sample to be quantified based on the refined fragments ion set. Last, replicates of each sample and proteotypic peptides from the same protein were considered to exclude unreliably quantified peptides and minimize missing values for protein quantification across the entire data set.

a

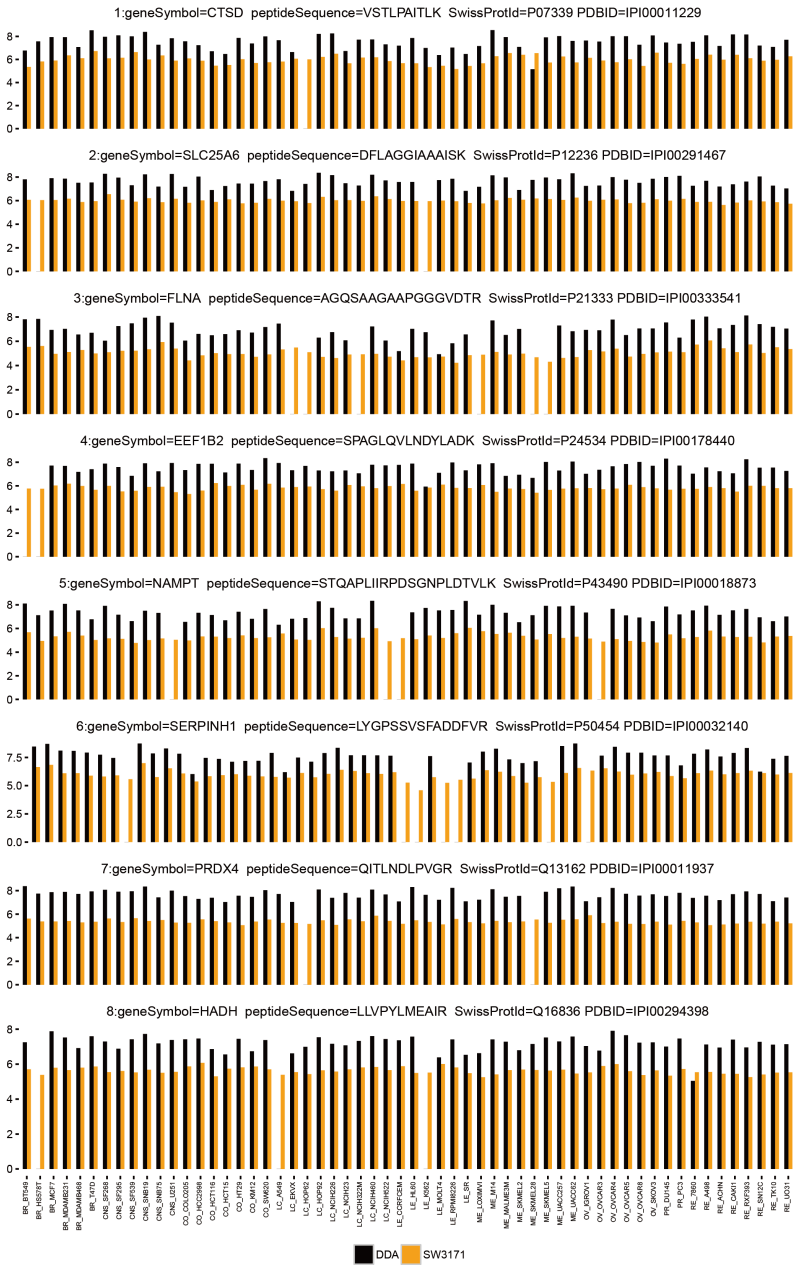
Min number of peptides	Number of proteins	R2
1	3171	0.9872
2	2200	0.9800
3	1741	0.9815
4	1428	0.9823



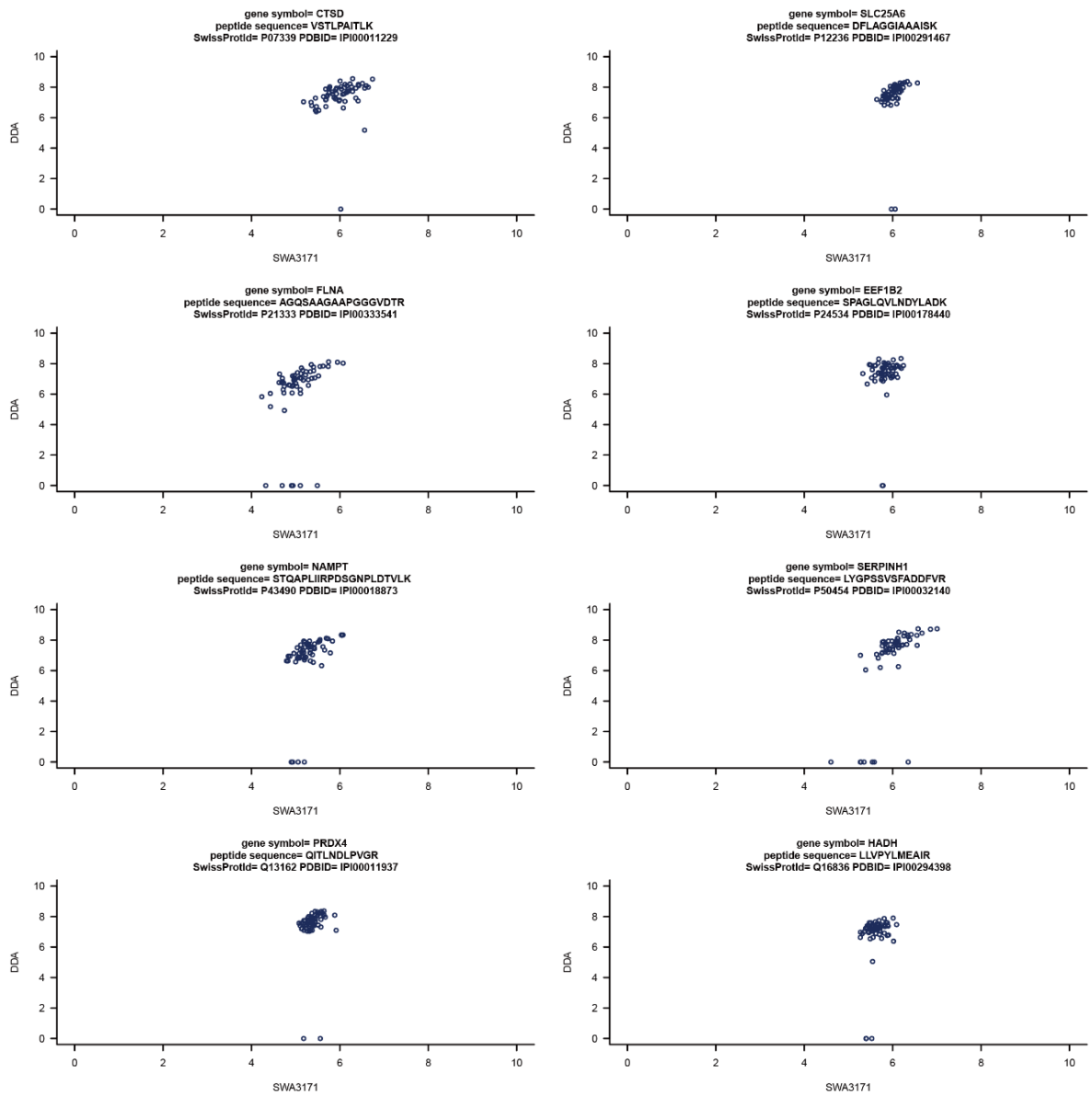
Supplementary Figure 4, Related to Figure 1. Quantitative accuracy of the NCI-60 proteome as a function of the number of peptides quantified per protein. (**a**) Number of proteins quantified when minimally 1, 2, 3 and 4 peptides were quantified per protein. The R2 values of technical replicates are computed. (**b**) Distribution of protein numbers based on increasing number of peptides. (**c**) The heatmap scatter plot of proteins quantified in two technical replicates when the minimal peptide number is limited to 1, 2, 3 and 4.



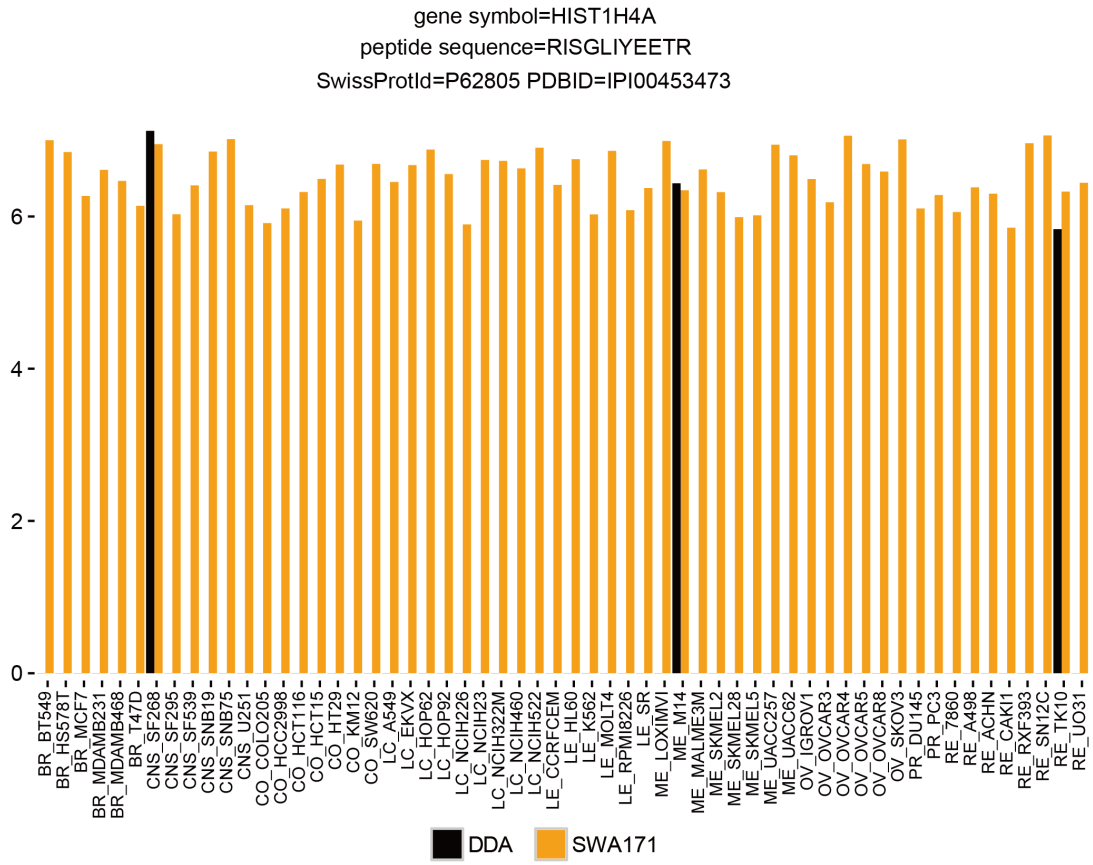
Supplementary Figure 5, Related to Figure 2. Count of proteins quantified in increasing number of cells. This plot shows the number of proteins quantified in the NCI-60 cells. DDA-LFQ denotes the LFQ-processed DDA data of the NCI-60 cells. SW3171 means the SWATH data set presented in this study. Most of the SW3171 proteins were quantified in all 60 cells. In DDA-LFQ data set (Gholami et al., 2013b), highest numbers of IPI protein groups were quantified in 1 and 59 cells.



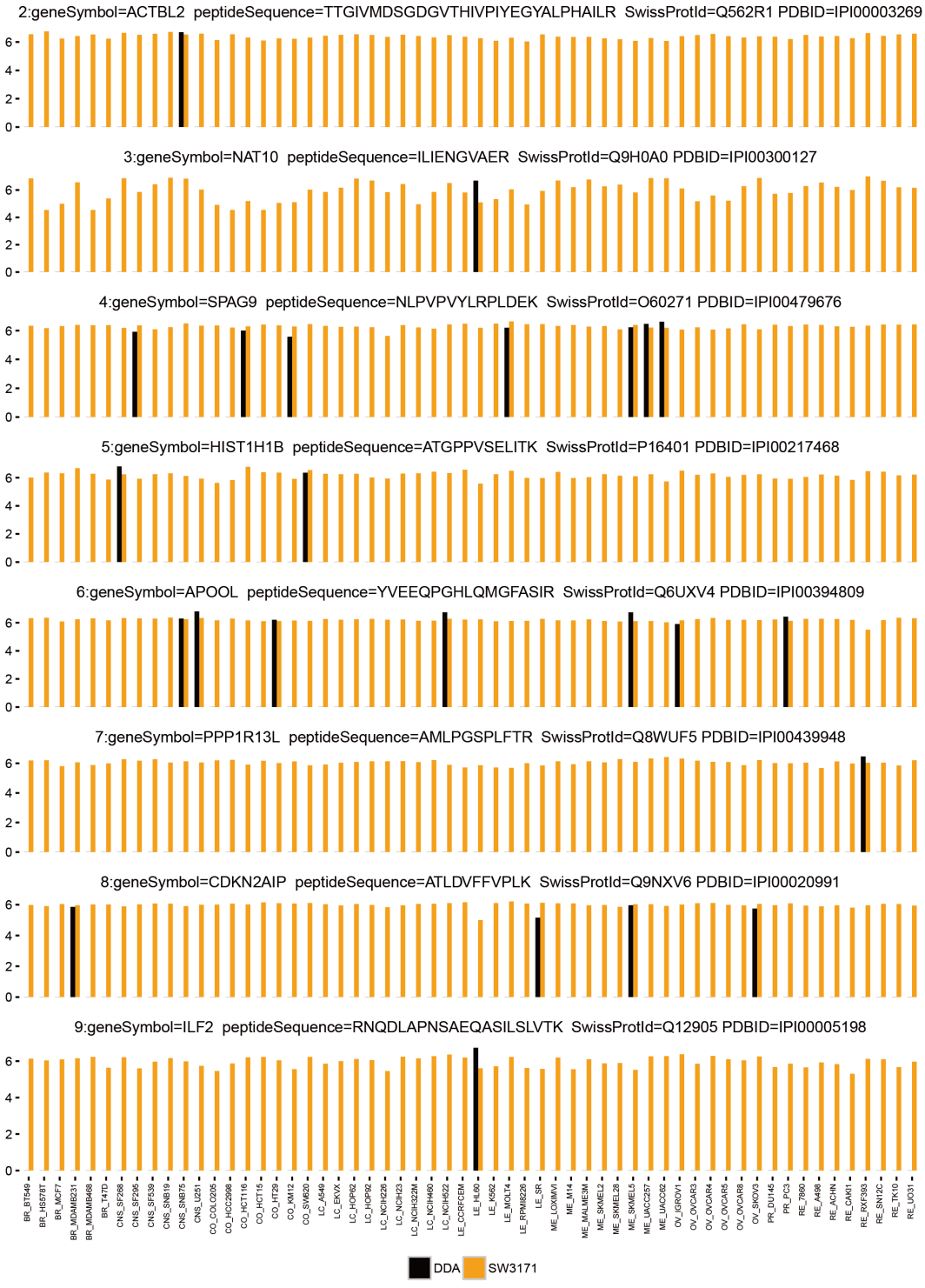
Supplementary Figure 6, Related to Figure 2. Comparison of 8 representative proteins which have been consistently quantified across nearly all NCI-60 cell lines by DDA. The data are shown in bar plots. Protein intensity values are log10 scaled



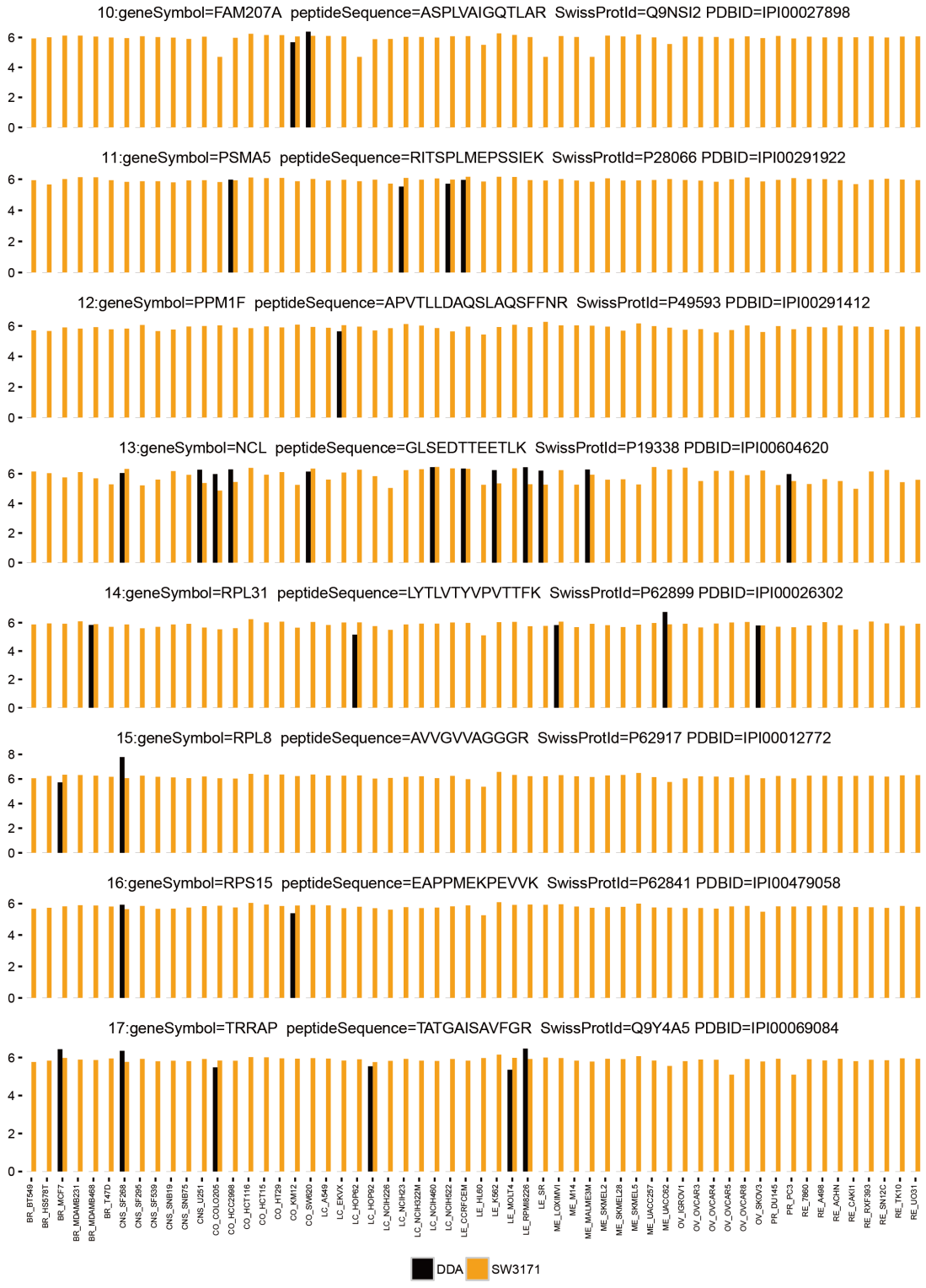
Supplementary Figure 7, Related to Figure 2. Comparison of 8 representative proteins which have been consistently quantified across nearly all NCI-60 cell lines by SWATH. The data are shown in scatter plots. Protein intensity values are log10 scaled.



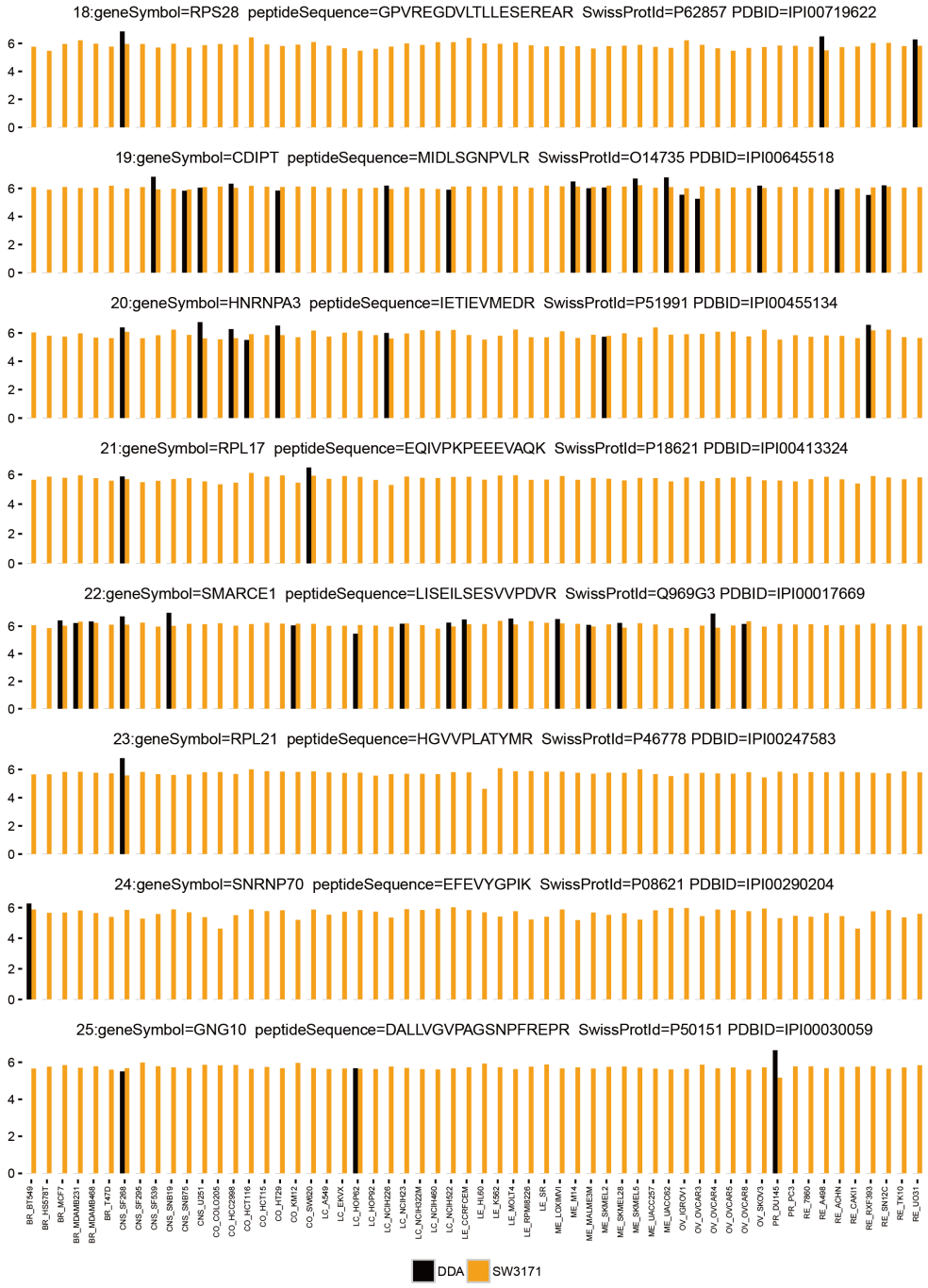
Supplementary Figure 8, Related to Figure 2. Bar plots for P62805, which is quantified cross all NCI-60 cell lines by SWATH but not quantified by DDA.



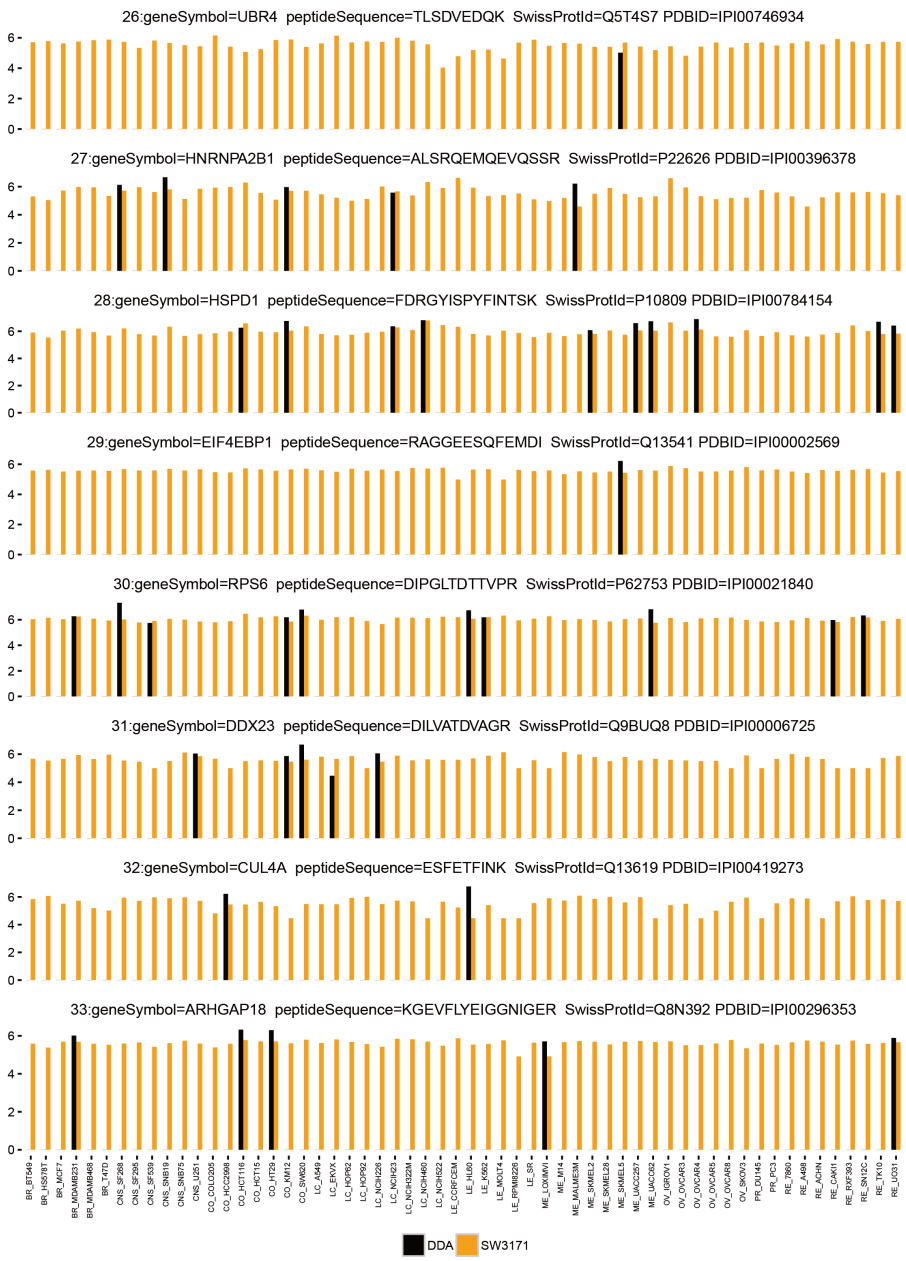
Supplementary Figure 9, Related to Figure 2. Bar plots for Q562R1, Q9H0A0, O60271, P16401, Q6UXV4, Q8WUF5, Q9NXV6 and Q12905, which are quantified cross all NCI-60 cell lines by SWATH but not quantified by DDA.



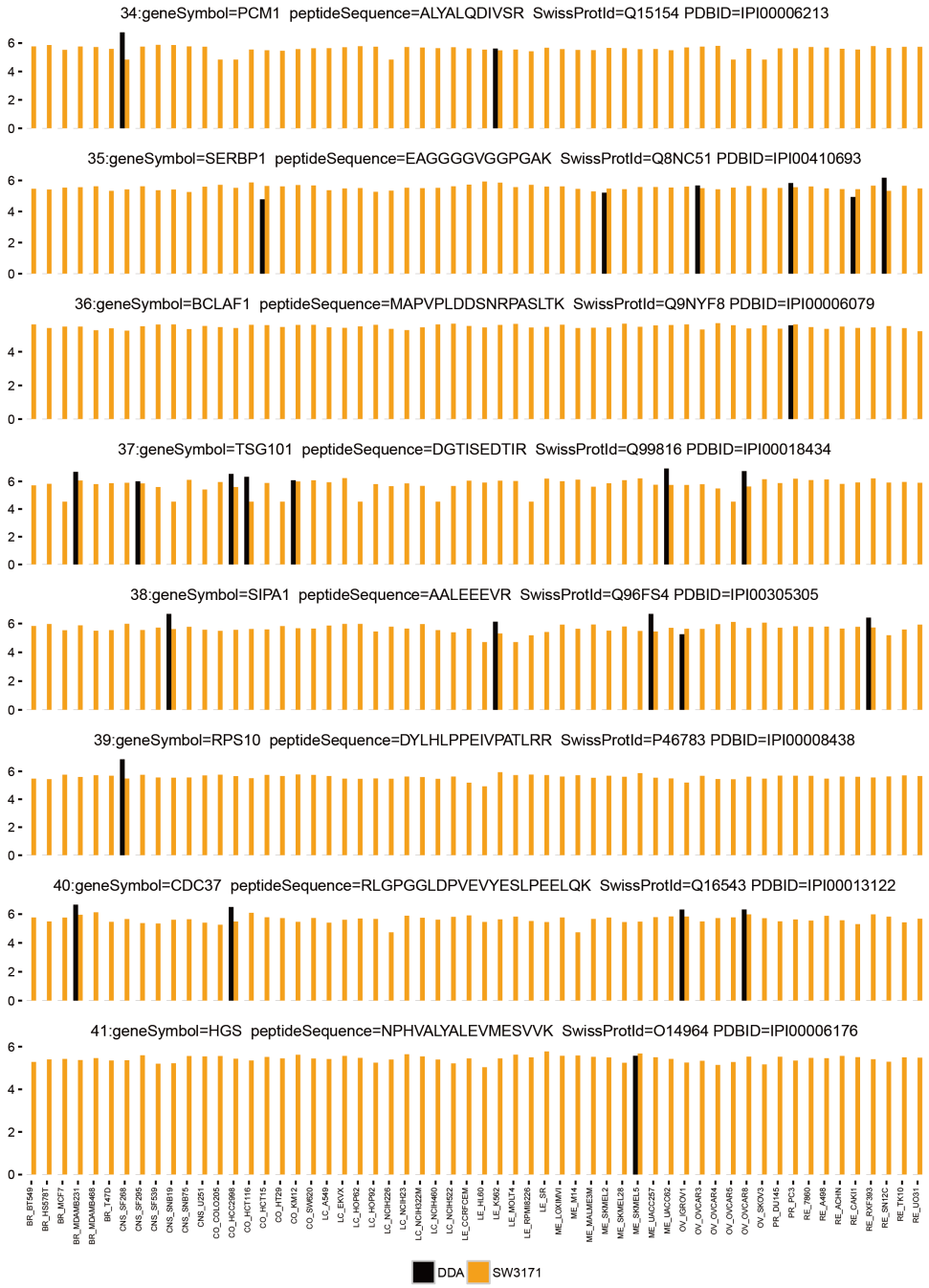
Supplementary Figure 10, Related to Figure 2. Bar plots for Q9NSI2, P28066, P49593, P19338, P62899, P62917, P62841, and Q9Y4A5, which are quantified cross all NCI-60 cell lines by SWATH but not quantified by DDA.



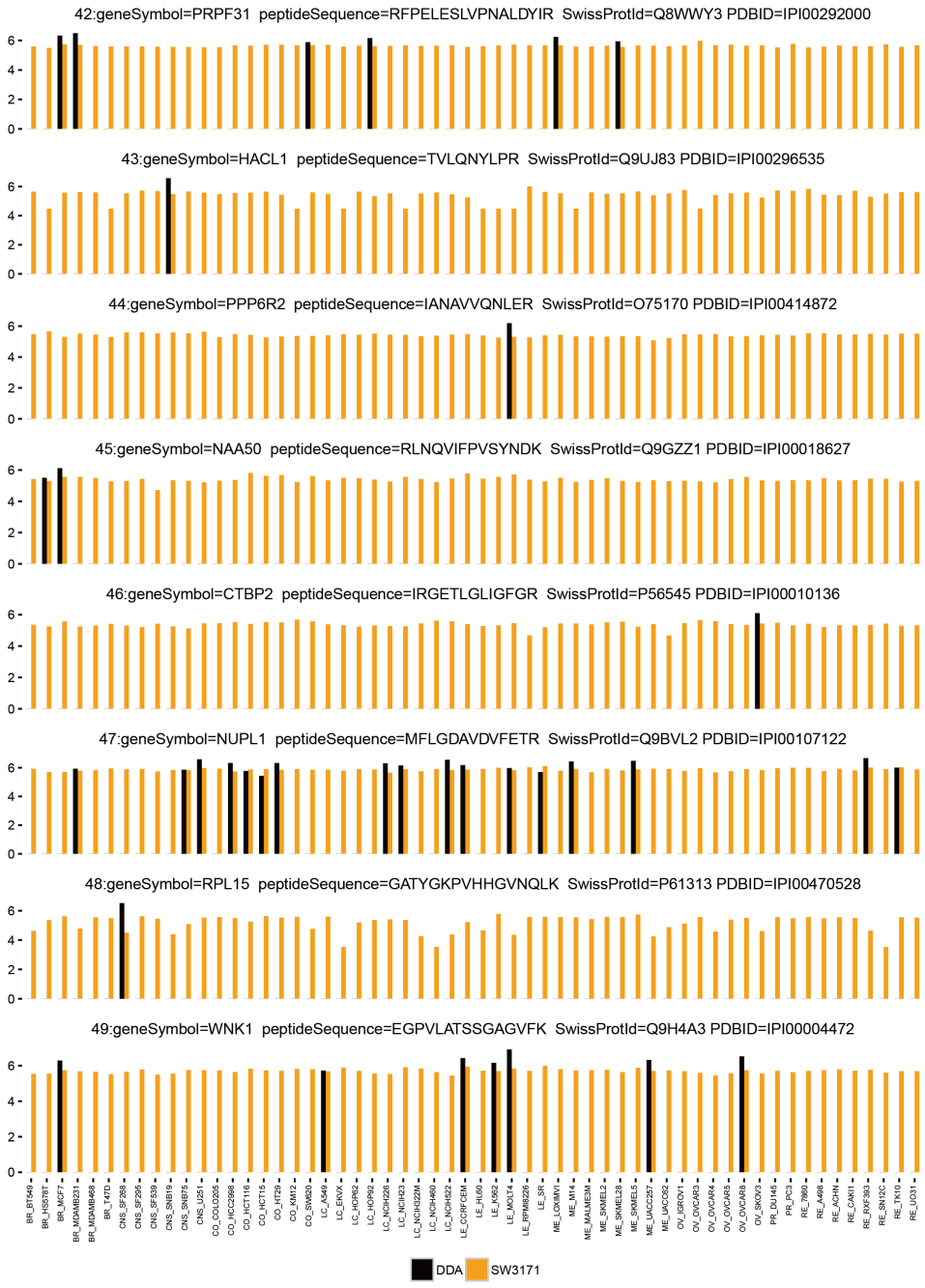
Supplementary Figure 11, Related to Figure 2. Bar plots for P62875, O14735, P51991, P18621, Q969G3, P46778, P08621, and P50151, which are quantified cross all NCI-60 cell lines by SWATH but not quantified by DDA.



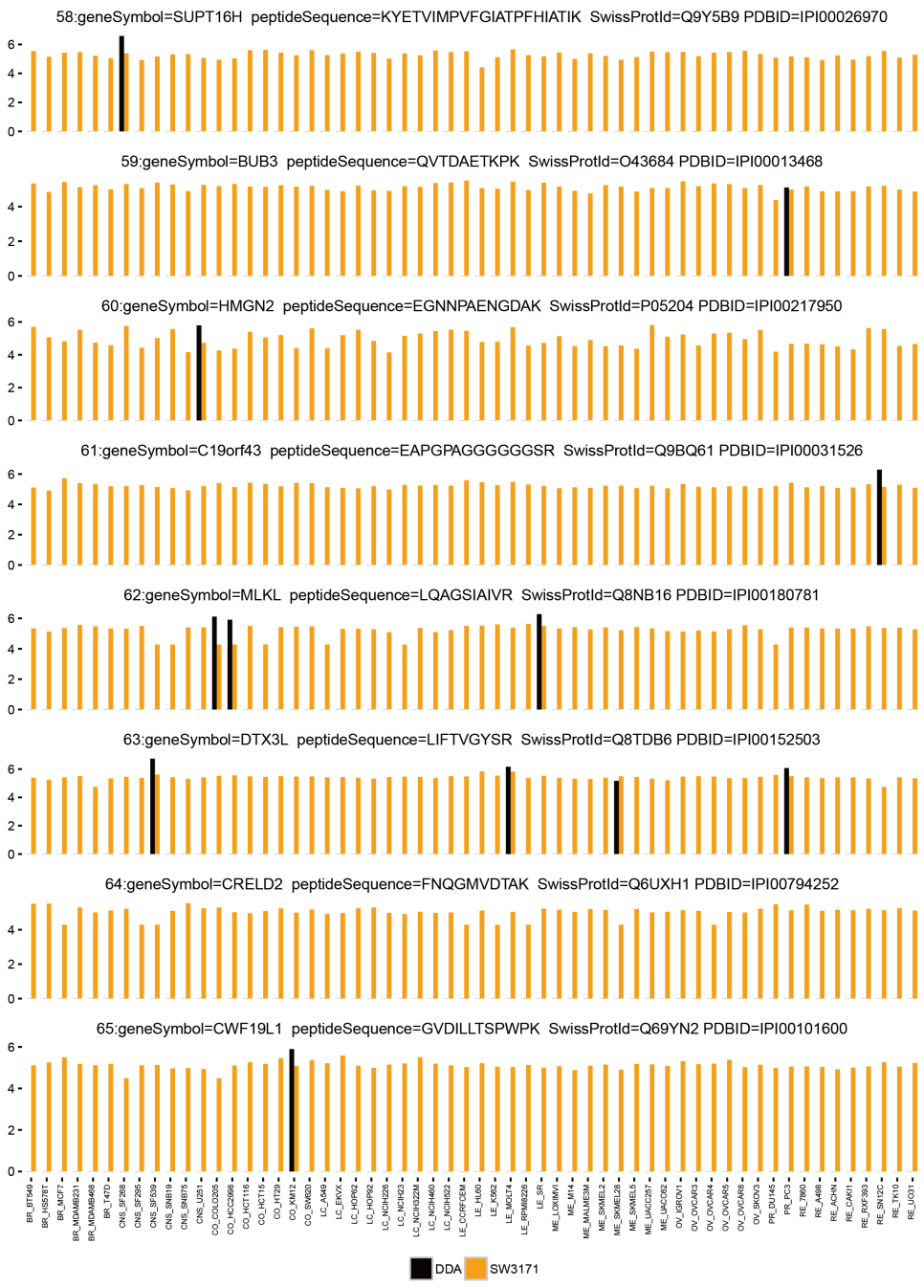
Supplementary Figure 12, Related to Figure 2. Bar plots for Q5T4S7, P22626, P10809, Q13541, P62753, Q9BUQ8, Q13619 and Q8N392, which are quantified cross all NCI-60 cell lines by SWATH but not quantified by DDA.



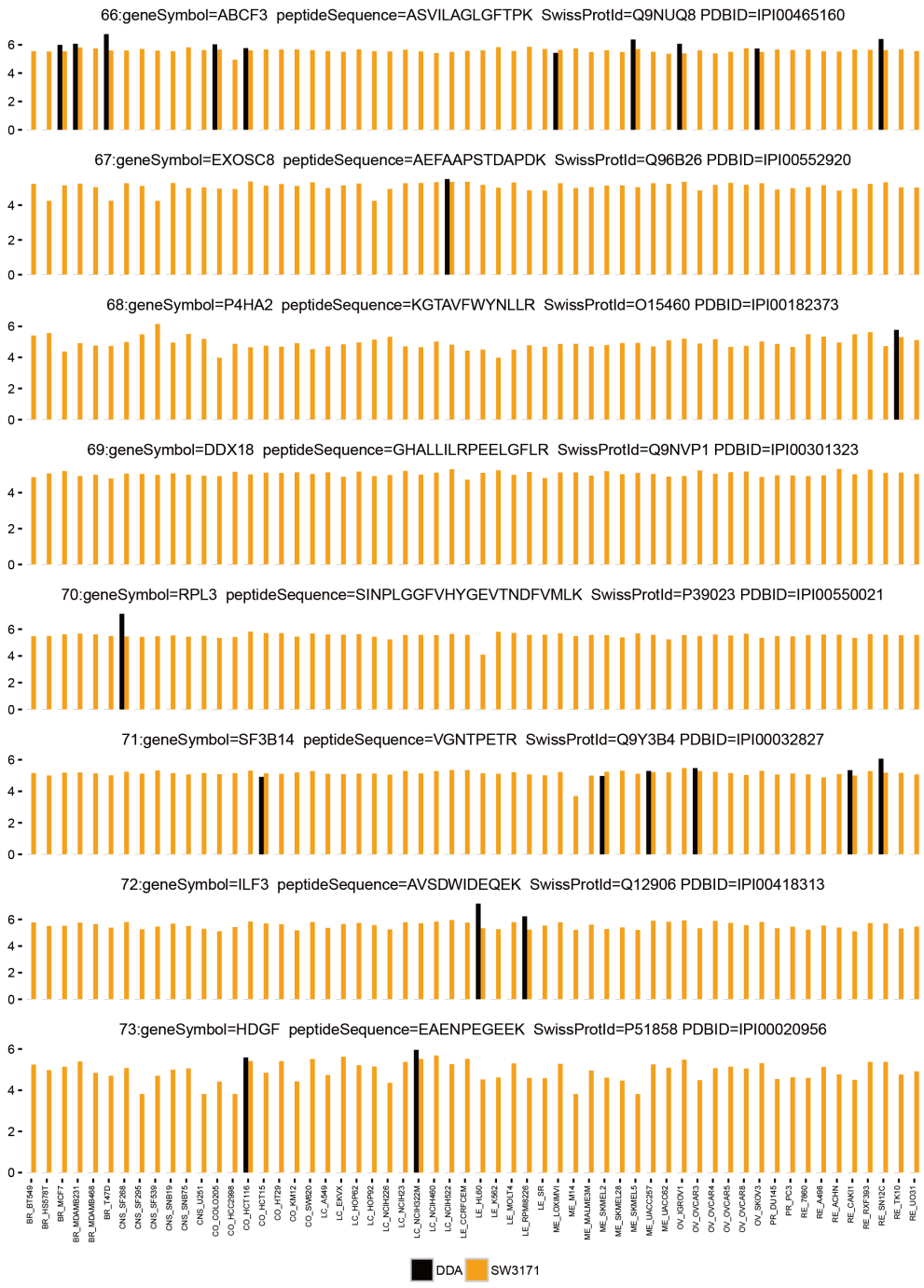
Supplementary Figure 13, Related to Figure 2. Bar plots for Q15154, Q8NC51, Q9NYF8, Q99816, Q96FS4, P46783, Q16543, O14964, which are quantified cross all NCI-60 cell lines by SWATH but not quantified by DDA.



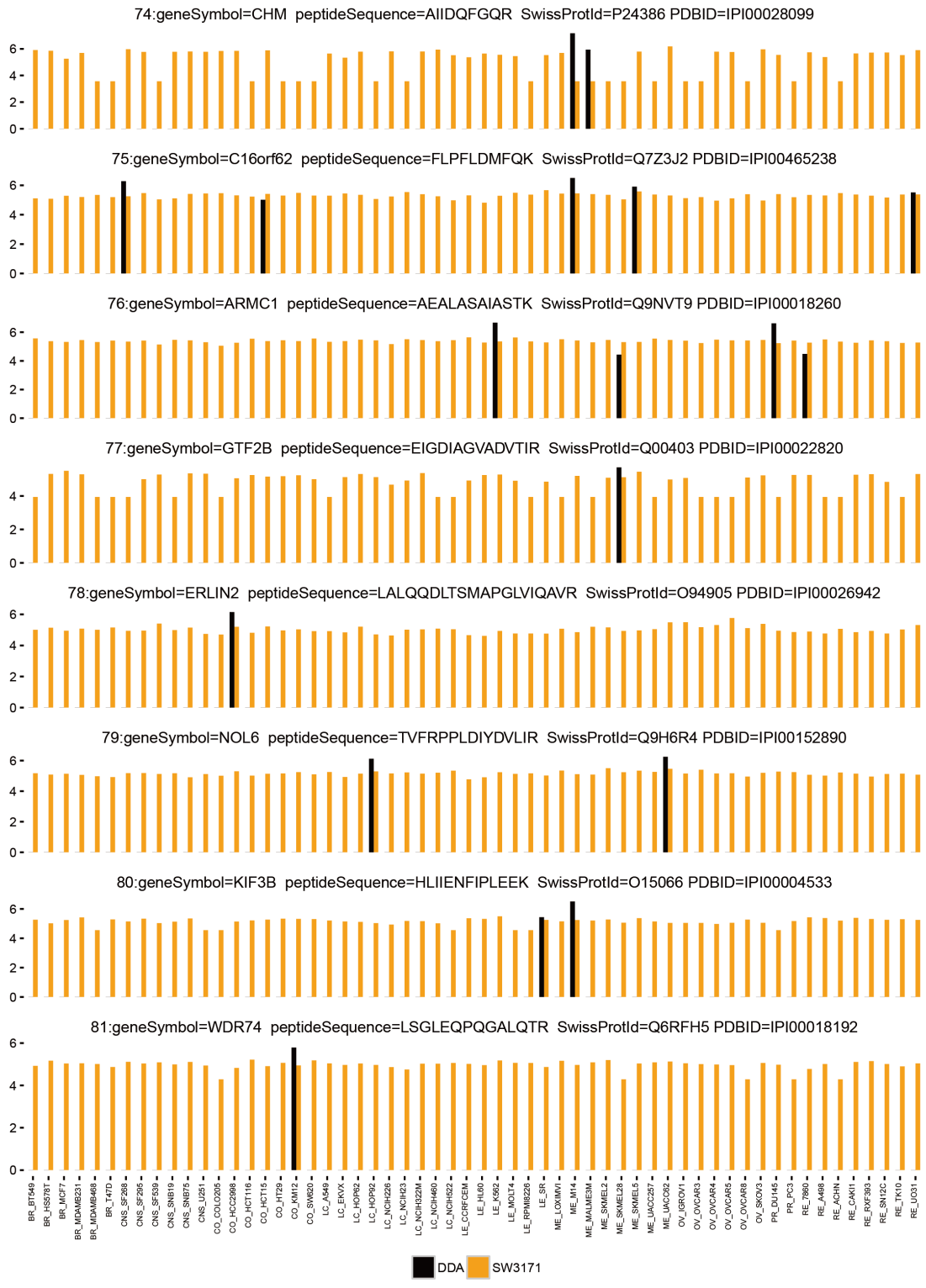
Supplementary Figure 14, Related to Figure 2. Bar plots for Q8WWY3, Q9UJB3, O75170, Q9GZZ1, P56545, Q9BVL2, P61313 and Q9H4A3, which are quantified cross all NCI-60 cell lines by SWATH but not quantified by DDA.



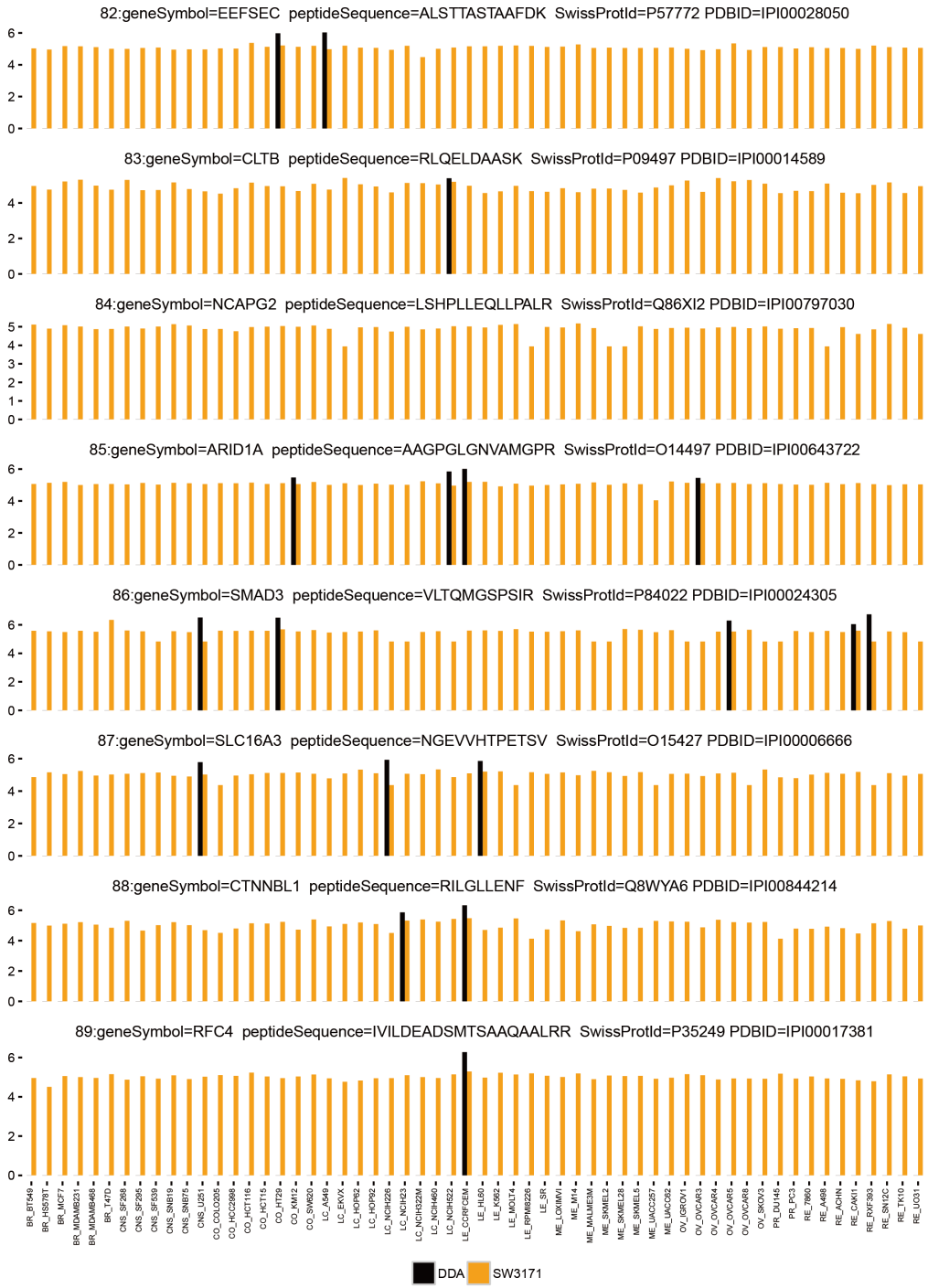
Supplementary Figure 15, Related to Figure 2. Bar plots for Q9Y5B9, Q436B4, P05204, Q9BQ61, Q8NB16, Q8TDB6, Q6UXH1, and Q69YN2, which are quantified cross all NCI-60 cell lines by SWATH but not quantified by DDA.



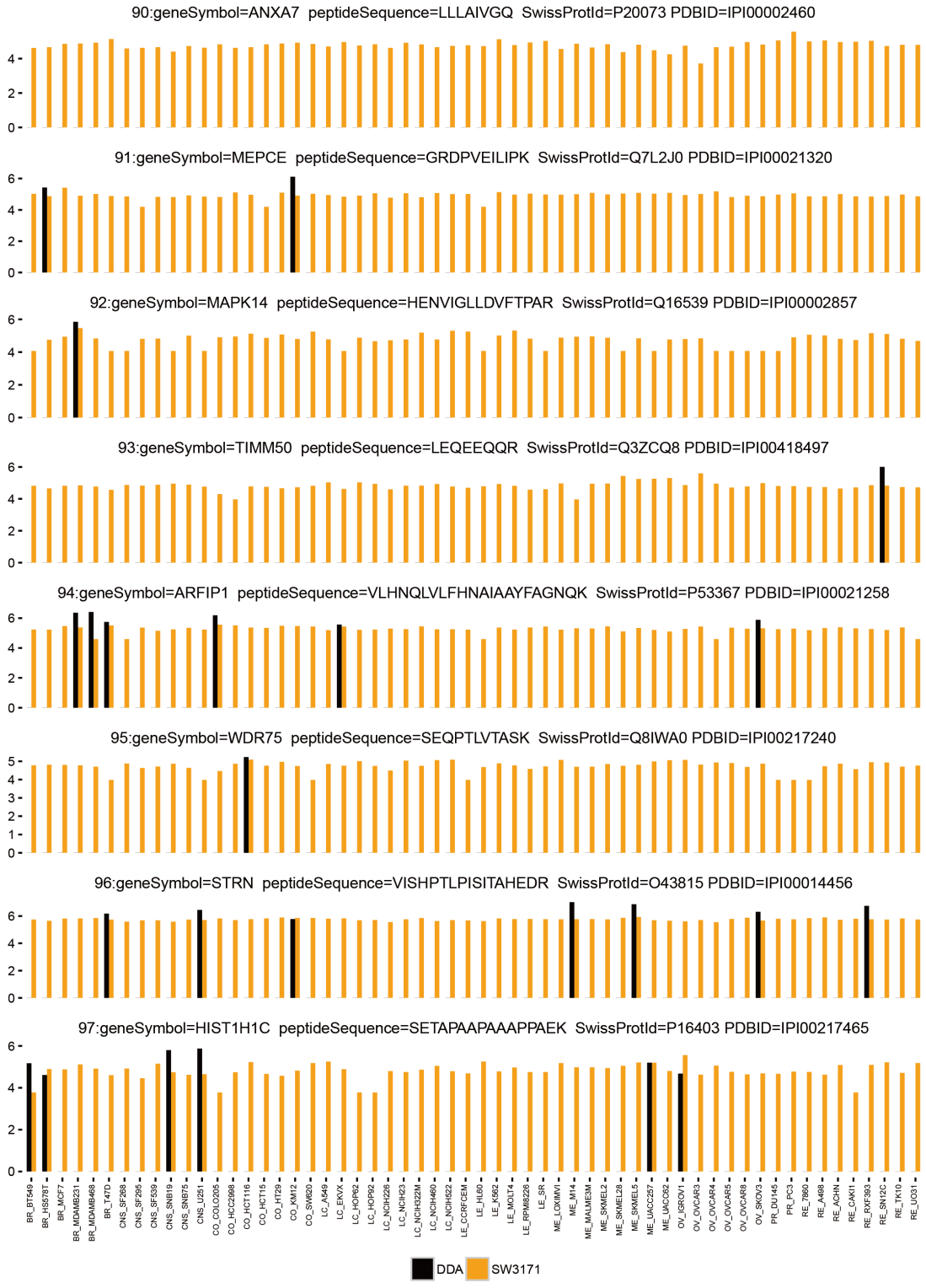
Supplementary Figure 16, Related to Figure 2. Bar plots for Q9NUQ8, Q96B26, O15460, Q9NVP1, P39D23, Q9Y3B4, Q12906, and P51858, which are quantified cross all NCI-60 cell lines by SWATH but not quantified by DDA.



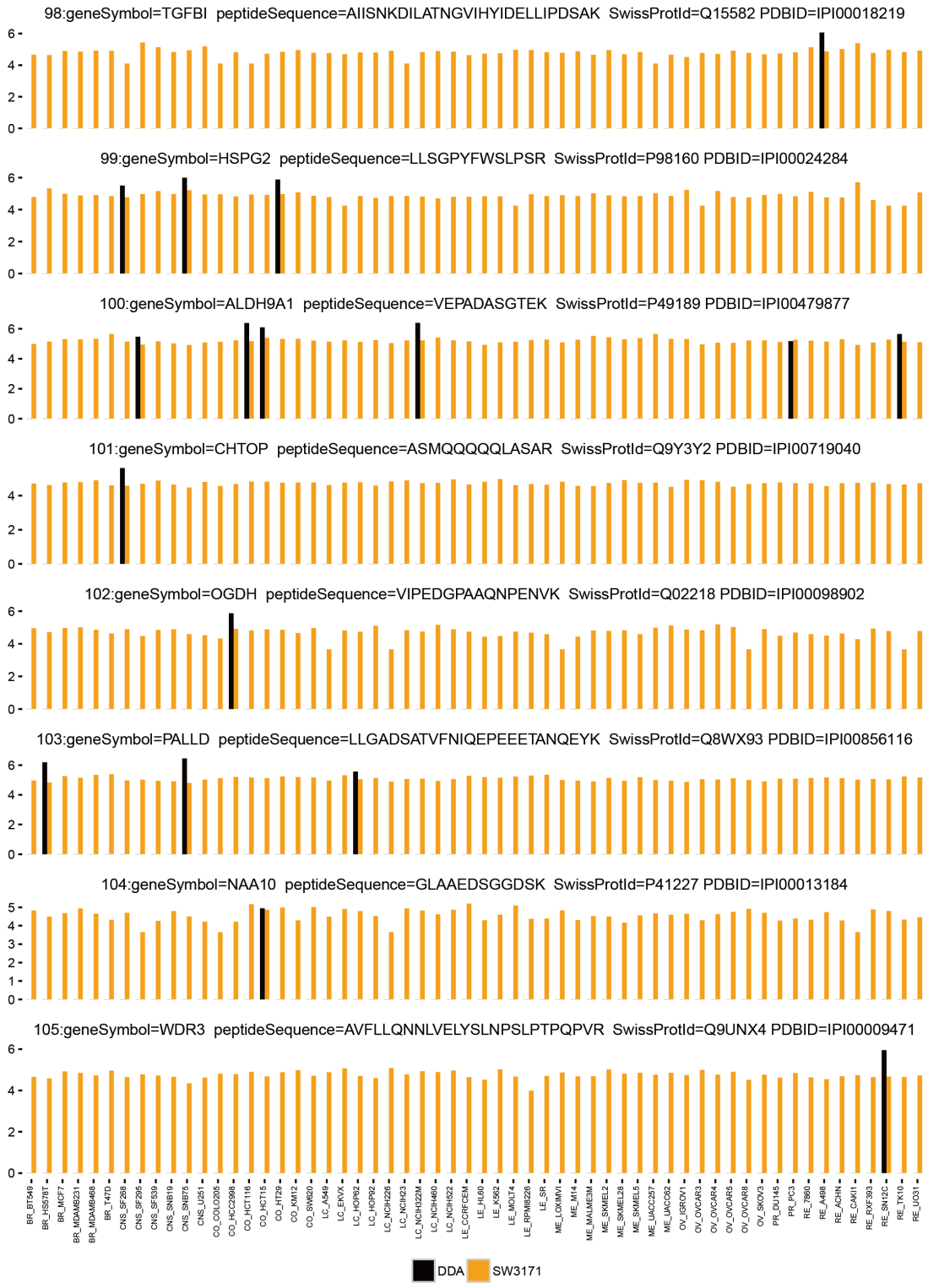
Supplementary Figure 17, Related to Figure 2. Bar plots for Bar plots for P24386, Q7Z3J2, Q9NVT9, Q00403, Q94905, Q9H6R4, O15066, and Q6RFH5, which are quantified cross all NCI-60 cell lines by SWATH but not quantified by DDA.



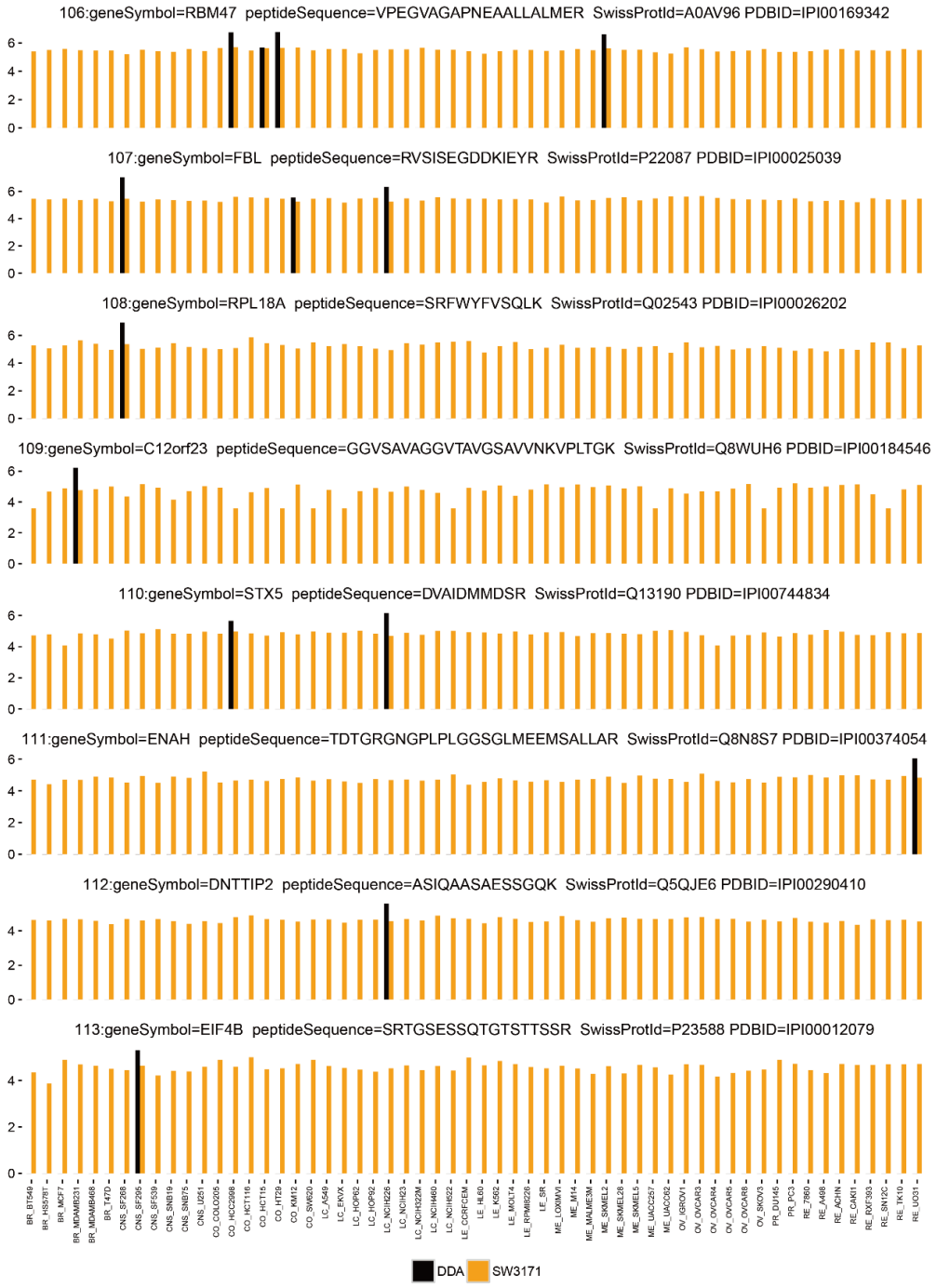
Supplementary Figure 18, Related to Figure 2. Bar plots for P57772, P09497, Q86X12, O14497, P84022, O15427, Q8WYA6, and P359249, which are quantified cross all NCI-60 cell lines by SWATH but not quantified by DDA.



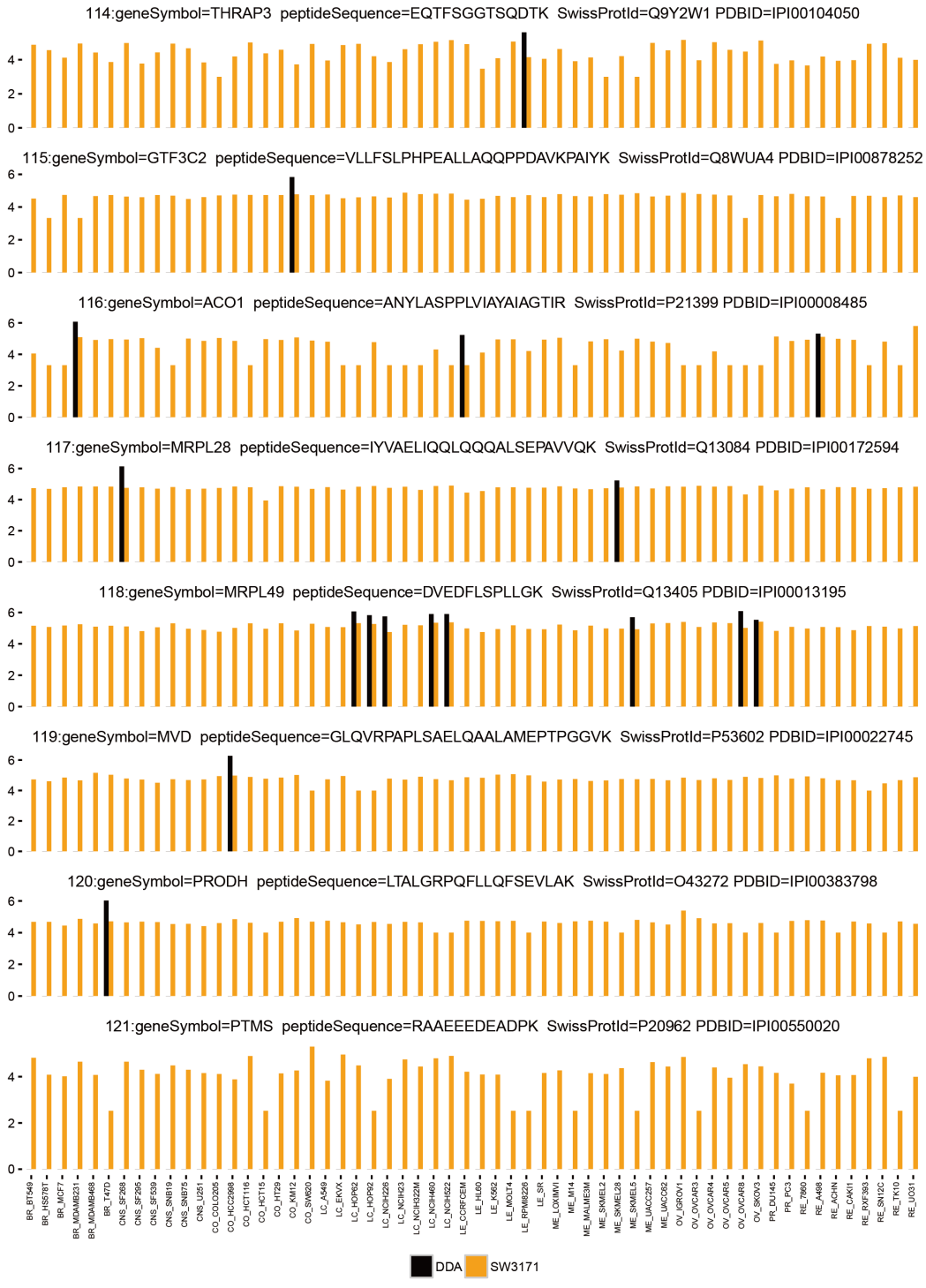
Supplementary Figure 19, Related to Figure 2. Bar plots for P20073, Q7L2J0, Q16539, Q3ZCQ8, P53367, Q8IWA0, 043815 and P16403, which are quantified cross all NCI-60 cell lines by SWATH but not quantified by DDA.



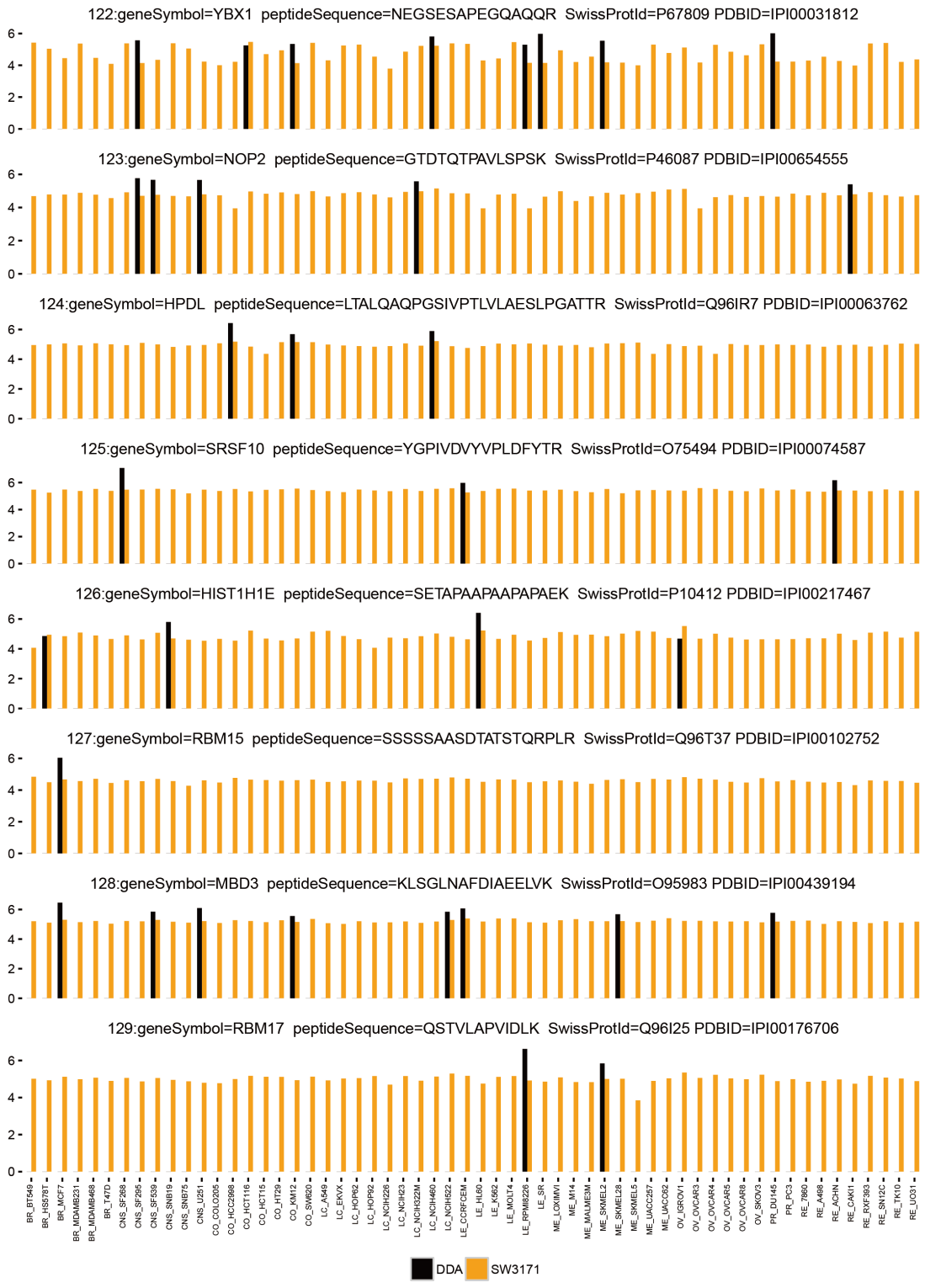
Supplementary Figure 20, Related to Figure 2. Bar plots for Q15582, P98160, P49189, Q9Y3Y2, Q02218, Q8WX93, P41227, and Q9UNX4, which are quantified cross all NCI-60 cell lines by SWATH but not quantified by DDA.



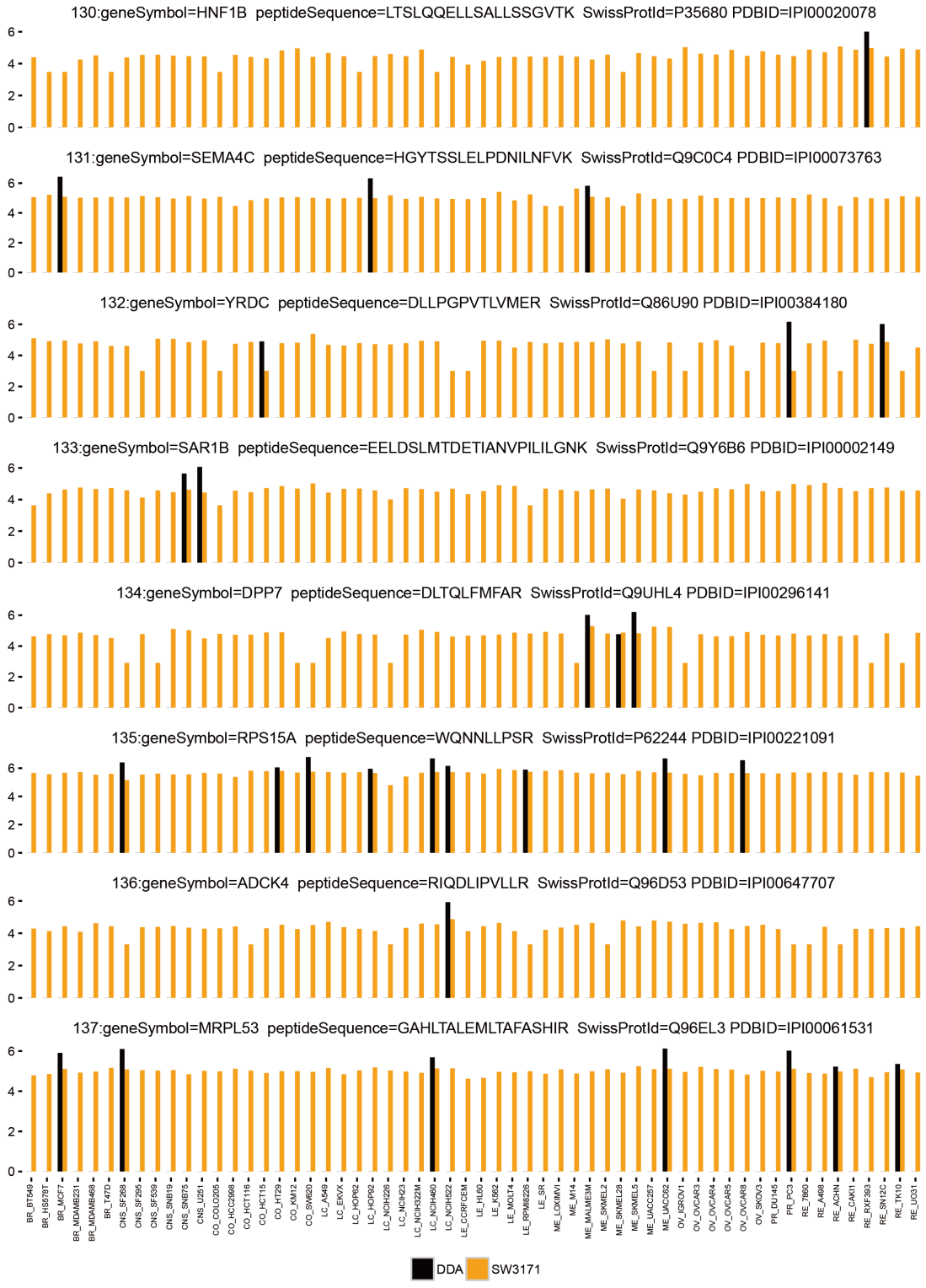
Supplementary Figure 21, Related to Figure 2. Bar plots for A0AV96, P22087, Q02543, Q8WUH6, Q13190, Q8N8S7, Q5QJE6 and P23588, which are quantified cross all NCI-60 cell lines by SWATH but not quantified by DDA.



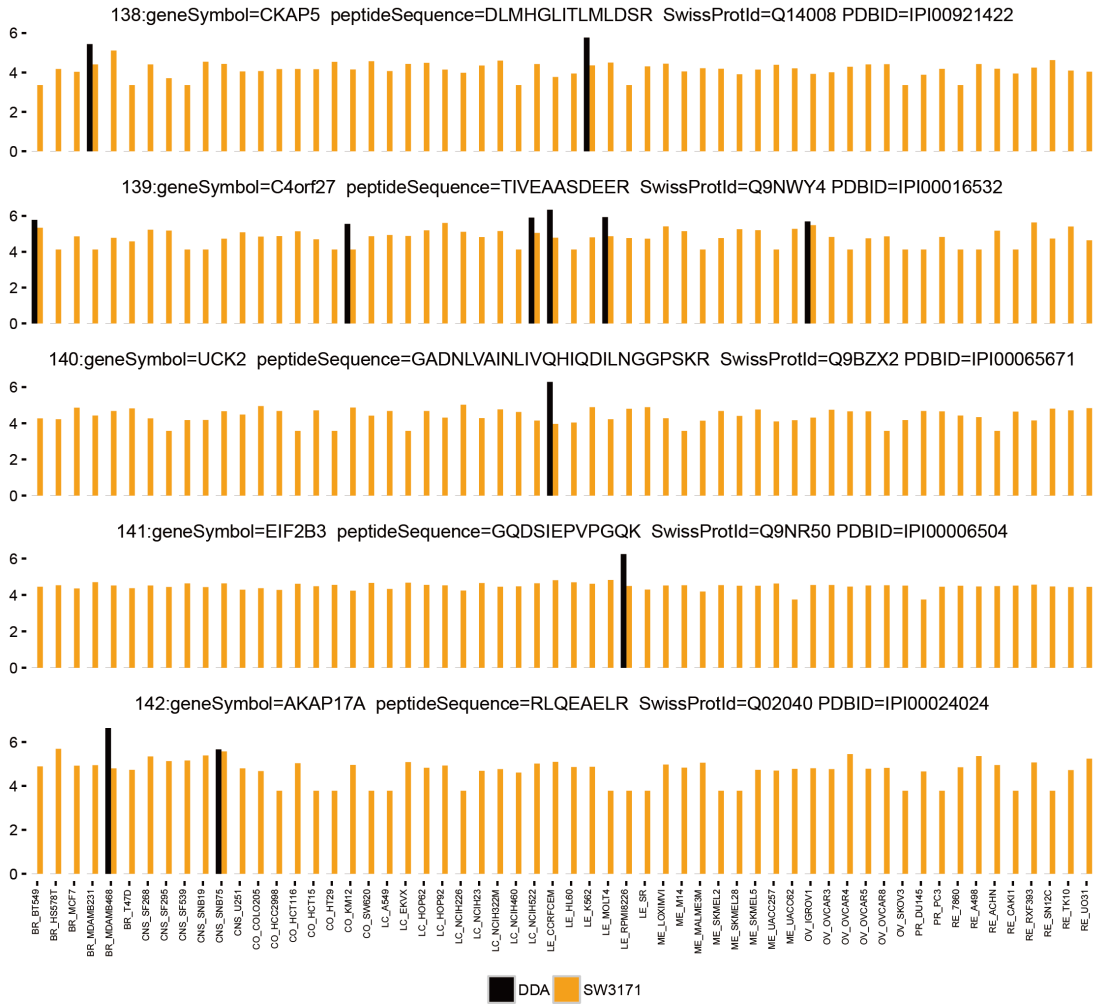
Supplementary Figure 22, Related to Figure 2. Bar plots for Q9Y2W1, Q8WUA4, P21399, Q13084, P53602, O43272 and P20962, which are quantified cross all NCI-60 cell lines by SWATH but not quantified by DDA.



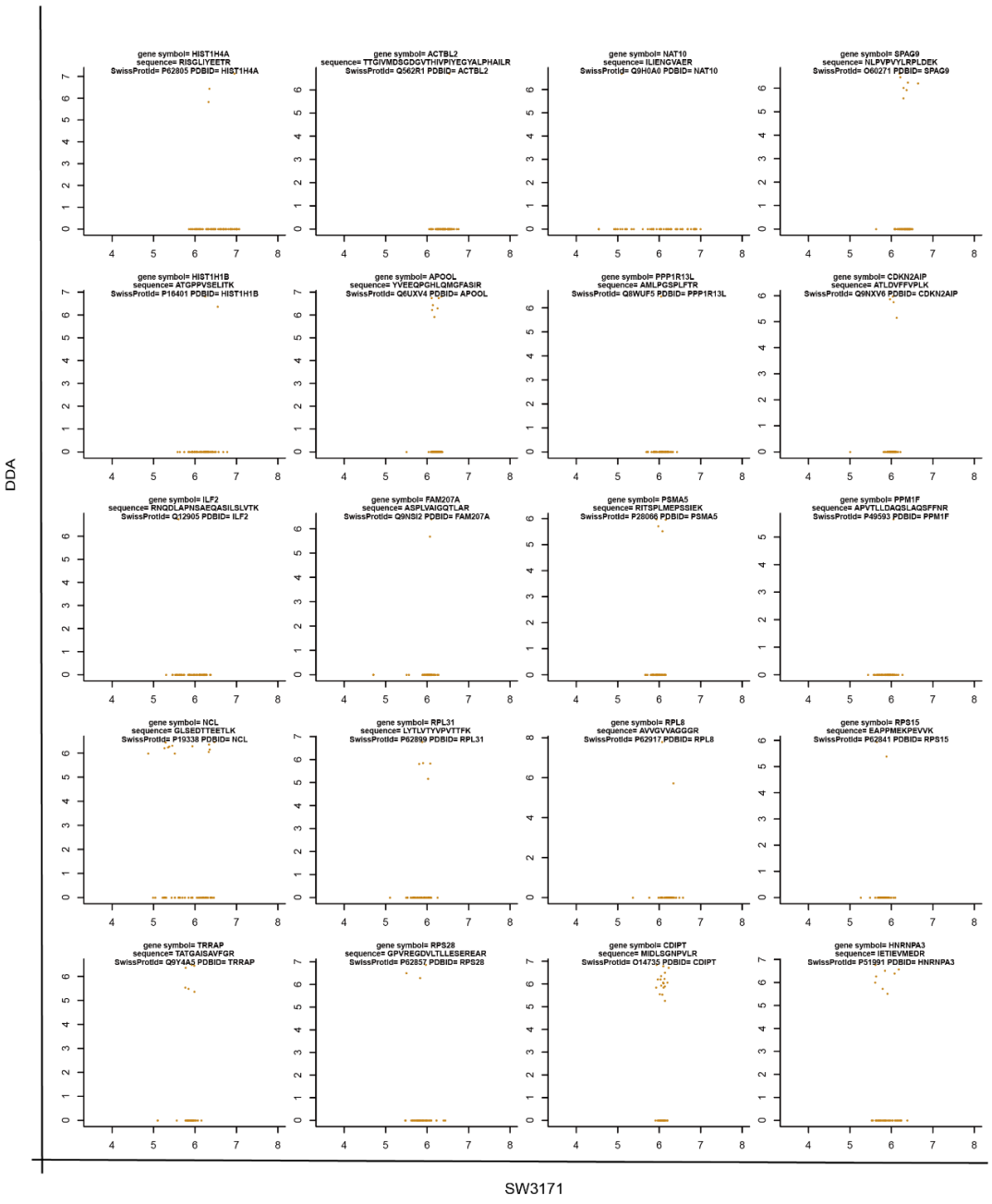
Supplementary Figure 23, Related to Figure 2. Bar plots for P67809, P46087, Q96IR7, O75494, P10412, Q96T37, O95983, and Q96I25, which are quantified cross all NCI-60 cell lines by SWATH but not quantified by DDA.



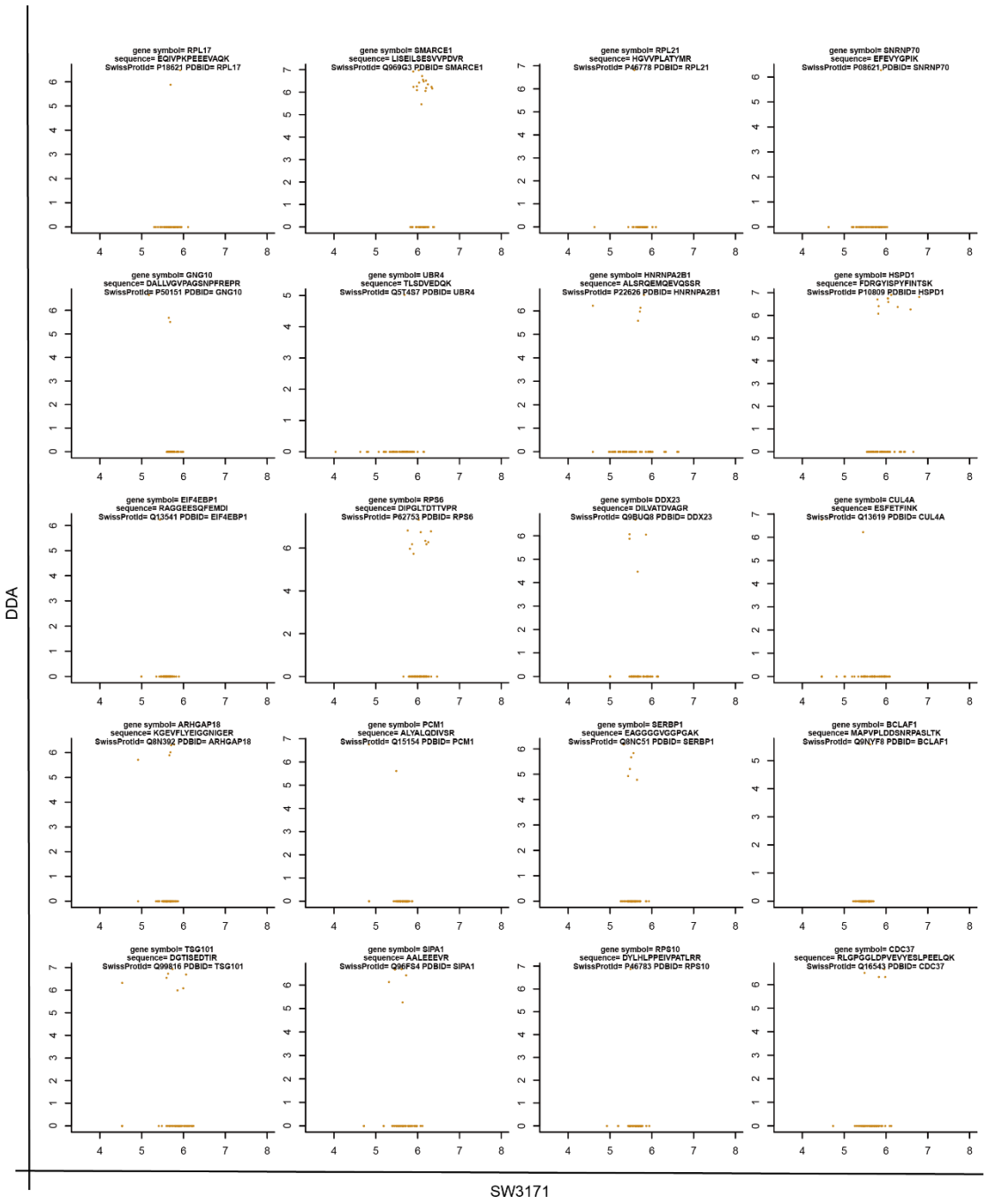
Supplementary Figure 24, Related to Figure 2. Bar plots for P35680, Q9C0C4, Q86U90, Q9Y6B6, Q9UHL4, P62244, Q96D53 and Q96EL3, which are quantified cross all NCI-60 cell lines by SWATH but not quantified by DDA.



Supplementary Figure 25, Related to Figure 2. Bar plots for Q14008, Q9NWY4, Q9BZX2, Q9NR50, and Q02040, which are quantified cross all NCI-60 cell lines by SWATH but not quantified by DDA.

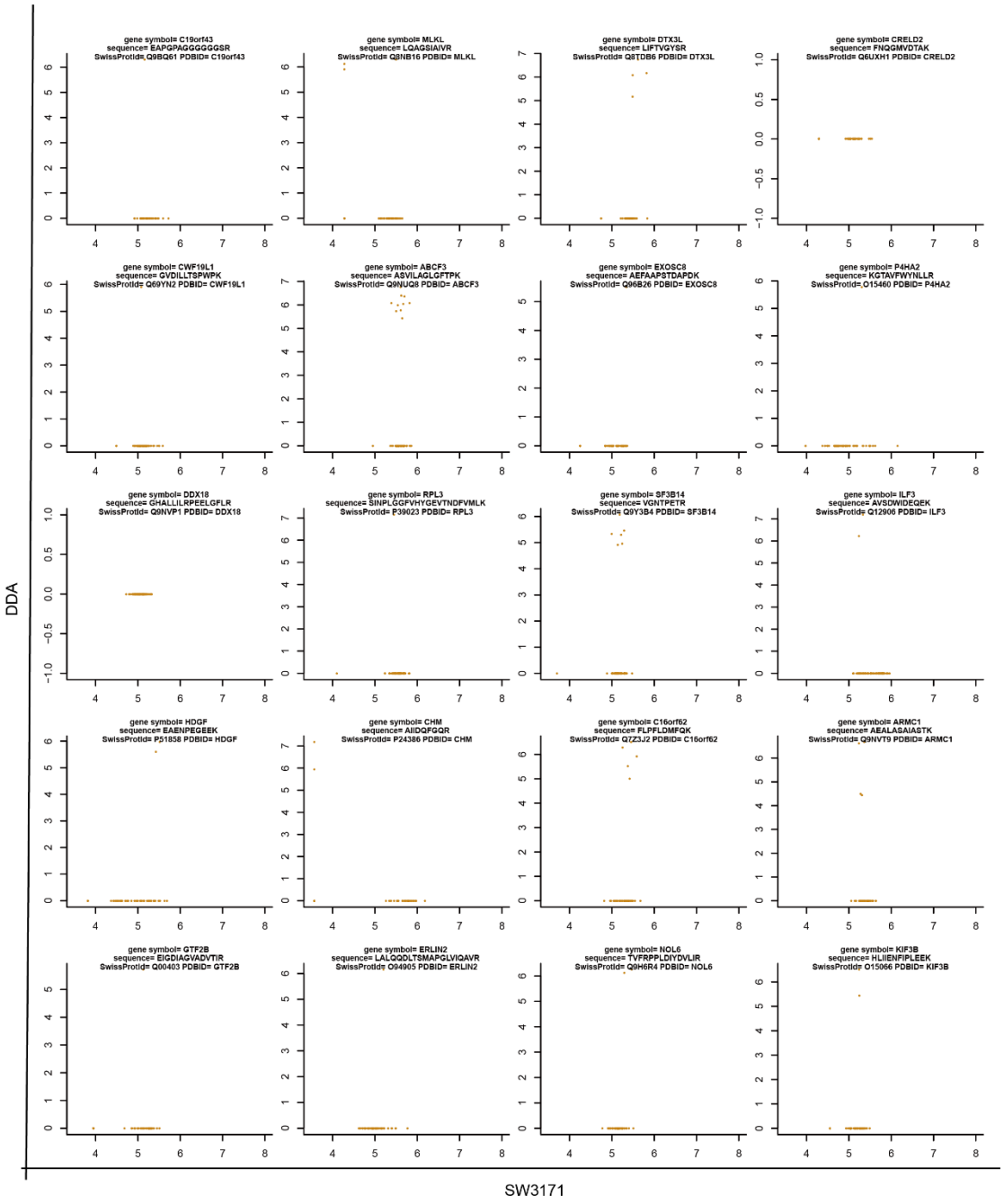


Supplementary Figure 26, Related to Figure 2. Scatter plots for P62805, Q562R1, Q9H0A0, O60271, P16401, Q6UXV4, Q8WUF5, Q9NXV6 , Q12905,Q9NSI2, P28066, P49593, P19338, P62899, P62917, P62841, , Q9Y4A5,P62875, O14735, and P51991 , which are all quantified cross all NCI-60 cell lines by SWATH but not quantified by DDA. This figure shows the data completeness difference of the two data sets.

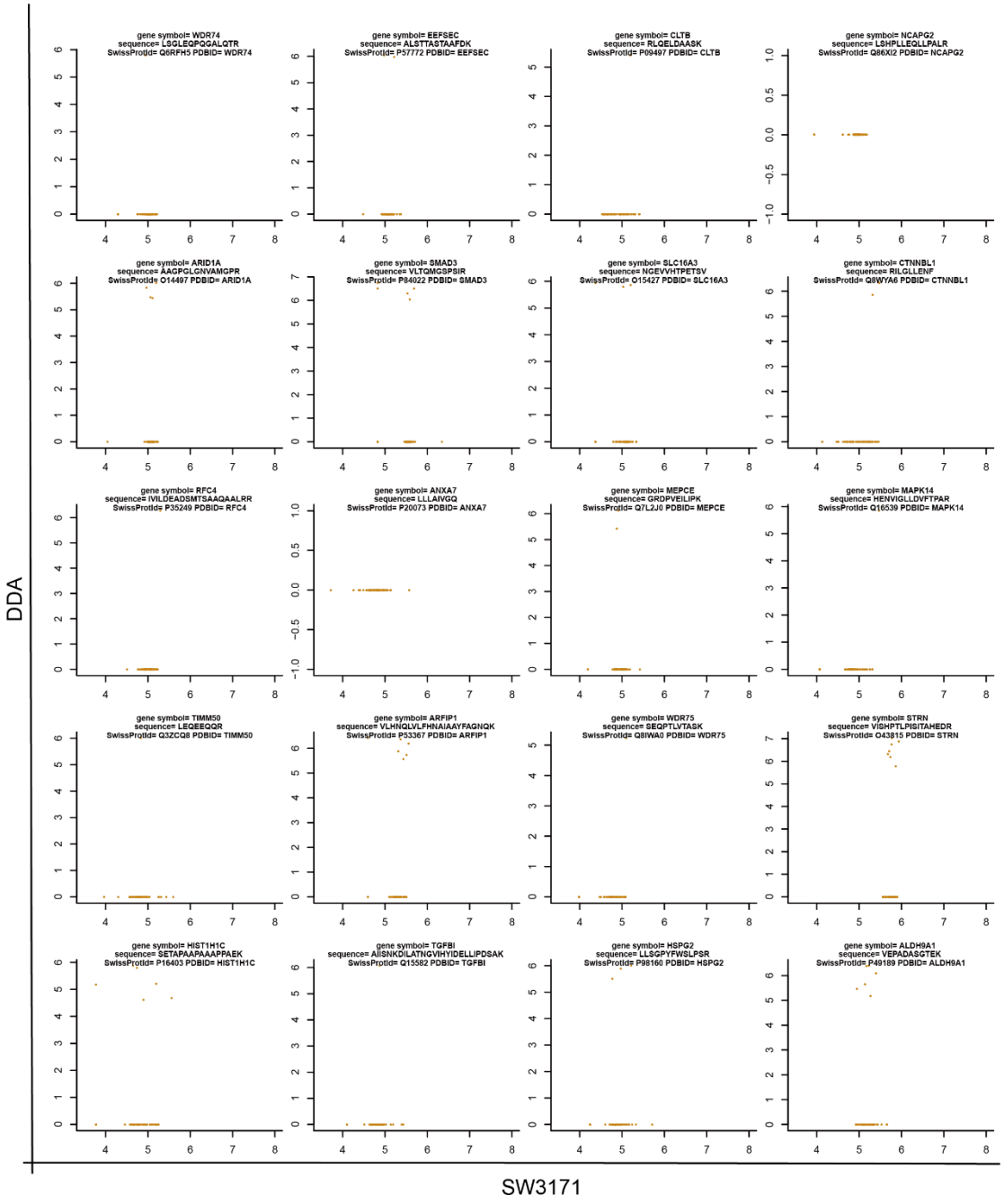


Supplementary Figure 27, Related to Figure 2. Scatter plots for P18621, Q969G3, P46778, P08621, P50151, Q5T4S7, P22626, P10809, Q13541, P62753, Q9BUQ8, Q13619 , Q8N392, Q15154, Q8NC51, Q9NYF8, Q99816, Q96FS4, P46783, and Q16543, which are all quantified cross all NCI-60 cell lines by SWATH but not quantified by DDA. This figure shows the data completeness difference of the two data sets.

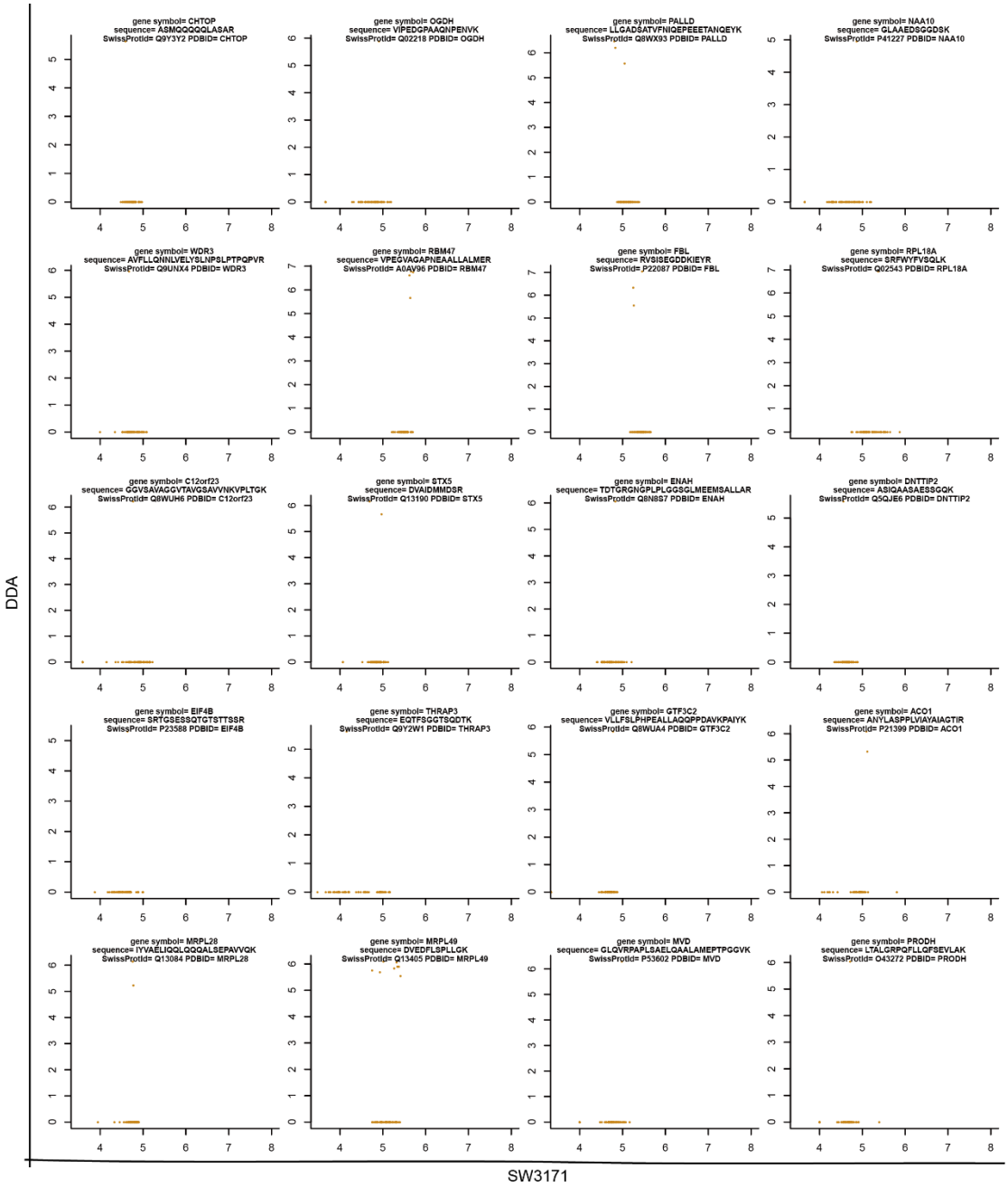
Supplementary Figure 28, Related to Figure 2. Scatter plots for O14964,Q8WWY3, Q9UJB3, O75170, Q9GZZ1, P56545, Q9BVL2, P61313, Q9H4A3,Q9NPE3, Q06023, Q92576, Q969S3, Q9H089, Q13435, Q8N5N7, Q9Y5B9, Q436B4, and P05204, which are all quantified cross all NCI-60 cell lines by SWATH but not quantified by DDA. This figure shows the data completeness difference of the two data sets.



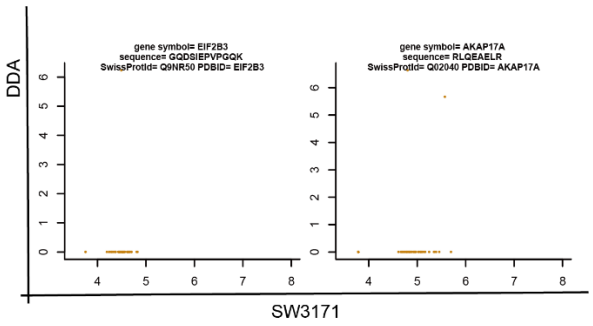
Supplementary Figure 29, Related to Figure 2. Scatter plots for Q9BQ61, Q8NB16, Q8TDB6, Q6UXH1, Q69YN2, Q9NUQ8, Q96B26, O15460, Q9NVP1, P39D23, Q9Y3B4, Q12906, , P51858, P24386, Q7Z3J2, Q9NVT9, Q00403, Q94905, Q9H6R4, and O15066, which are all quantified cross all NCI-60 cell lines by SWATH but not quantified by DDA. This figure shows the data completeness difference of the two data sets.



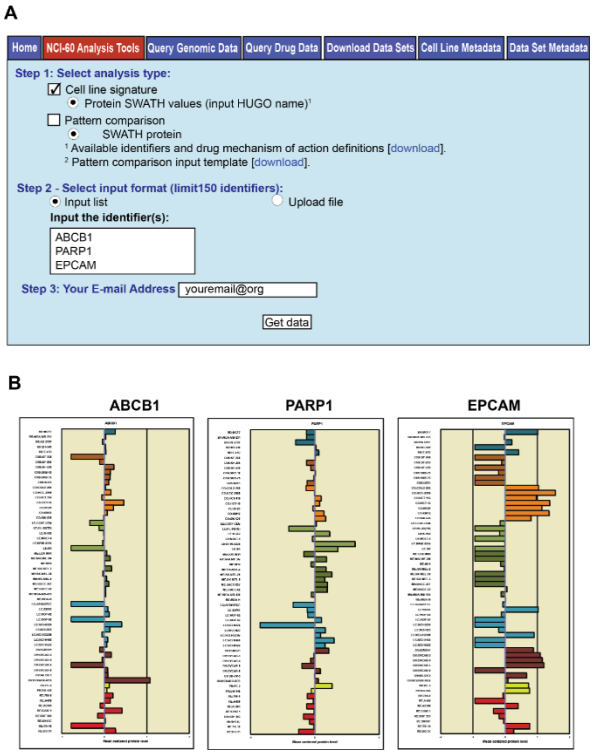
Supplementary Figure 30, Related to Figure 2. Scatter plots for Q6RFH5, P57772, P09497, Q86X12, O14497, P84022, O15427, Q8WYA6, P359249, P20073, Q7L2J0, Q16539, Q3ZCQ8, P53367, Q8IWA0, 043815 , P16403,Q15582, P98160, and P49189, which are all quantified cross all NCI-60 cell lines by SWATH but not quantified by DDA. This figure shows the data completeness difference of the two data sets.



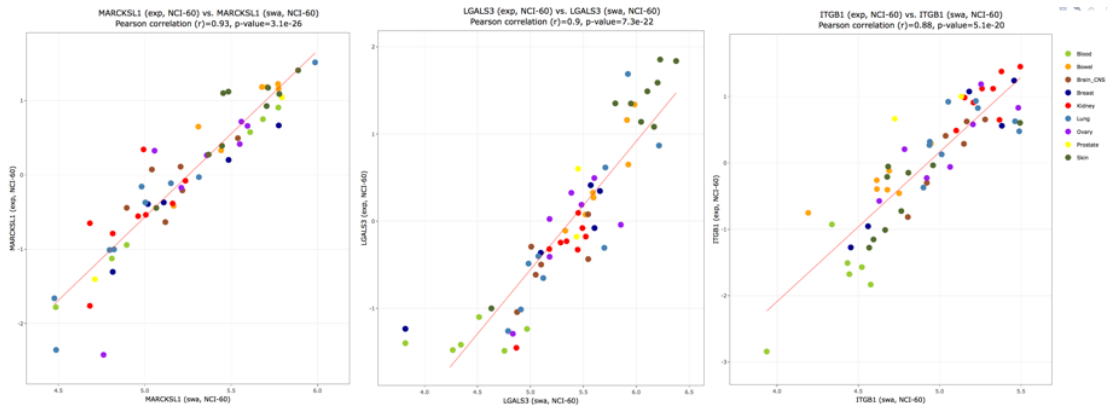
Supplementary Figure 31, Related to Figure 2. Scatter plots for Q9Y3Y2, Q02218, Q8WX93, P41227, Q9UNX4, A0AV96, P22087, Q02543, Q8WUH6, Q13190, Q8N8S7, Q5QJE6 , P23588,Q9Y2W1, Q8WUA4, P21399,Q13084, P53602, and O43272, which are all quantified cross all NCI-60 cell lines by SWATH but not quantified by DDA. This figure shows the data completeness difference of the two data sets.



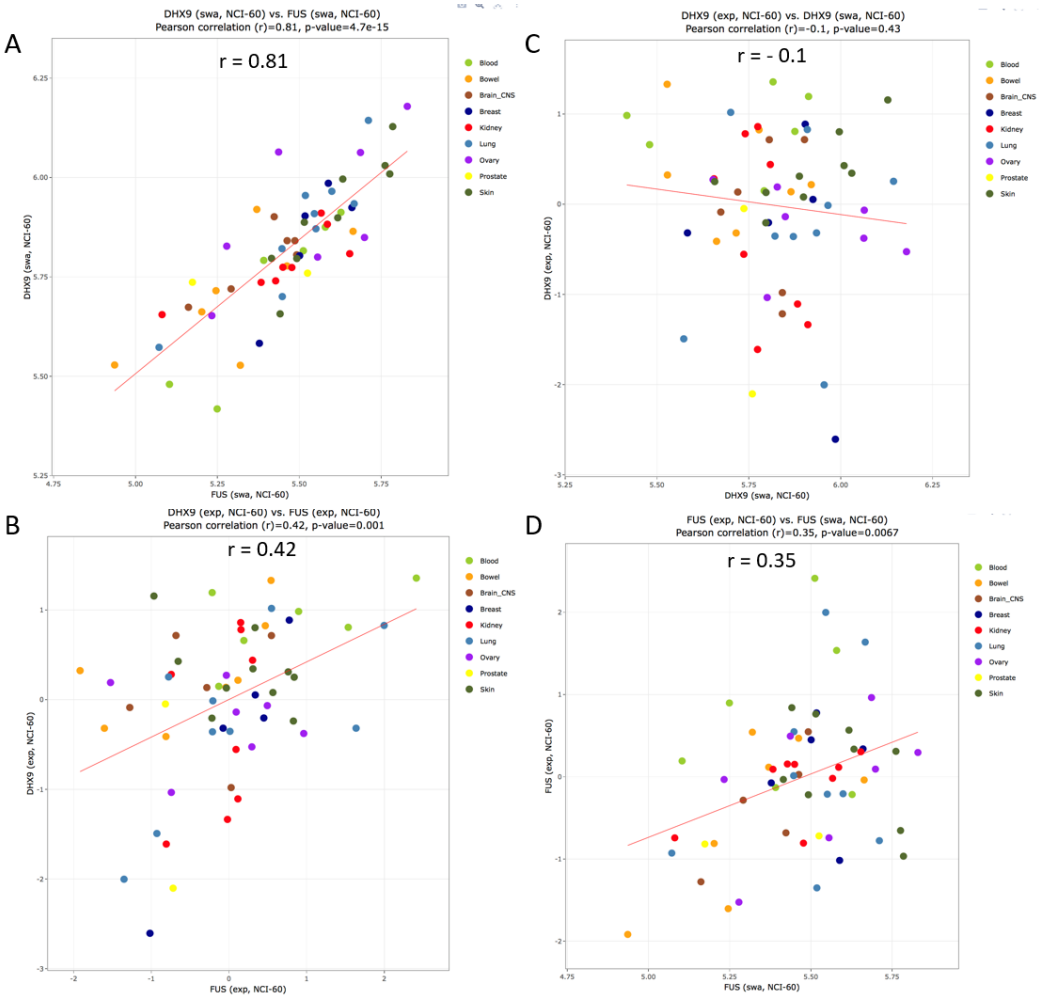
Supplementary Figure 32, Related to Figure 2. Scatter plots for P20962,P67809, P46087, Q96IR7, O75494, P10412, Q96T37, O95983, Q96I25,P35680, Q9C0C4, Q86U90, Q9Y6B6, Q9UHL4, P62244, Q96D53, Q96EL3, Q14008, Q9NWY4, Q9BZX2, Q9NR50, and Q02040, which are all quantified cross all NCI-60 cell lines by SWATH but not quantified by DDA. This figure shows the data completeness difference of the two data sets.



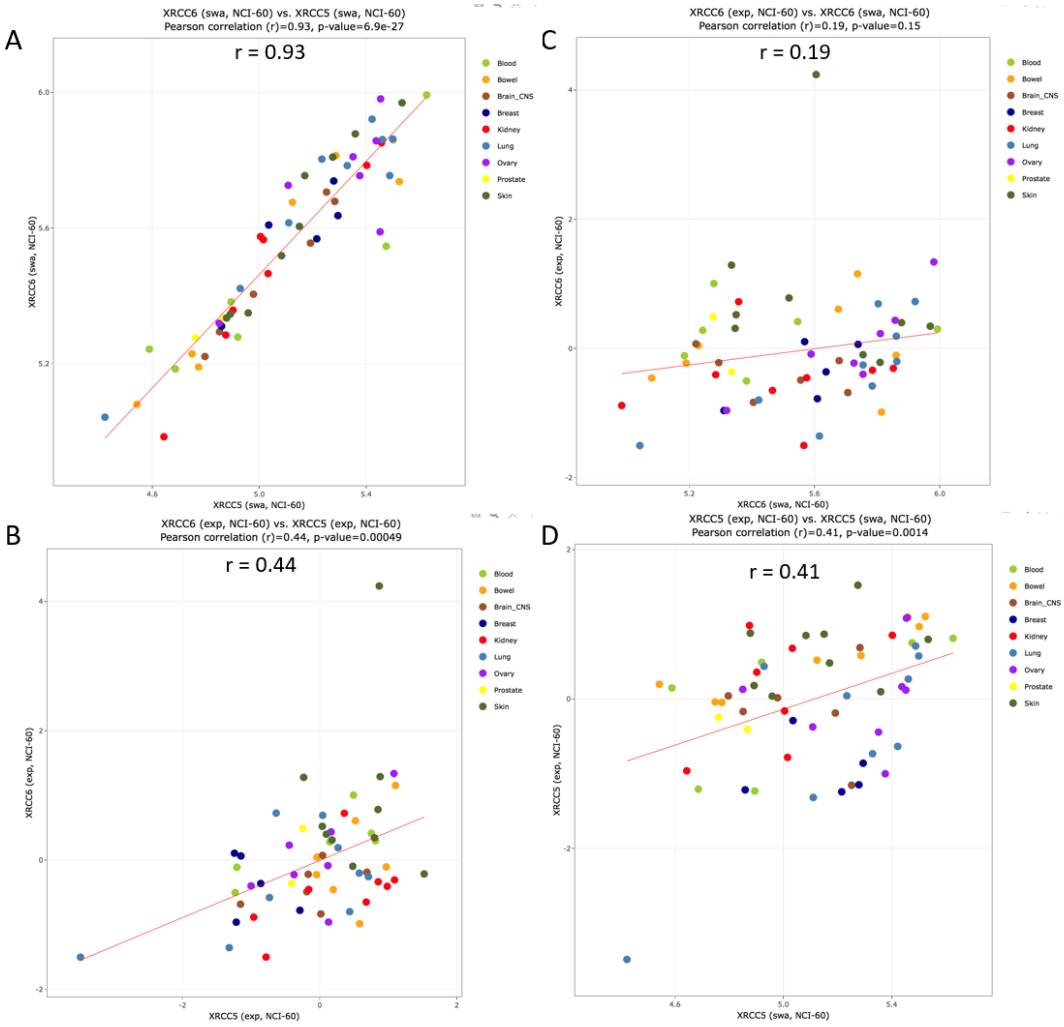
Supplementary Figure 33, Related to Figure 2. Access to NCI-60 proteotype in Cellminer. To facilitate data access, visualization, and comparison with other forms of genomic and pharmacological data for the NCI-60 cancer cell lines, we have incorporated the SWATH data within CellMiner 4 5. The CellMiner web site allows the data to be retrieved or used in several ways 6. (**A**) The "Download Data Sets" tab allows either the total 3,171 proteins, or the 22,554 peptides data sets to be downloaded. This data will primarily be of use in computational biology pipelines. The "Query Genomic Data" tab allows up to 150 proteins or peptides to be accessed (using the "Gene" or "Peptide" pull downs), queryable by gene name or peptide peak identifier, chromosomal or genomic location. Data is sent in both Excel (.xls) and text (.txt) format. The "NCI-60 Analysis Tools" tab (**A**) provides "Cell line signatures". To obtain "Cell line signatures" for genes, select "Cell line signature" in Step 1, and then "Protein SWATH values". In Step 2, up to 150 genes of interest may be input by either typing in the gene names in the "Input the identifier" box, or uploading them as a text or Excel file using the "Upload file" radio button. In Step 3, enter your e-mail address, and click "Get data". Results will be sent by e-mail for each gene, with a link to download the results. This file contains three worksheets: i) tabular mean centered protein levels ratios as a both a bar plot and tabular data, and the peptide peak information for that gene ii) "Bin protein levels" with a histogram of the protein levels and iii) and "Footnotes". (**B**) provides examples of three genes of interest. These "Cell line signatures" can also be used as input for the Pattern Comparison tool (also within the NCI-60 Analysis tools section) which provides correlated molecular and compound activity data. All available gene and peptide identifiers are available as a list within the "Available identifiers and drug mechanism of action definitions" as a download within the "NCI-60 Analysis Tools" tab.



Supplementary Figure 34, Related to Figure 2. CellminerCDB snapshot views of three genes with highest correlation between expression in SWATH and transcriptome data: Myristyolated Alanine-Rich C Kinase (MARCKSL1), Galectin 3 (LGALS3) and Integrin β1 (ITGB1) (see Supplementary Table 3). Data are snapshots from https://discover.nci.nih.gov/CellMinerCDB.

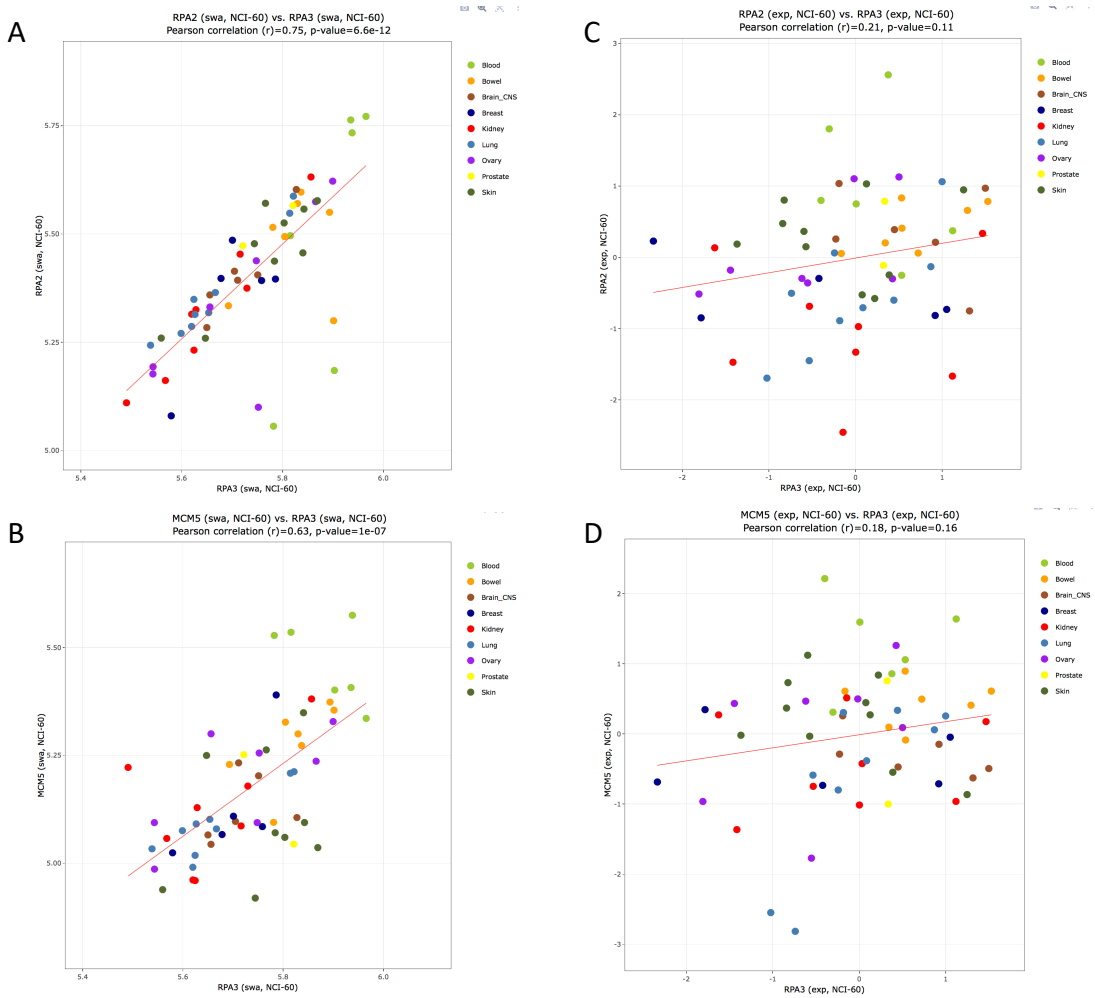


Supplementary Figure 35, Related to Figure 2 and 4. CellminerCDB view of the stoichiometric relationship of the expression of two RNA binding proteins DHX9 (RNAse A) and FUS (Fused in Sarcoma) in SWATH and transcriptome data. https://discover.nci.nih.gov/CellMinerCDB snapshots showing: **A.** The high stoichiometric correlation for both DHX9 and FUS across the NCI-60. **B.** The lower stoichiometric relationship between DHX9 and FUS transcripts. **C.** The lack of correlation between DHX9 protein and transcripts across the NCI-60. **D.** The low stoichiometric relationship between FUS protein and transcripts across the NCI-60.



Supplementary Figure 36, Related to Figure 2. CellminerCDB view of XRCC5 and XRCC6 expression in SWATH and transcriptome data.

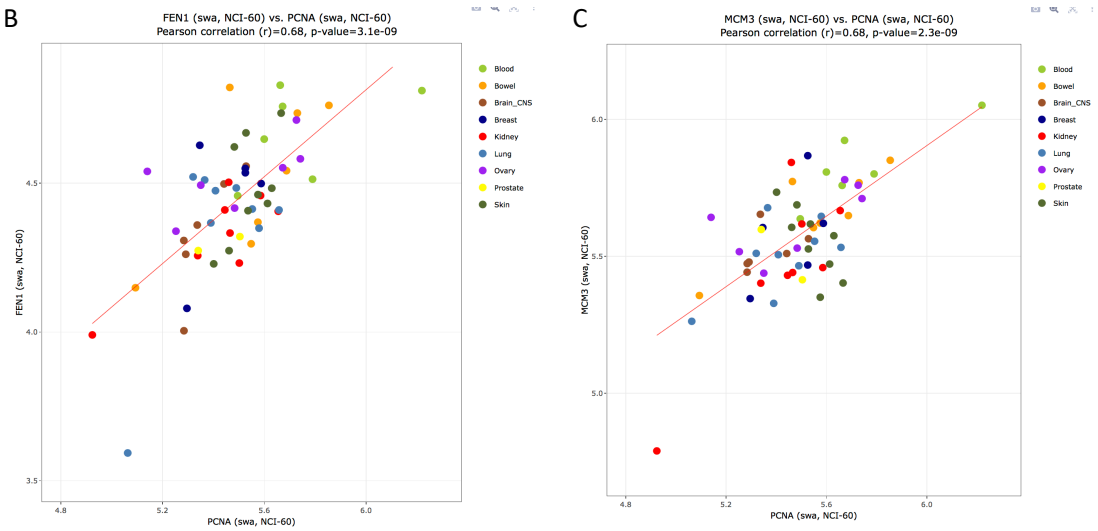
https://discover.nci.nih.gov/CellMinerCDB snapshots showing: **A.** The high stoichiometric correlation for both Ku subunits XRCC6 (KU70) and XRCC5 (KU80) across the NCI-60. **B.** The lower stoichiometric relationship between XRCC6 and XRCC5 transcripts. **C.** The lack of correlation between XRCC6 protein and transcripts across the NCI-60. **D.** The lower stoichiometric relationship between XRCC5 protein and transcripts across the NCI-60.



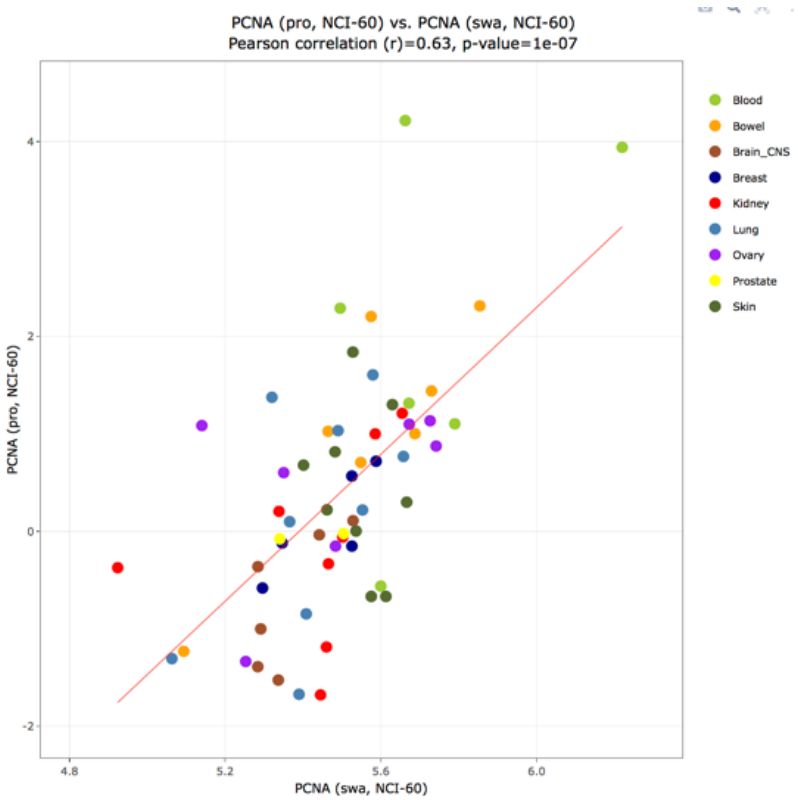
Supplementary Figure 37, Related to Figure 2. CellminerCDB snapshot views showing co-expression of replication proteins: RPA3 with RPA2 (A) and MCM5 (C) whereas transcripts do not show significant correlations (C-D). Data are snapshots from https://discover.nci.nih.gov/CellMinerCDB.

A

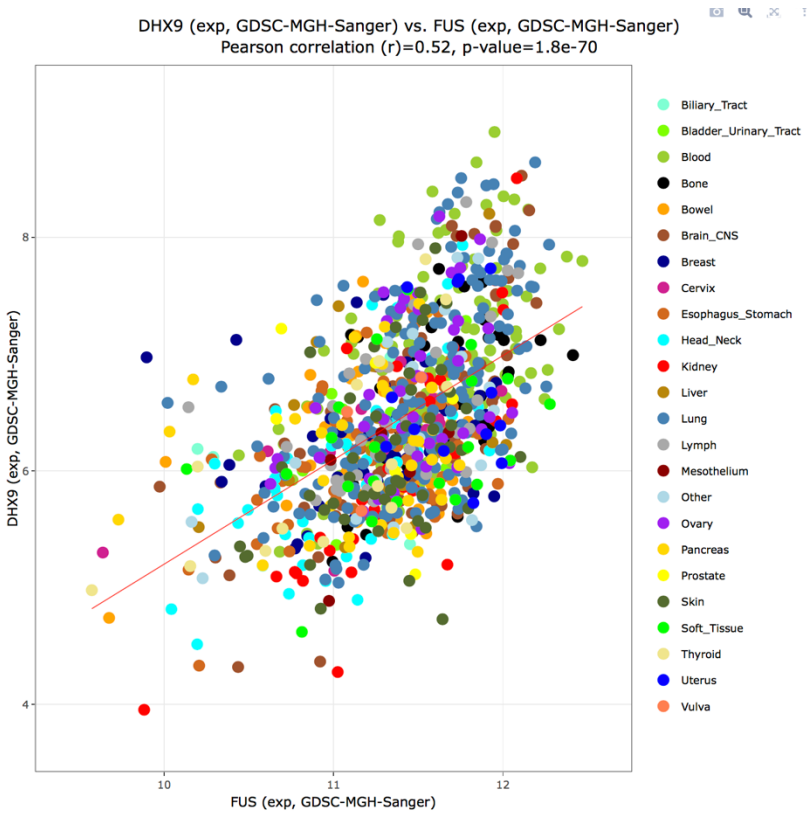
SWA	All	All	All	All	All	All
swa	PCNA	20pter-p12	1	0	0	DNA Damage Response (DDR); DDR (BER); DDR (DNA replication)
swa	MCM3	6p12	0.684	2.3e-9	0.0000904	DDR (DNA replication)
swa	FEN1	11q12	0.68	3.15e-9	0.0000904	DNA Damage Response (DDR); DDR (BER); DDR (DNA replication)
swa	MTHFD1	14q24	0.661	1.19e-8	0.00025	water-soluble vitamin metabolic process;cellular amino acid biosynthetic process
swa	ATIC	2q35	0.65	2.5e-8	0.000448	purine ribonucleoside monophosphate biosynthetic process;organ regeneration



Supplementary Figure 38, Related to Figure 2. CellminerCDB snapshots showing co-expression of replication proteins determined by SWATH and detailing the coexpression of FEN1 with PCNA and of PCNA with the replication helicase protein MCM3. A. The "Compare pattern" tool was used with PCNA as the "x-axis entry". Snapshots from https://discover.nci.nih.gov/CellMinerCDB showing only the top correlates with RPA3 including MCM3 and FEN1. **B.** Stoichiometric relationship between PCNA and FEN1 proteins across the NCI-60. **C.** Stoichiometric relationship between PCNA and MCM3 proteins across the NCI-60.



Supplementary Figure 39, Related to Figure 2. CellminerCDB snapshot showing reproducible expression of PCNA determined by SWATH and RPPA (Reverse Phase Protein Array). The snapshot from https://discover.nci.nih.gov/CellMinerCDB shows PCNA protein levels across the NCI-60.

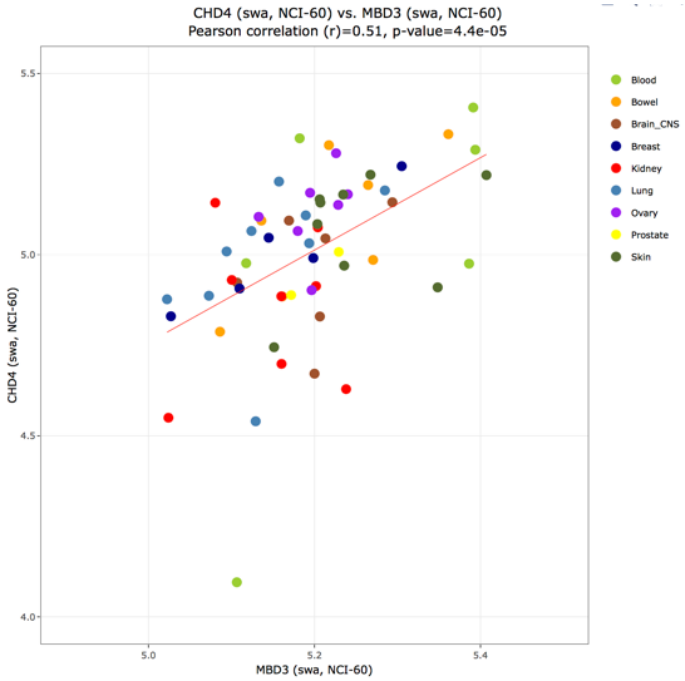


Supplementary Figure 40, Related to Figure 2 and 4. CellminerCDB snapshot of FUS and DHX9 transcript expression (https://discover.nci.nih.gov/CellMinerCDB) **across the MGH-Sanger cell lines**.

SWATH											
Pearsons correlation											
Identifier	RBBP7	RBBP4	MTA3	MTA1	HDAC2	HDAC1	GATAD2B	GATAD2A	MBD3	CHD4	ZMYND8
RBBP7	1	0.226	-0.031	0.064	-0.01	0.197	0.127	-0.053	-0.019	0.12	-0.136
RBBP4	0.178	1	-0.091	0.245	0.271	0.539	0.073	0.18	0.168	0.343	-0.063
MTA3	-0.245	-0.099	1	0.325	0.039	0.076	-0.036	-0.254	-0.159	-0.083	0.016
MTA1	0.228	0.047	-0.173	1	0.457	0.214	0.147	-0.017	0.125	0.106	-0.364
HDAC2	0.133	0.026	-0.043	**0.518**	1	0.419	-0.067	0.307	0.286	0.351	-0.118
HDAC1	**0.391**	**0.497**	-0.127	0.189	-0.004	1	0.05	0.36	0.049	0.497	0.027
GATAD2B	0.24	0.29	-0.088	0.226	0.219	0.271	1	0.218	-0.069	0.366	0.047
GATAD2A	0.274	0.179	0.061	**0.306**	0.277	0.294	0.095	1	0.248	0.451	0.297
MBD3	**0.316**	**0.447**	-0.086	**0.443**	**0.313**	**0.434**	**0.438**	**0.421**	1	0.171	-0.084
CHD4	0.269	0.238	-0.201	**0.311**	0.138	**0.382**	**0.402**	**0.419**	**0.506**	1	0.316
ZMYND8	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	1

Transcripts											
Pearsons correlation											
Identifier	RBBP7	RBBP4	MTA3	MTA1	HDAC2	HDAC1	GATAD2B	GATAD2A	MBD3	CHD4	ZMYND8
RBBP7	1	0.226	-0.031	0.064	-0.01	0.197	0.127	-0.053	-0.019	0.12	-0.136
RBBP4	0.226	1	-0.091	0.245	0.271	0.539	0.073	0.18	0.168	0.343	-0.063
MTA3	-0.031	-0.091	1	0.325	0.039	0.076	-0.036	-0.254	-0.159	-0.083	0.016
MTA1	0.064	0.245	**0.325**	1	0.457	0.214	0.147	-0.017	0.125	0.106	-0.364
HDAC2	-0.01	0.271	0.039	**0.457**	1	0.419	-0.067	0.307	0.286	0.351	-0.118
HDAC1	0.197	**0.539**	0.076	0.214	**0.419**	1	0.05	0.36	0.049	0.497	0.027
GATAD2B	0.127	0.073	-0.036	0.147	-0.067	0.05	1	0.218	-0.069	0.366	0.047
GATAD2A	-0.053	0.18	-0.254	-0.017	0.307	**0.36**	0.218	1	0.248	0.451	0.297
MBD3	-0.019	0.168	-0.159	0.125	0.286	0.049	-0.069	0.248	1	0.171	-0.084
CHD4	0.12	**0.343**	-0.083	0.106	**0.351**	**0.497**	**0.366**	**0.451**	0.171	1	0.316
ZMYND8	-0.136	-0.063	0.016	-0.364	-0.118	0.027	0.047	0.297	-0.084	**0.316**	1

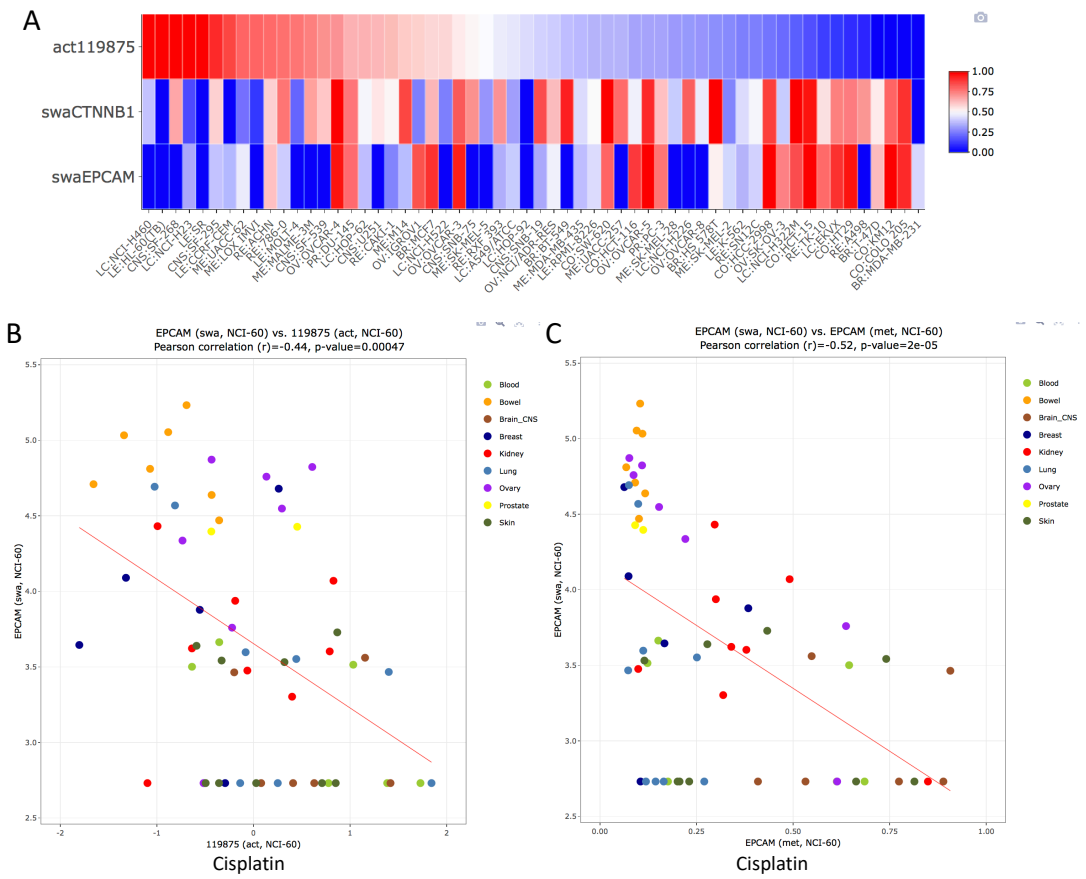
Supplementary Figure 41, Related to Figure 2. Snapshot views showing co-expression of the NuRF (Nucleosome Remodeling Factors) proteins determined by SWATH (top) and transcripts (bottom). The "Cross-correlations" tool of CellMiner was used with the listed proteins or genes (left column). Snapshots from the Excel files obtained from http://discover.nci.nih.gov/cellminer.



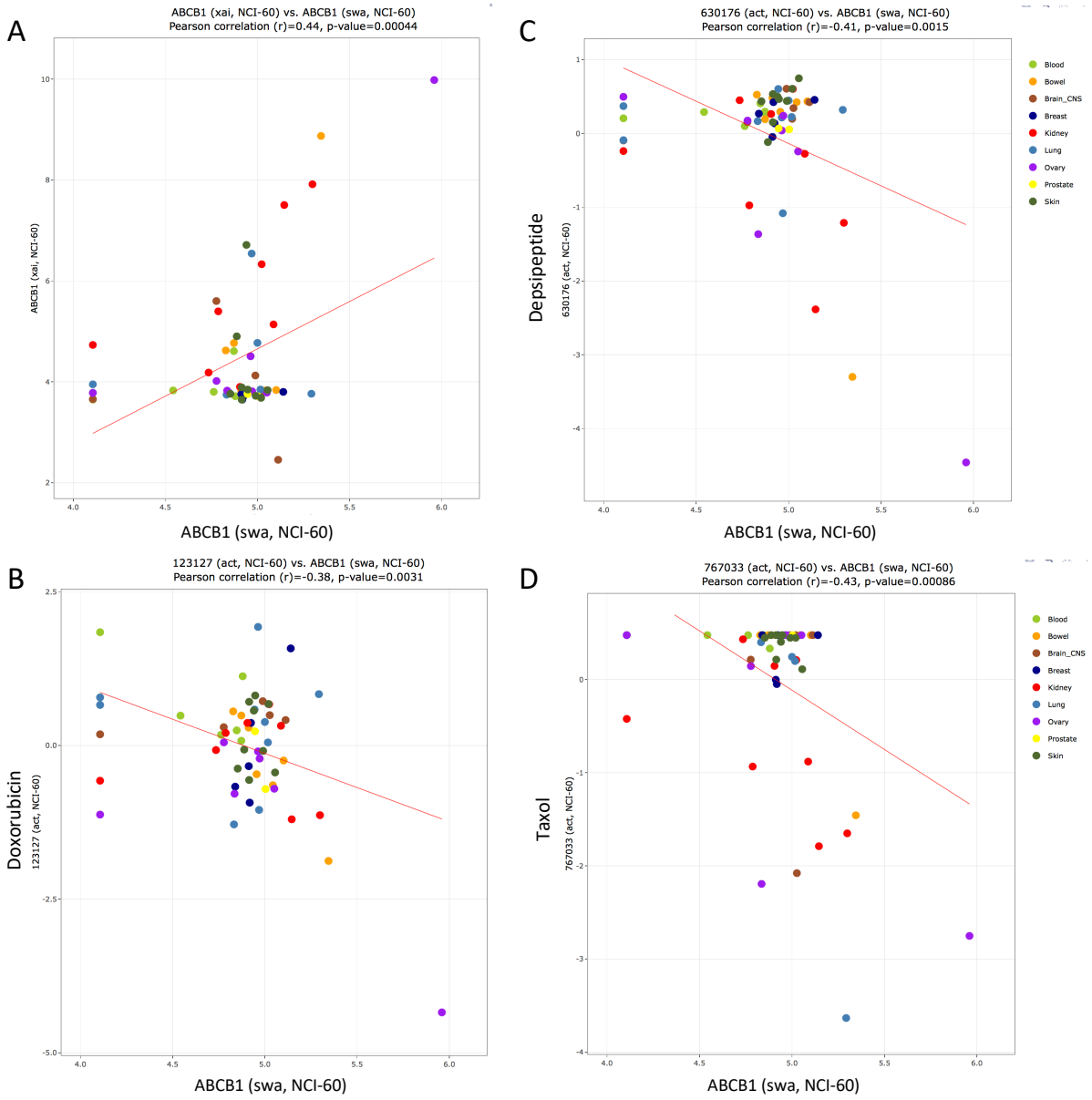
Supplementary Figure 42, Related to Figure 2. CellminerCDB snapshot showing stoichiometry expression of MBD3 and CHD4 determined by SWATH. The https://discover.nci.nih.gov/CellMinerCDB snapshot show MBD3 and CHD4 protein levels across the NCI-60.

Data Type	ID	Location	Correlation	P-Value	FDR	Annotation
swa	All	All	All	All	All	All
swa	CTNNB1	3p21	1	0	0	Apoptosis; Cell Signaling; Oncogenes
swa	CTNND1	11q11	0.814	4.42e-15	2.78e-10	morphogenesis of a polarized epithelium;adherens junction organization
swa	CTNNA1	5q31.2	0.784	2.18e-13	9.15e-9	Apoptosis; Tumor Suppressors
swa	RAB6A	11q13.3	0.531	0.0000154	0.0303	peptidyl-cysteine methylation;protein localization in Golgi apparatus
swa	CTNNA2	2p12-p11.1	0.523	0.0000216	0.0321	radial glia guided migration of Purkinje cell;muscle cell differentiation
swa	TROVE2	1q31	-0.505	0.0000457	0.0384	transcription from RNA polymerase III promoter
swa	SLC1A5	19q13.3	0.502	0.0000502	0.0395	Solute Carriers
swa	S100A16	1q21	0.49	0.0000803	0.0455	response to calcium ion
swa	COL5A2	2q14-q32	0.466	0.000199	0.0606	collagen fibril organization;skin development
swa	CD9	12p13.3	0.459	0.000258	0.0661	response to water deprivation;platelet degranulation
swa	PGK1	Xq13.3	-0.457	0.000276	0.0684	glucose metabolic process;gluconeogenesis
swa	RAC2	22q13.1	-0.456	0.000282	0.069	GTP catabolic process;actin cytoskeleton organization
swa	SIPA1	11q13	0.456	0.000286	0.069	Apoptosis
swa	EPCAM	2p21	0.454	0.000301	0.0697	positive regulation of cell proliferation;ureteric bud development
swa	DHRS7	14q23.1	0.453	0.000313	0.0702	
swa	INF2	14q32.33	0.447	0.000388	0.0756	cellular component organization;actin cytoskeleton organization
swa	IBA57	1q42.13	0.443	0.000442	0.079	glycine catabolic process;heme biosynthetic process
swa	VAPB	20q13.33	0.436	0.000555	0.0845	endoplasmic reticulum unfolded protein response;small molecule metabolic process
swa	NDUFAF3	3p21.31	0.433	0.000619	0.0871	mitochondrial respiratory chain complex I assembly
swa	DSP	6p24	0.431	0.000656	0.0873	cell-cell adhesion;peptide cross-linking
swa	ANXA2	15q22.2	0.43	0.000683	0.0889	positive regulation of vesicle fusion;fibrinolysis

Supplementary Figure 43, Related to Figure 2 and 4. CellminerCDB snapshot showing co-expression of cell adhesion proteins determined by SWATH. The "Compare pattern" tool was used with β-catenin (CTNNB1) as the "x-axis entry" (https://discover.nci.nih.gov/CellMinerCDB). Only the top correlates are shown among over 3,000 proteins in the database.



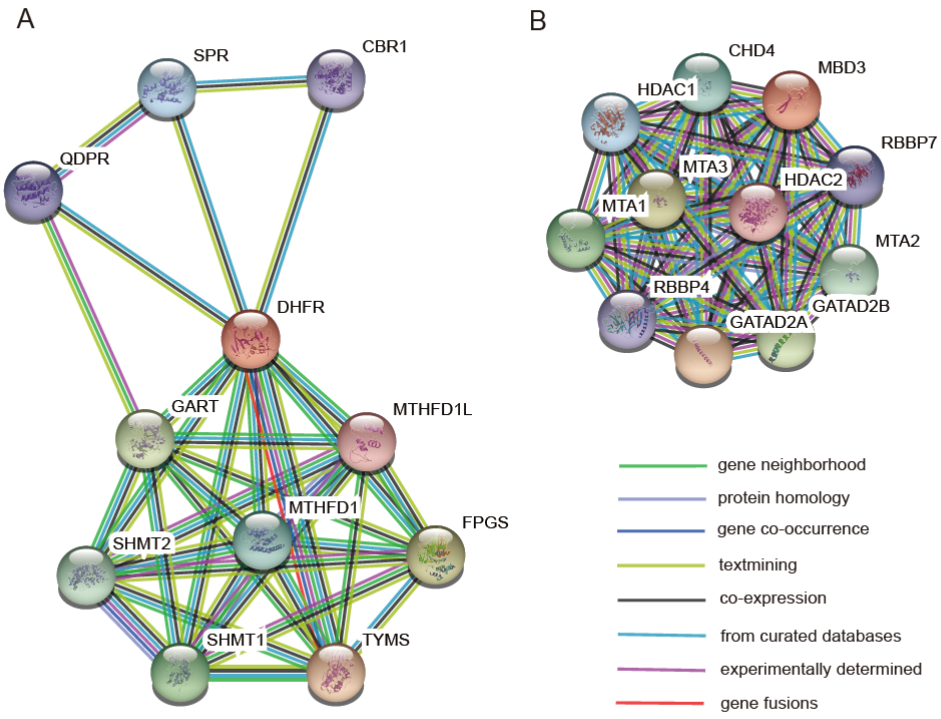
Supplementary Figure 44, Related to Figure 4. Predictive protein biomarkers for cisplatin (NSC119875) activity. The snapshots from https://discover.nci.nih.gov/CellMinerCDB show: **A.** Results obtained with the "Regression Model" tool of CellMinerCDB using cisplatin as "Response Identifier" for the query. **B.** Significant negative correlation between EPCAM protein expression determined by SWATH and activity of cisplatin. **C.** Highly significant negative correlation between EPCAM protein expression and EPCAM promoter methylation across the NCI-60.



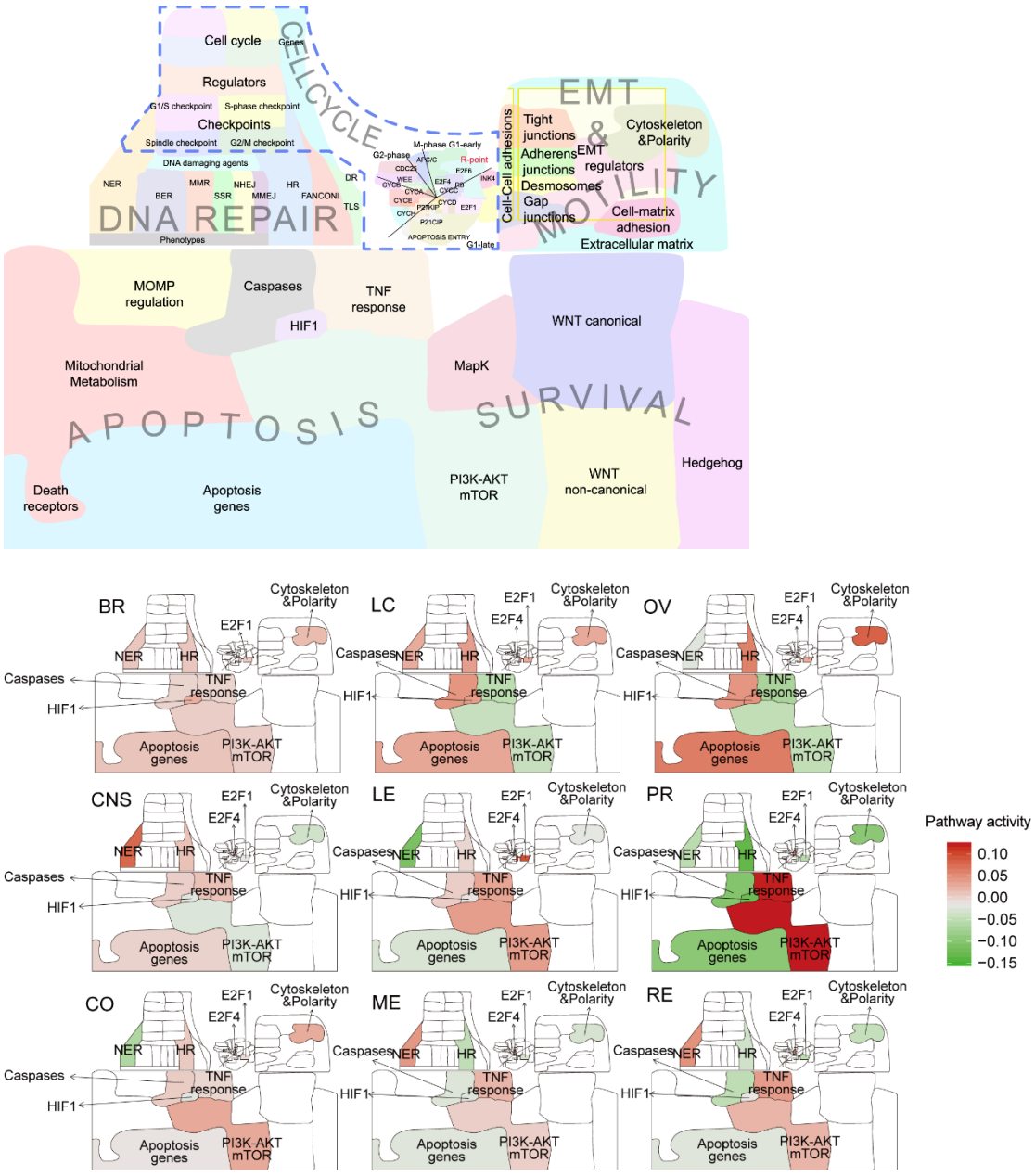
Supplementary Figure 45, Related to Figure 4. ABCB1 (PGP; P-glycoprotein) protein levels across the NCI-60 and prediction of drug response. https://discover.nci.nih.gov/CellMinerCDB snapshots showing: **A.** The correlation between ABCB1 protein and gene expression across the NCI-60. **B-D.** The significant negative correlations between ABCB1 protein levels and response to doxorubicin, depsipeptide and taxol across the NCI-60.

Data Type	ID	Location	Correlation	P-Value	FDR	Annotation
swa	All	All	All	All	All	All
swa	RPA3	7p22	1	0	0	DNA Damage Response (DDR); DDR (MMR); DDR (DNA replication)
swa	RPA2	1p35	0.752	6.58e-12	2.07e-7	DNA Damage Response (DDR); DDR (MMR); DDR (DNA replication)
swa	MCM5	22q13.1	0.628	1.02e-7	0.000985	DDR (DNA replication)
swa	MCM7	7q21.3-q22.1	0.625	1.22e-7	0.00102	DDR (DNA replication)
swa	PFDN2	1q23.3	0.615	2.25e-7	0.00142	protein folding;cellular protein metabolic process
swa	ITGB1	10p11.2	-0.612	2.56e-7	0.00153	Apoptosis
swa	DUT	15q21.1	0.6	5.02e-7	0.00208	DNA Damage Response (DDR)
swa	NUDC	1p35-p34	0.598	5.63e-7	0.00208	mitotic cell cycle;cytokinesis
swa	LRRC47	1p36.32	0.598	5.85e-7	0.00208	
swa	MCM3	6p12	0.595	6.71e-7	0.00208	DDR (DNA replication)
swa	TBCA	5q14.1	0.592	7.73e-7	0.00214	'de novo' posttranslational protein folding;protein folding
swa	PAICS	4q12	0.591	8.21e-7	0.00217	purine base metabolic process;purine nucleotide biosynthetic process
swa	UBE2K	4p14	0.588	9.67e-7	0.0023	ubiquitin-dependent protein catabolic process;protein ubiquitination
swa	RFC4	3q27	0.588	9.71e-7	0.0023	DNA Damage Response (DDR); DDR (MMR); DDR (DNA replication)
swa	PLEC	8q24	-0.585	0.00000116	0.00257	cell junction assembly;apoptotic process
swa	EIF3G	19p13.2	0.583	0.00000126	0.0026	cellular protein metabolic process;translation
swa	PFDN6	6p21.3	0.581	0.00000143	0.0027	
swa	HMGB2	4q31	0.581	0.00000144	0.0027	DNA Damage Response (DDR); DDR (BER)
swa	RFC2	7q11.23	0.578	0.00000163	0.00296	DNA Damage Response (DDR); DDR (MMR); DDR (DNA replication)
swa	ETF1	5q31.1	0.576	0.0000018	0.00305	nuclear-transcribed mRNA catabolic process, nonsense-mediated decay;cellular protein metabolic process
swa	PSMF1	20p13	0.575	0.00000186	0.00305	S phase of mitotic cell cycle;apoptotic process

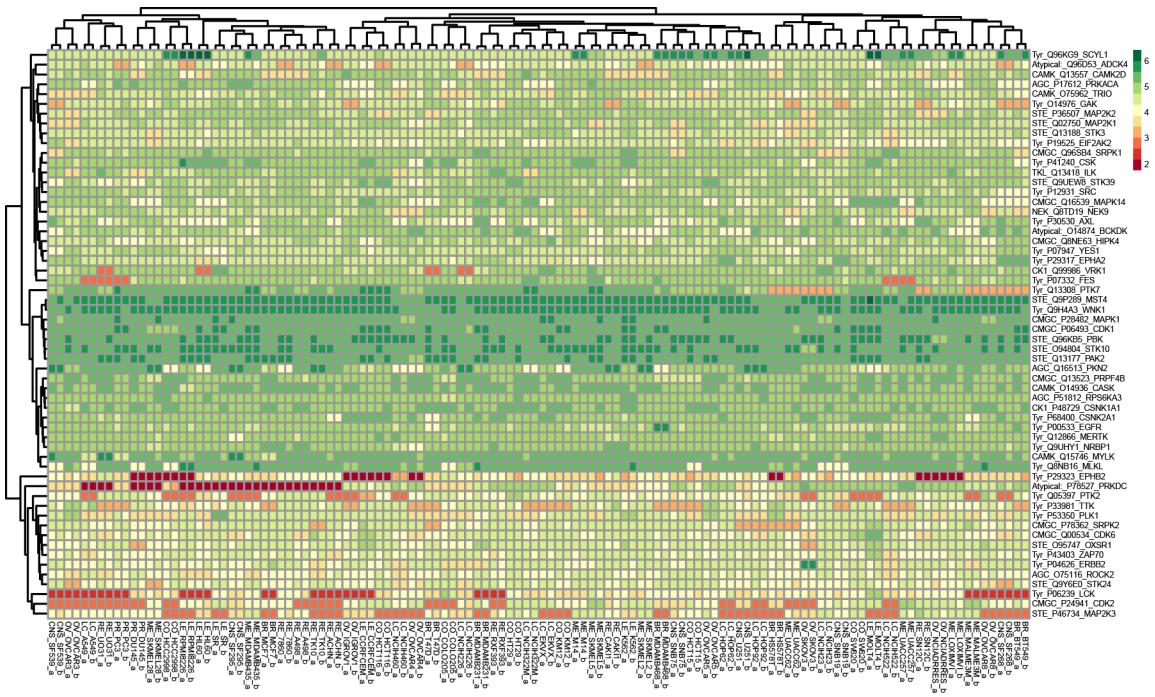
Supplementary Figure 46, Related to Figure 2. CellminerCDB snapshot views showing co-expression of replication proteins determined by SWATH. The "Compare pattern" tool was used with RPA3 as the "x-axis entry". Snapshots from https://discover.nci.nih.gov/CellMinerCDB showing the top correlates with RPA3.



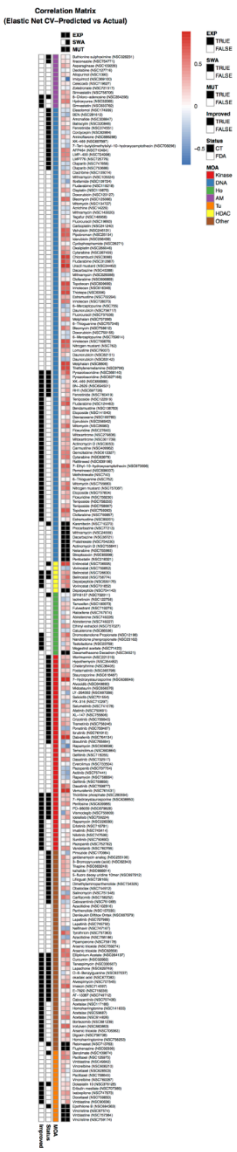
Supplementary Figure 47, Related to Figure 2. Proteins interacting with DHFR and MBD3. Proteins interacting with DHFR (A) and MBD3 (B) from STRING.



Supplementary Figure 48, Related to Figure 3. Global cancer signaling pathway maps based on Atlas of Cancer Signaling Network (ACSN) pathways (Kuperstein et al., 2015) (www. acsn.curie.fr). The annotations of the pathway map are shown in the upper panel. ROMA representation of the pathway activities are shown in the lower panel.



Supplementary Figure 49, Related to Figure 3. Fifty-eight protein kinases quantified in the NCI-60 cell panel. Expression of 58 protein kinases, represented by log10 transformed protein intensity values, in the NCI-60. Values are clustered without supervision across both proteins and cell lines.



Supplementary Figure 50, Related to Figure 3. Predictive power of different omics data combinations for the activity of 224 compounds, based on elastic net (multivariate linear regression) modeling of the drug response. Each column indicates the input data gene expression and mutation alone and in combination with proteomic abundances; each row represents a compound. The color indicates the predictive power, measured by Pearson correlation of cross-validation predicted and observed drug response values. Rows specifying compound-specific response prediction accuracies are sorted by mechanism of action and additional annotations are provided 1) whether the inclusion of the SWATH data improved the overall model and 2) the clinical status of the compound whether FDA approved or in clinical trial.

Supplementary Table Legends

All the supplementary tables are provided as separate Excel spreadsheet.

Supplementary Table 1, Related to Figure 1. Quantitative proteome maps of the NCI-60 cell lines. (A) Information for PCT-SWATH analysis of the NCI-60 cells (B) List of peptide precursors appeared in the library. (C) Quantitative values of 22,554 peptide precursors in the NCI-60 cells in duplicates. (D) Quantitative values of 3,171 proteins in the NCI-60 cells in duplicates. (E) Averaged protein intensity in the NCI-60 cells.

Supplementary Table 2, Related to Figure 2. Count of proteins in each cell line in the DDA data and the SWATH data. This table shows the count of IPI protein group number from the DDA data set quantified using iBAQ and LFQ algorithms 1, and the count of SwissProt proteotypic proteins from the SWATH data as reported in this data(Gholami et al., 2013a).

Supplementary Table 3, Related to Figure 2. Correlation between NCI-60 transcript expression and SWATH-MS protein expression for indicated gene.

Supplementary Table 4, Related to Figure 2. Stoichiometry of 101 protein complexes in the NCI-60 proteotype. Average Abundance (log10) means the averaged log 10 scaled protein abundance signal for proteins in a complex. Standard Deviation means the standard deviation of log 10 scaled protein abundance signal for proteins in a complex. Average Pearson Correlation means the averaged Pearson corelation value for each pair of proteins in a complex.

Supplementary Table 5, Related to Figure 2. The activity of apoptosiswas found significantly higher in ovarian cell lines. (A) The modules that show a significant dispersion are reported here. (B) A t-test is performed for cell lines from one cancer type vs. all other cancer cell lines.

Supplementary Table 6, Related to Figure 3. Cellminer data for the NCI-60 cells used in this study. (A) Exome data of the NCI-60 cells. (B) Log2 scaled mRNA expression data of the NCI-60 cells. (C) Common features at three different levels, i.e. DNA, mRNA and protein.

Supplementary Table 7, Related to Figure 3. Elastic net results.

Transparent Methods

PCT-assisted sample preparation for MS analyses

The NCI-60 cells were obtained as frozen, non-viable cell pellets from the Developmental Therapeutics Program (DTP), National Cancer Institute (NCI-NIH) and processed using Barocycler® NEP2320 (PressureBioSciences Inc, South Easton, MA). The IDs of the NCI-60 cells in our study matching to the IDs in Cellminer and a previous proteomic study by the Kuster group are provided in **Supplementary Table 1**. Briefly, cell pellets were lysed in a buffer containing 8M urea, 0.1M ammonium bicarbonate, and Complete^{TM} protease inhibitor using barocycler program (20 seconds 45 kpsi, 10 seconds 0 kpsi, 120 cycles) at 35°C (Guo et al., 2015). Whole cell lysates were sonicated for 25 seconds with 1 min interval on ice for 3 times. Cellular debris was removed by centrifugation and sample protein concentration was determined by BCA assay prior to protein reduction with 10 mM TCEP for 20 min at 35°C, and alkylation with 40 mM iodoacetamide in the dark for 30 min at room temperature. Lys-C digestion (1/50, w/w) was performed in 6 M urea using PCT program (25 seconds 25 kpsi, 10 seconds 0 kpsi 75 cycles) at 35°C; whereas trypsin digestion (1/30, w/w) was performed in further diluted urea (1.6M) using PCT program (25 seconds 25 kpsi, 10 seconds 0 kpsi, 160 cycles) at 35°C. Digestion was stopped by acidification with trifluoroacetic acid to a final pH of around 2 before C18 column desalting using SEP-PAK C18 cartridges (Waters Corp., Milford, MA, USA).

Off-gel electrophoresis

To create a comprehensive spectral library for SWATH-MS analysis, we pooled 20-40% of desalted peptide solutions from each NCI-60 sample and performed off-gel fractionation. Briefly, pooled peptides were resolubilised in OGE buffer containing 5% (v/v) glycerol, 0.7% (v/v) acetonitrile (ACN) and 1% (v/v) carrier ampholytes mixture (IPG buffer pH 3.0-10.0, GE Healthcare). Fractionation was performed on a 3100 OFFGEL (OGE) Fractionator (Agilent Technologies) using a 24 cm pH3-10 IPG strip (Immobilised pH Gradient strip from GE Healthcare) according to manufacturer's instructions using a program of 1 h rehydration at a maximum of 500 V, 50 µA and 200 mW followed by separation at a maximum of 8000 V, 100 µA and 300 mW until 50 kVh were reached. Each of 24 fraction was recovered and cleaned up by C18 reversed-phase MicroSpin columns (The Nest Group Inc.). Based on the sample complexity (based on Nanodrop, A280 measurement), for each strip, the following fractions were pooled into 12 samples for MS injections: pool 1 (fraction 1-2), pool 2 (fraction 3), pool 3 (fraction 4), pool 4 (fraction 5), pool 5 (fraction 6-7), pool 6 (fraction 8-9), pool 7 (fraction 10-11), pool 8 (fraction 12-15), pool 9 (fraction 16-19), pool 10 (fraction 20-

21), pool 11 (fraction 22), pool 12 (fraction 23-24). Those were injected in quadruplicate, resulting in 48 DDA injections of fractionated samples.

DDA MS for spectral library generation

For spectral library generation, a SCIEX TripleTOF 5600 System mass spectrometer was operated essentially as described before (Schubert et al., 2015): all samples were analyzed on an Eksigent nanoLC (AS-2/1Dplus or AS-2/2Dplus) system coupled with a SWATH-MS-enabled AB SCIEX TripleTOF 5600 System. The HPLC solvent system consisted of buffer A (2% ACN and 0.1% formic acid, v/v) and buffer B (95% ACN with 0.1% formic acid, v/v). Samples were separated in a 75 µm diameter PicoTip emitter (New Objective) packed with 20 cm of Magic 3 µm, 200A C18 AQ material (Bischoff Chromatography). The loaded material was eluted from the column at a flow rate of 300 nL min^{-1} with the following gradient: linear 2 - 35% B over 120 min, linear 35 - 90% B for 1 min, isocratic 90% B for 4 min, linear 90 - 2% B for 1 min and isocratic 2% solvent B for 9 min. The mass spectrometer was operated in DDA mode using a top20 method, with 500 ms and 150 ms acquisition time for the MS1 and MS2 scans respectively, and 20 s dynamic exclusion for the fragmented precursors. Rolling collision energy using the following equation (0.0625 × m/z - 3.5) with a collision energy spread of 15 eV was used for fragmentation regardless of the charge state of the precursors, to mimic as close as possible the fragmentation conditions of the precursors in SWATH-MS mode. Altogether, we had 66 DDA-MS injections, including the 48 OGE samples and another 18 pooled peptide samples from the unfractionated cell lysate of the NCI-60 cells.

Spectral and assay library generation

All raw instrument data were centroided using Proteowizard msconvert (version 2.0). The assay library was generated using an established protocol (Schubert et al., 2015). In short, the shotgun data sets were searched individually using X!Tandem (Craig and Beavis, 2003) (2011.12.01.1) with k-score plugin (MacLean et al., 2006), Myrimatch (Tabb et al., 2007) (2.1.138), OMSSA (Geer et al., 2004) (2.1.8) and Comet (Eng et al., 2013) (2013.02r2) against the reviewed UniProtKB/Swiss-Prot (2014_02) protein sequence database containing 20,270 proteins appended with 11 iRT peptides and decoy sequences. Carbamidomethyl was used as a fixed modification and oxidation as the variable modification. Maximally two missed cleavages were allowed. Peptide mass tolerance was set to 50 ppm, fragment mass error to 0.1 Da. The search identifications were combined and statistically scored using PeptideProphet (Keller et al., 2002) and iProphet (Shteynberg et al., 2011) available within the Trans-Proteomics Pipeline (TPP) toolset (version 4.7.0) (Keller et al., 2005). MAYU

(Reiter et al., 2009) (v. 1.07) was used to determine the iProphet cutoff (0.999354) corresponding to a protein FDR of 1.03%. SpectraST was used in library generation mode with CID-QTOF settings and iRT normal-isation at import against the iRT Kit (Escher et al., 2012) peptide sequences (-c_IRTirt.txt -c_IRR) and a consensus library was consecutively generated. An in-house python script, spec-trast2tsv.py31 (msproteomicstools 0.2.2) was then used to generate the assay library with the following settings: -l 350,2000 -s b,y -x 1,2 -o 6 -n 6 -p 0.05 -d -e -w swath32.txt -k openswath (fragment ions between 350 and 2000 m/z, b and y ions authorized, fragment charges 1+ and 2+, 6 most intense transitions, precision of fragment ion retrieved 0.05 Da, exact fragment ion mass calculated, exclude fragments in the swath window). The OpenSWATH tool, ConvertTSVToTraML converted the TSV file to TraML format; Open-SwathDecoyGenerator generated the decoy assays in shuffle mode and appended them to the TraML assay library. In this study, we built a SWATH assay library containing 86,209 proteotypic peptide precursors in 8,056 proteotypic SwissProt proteins. This library is supplied in PRIDE project PXD003539.

SWATH-MS

The SWATH-MS data acquisition in a Sciex TripleTOF 5600 mass spectrometer was performed as described before (Gillet et al., 2012), using 32 windows of 25 Da effective isolation width (with an additional 1 Da overlap on the left side of the window) and with a dwell time of 100 ms to cover the mass range of 400 - 1200 *m/z* in 3.3 s. The collision energy for each window was set using the collision energy of a 2+ ion centered in the middle of the window (equation: 0.0625 x *m/z* - 3.5) with a spread of 15 eV. The sequential precursor isolation window setup was as follows: [400-425], [424-450], [449-475], …, [1174-1200].

Protein identification using OpenSWATH

We analyzed the SWATH data using OpenSWATH software (Rost et al., 2014) using parameters as described previously (Ori et al., 2016). We identified 48,374 peptides from 6,556 protein groups from the NCI-60 panel with < 1% false discovery rate at both peptide and protein level evaluated by OpenSWATH (Rost et al., 2014)and Mayu (Reiter et al., 2009) (supplied in PRIDE project PXD003539).

DIA-expert analyses

The DIA-expert software read OpenSWATH output result file which contains statistical scores (*i.e.* mProphet score or mScore) indicating the confidence of identification for each

peptide precursor in each sample, and from there selected the sample in which a peptide precursor was identified with highest confidence. It then obtained extracted ion chromatograms (XICs) for the target peptide precursor and all associated annotated *b* and *y* fragments in the reference sample, and refined fragments based on the peak shape of each fragment and its peak boundary. The refined fragments and precursor XIC traces from each of the rest samples were subsequently compared with the reference peak group using empirical expert rules, based on which the best matched peak group in each sample was picked and visualized. Duplicated measurements were used to evaluate the accuracy of peptide and protein quantification. The protein quantity was normalized based on total ion chromatography of the MS1 spectra from each raw SWATH file. All codes are provided in Github https://github.com/tiannanguo/dia-expert.

PRM analysis

PRM quantification strategy was used to quantify selected proteins. Biognosys-11 iRT peptides (Biognosys, Schlieren, CH) were spiked into peptide samples at the final concentration of 10% prior to MS injection for RT calibration. Peptides were separated at 300 nL/min along a 45min 8–35% linear LC gradient (buffer A: 2% ACN, 0.1% formic acid; buffer B: 20% ACN, 0.1% formic acid). The Q Exactive HF-X Hybrid Quadrupole-Orbitrap Mass Spectrometer was operated in the MS/MS mode with time-scheduled acquisition for 54 peptides in a +/- 5 min retention time window. The full MS mode was measured at resolution 60,000 at m/z 200 in the Orbitrap, with AGC target value of 3E6 and maximum IT of 55ms. Target ions were submitted to MS/MS in the HCD cell (1.2 amu isolation width, 30% normalized collision energy). MS/MS spectra were acquired at resolution 30,000 (at m/z 200) in the Orbitrap using AGC target value of 2E5, a max IT of 100ms.

Quantitative proteomics and transcriptomics analysis of protein complexes components

Technical replicates were averaged to generate the NCI-60 proteotypes. To assess the coverage of protein complexes by NCI-60 proteotypes, we first retrieved a large resource of mammalian protein complexes assembled from CORUM (Ruepp et al., 2010), COMPLEAT (Vinayagam et al., 2013) and literature-curated complexes (Ori et al., 2013; Ori et al., 2016). This resource contains 2,041 proteins as members of 279 distinct complexes and it is available at http://variablecomplexes.embl.de/. 101 complexes were represented in the NCI-60 proteotypes with at least 5 members quantified. These complexes, in total, contain 1,045 distinct proteins quantified in the NCI-60 proteotypes. Pearson's correlation coefficient was calculated for all the pairwise comparisons of 3,171 proteins across the NCI-60 cell lines. All

pairwise comparisons were classified into two categories: either two proteins were members of the same complex or not. Average abundance, standard deviation and average Pearson correlation of each complex were calculated based on the abundance of complex members in the NCI-60 proteotypes.

An extended list of protein-protein interactions (PPIs) was generated based on information acquired from 6 resources: 1) 17,556 PPIs were retrieved from the CORUM database of human protein complexes (Ruepp et al., 2010); 2) 16'345 PPIs were composed from the interaction pairs annotated as 'complex' members in the Reactome database (Fabregat et al., 2018); 3) 12,664 PPIs were retrieved from the STRING database (Szklarczyk et al., 2015) considering only high confidence interactions (score ≥ 700). 4) 1'378 interaction pairs were obtained from Interactome3D (Mosca et al., 2013). These interactors corresponded to experimentally observed interactions with a support in the form of structural data or structural models. 5) 309 PPIs were obtained by considering interactions identified in at least 3 independent APMS experiments. For this, we included studies deposited in the BioGrid database (Chatr-Aryamontri et al., 2017), interactions listed in the BioPlex portal (Huttlin et al., 2015), and interactions observed in the large-scale Polycomb (Hauri et al., 2016) and Kinome studies (manuscript in preparation). 6) 122 PPIs were retrieved from the EMBL-EBI complex portal (Park et al., 2017). The latter (smallest) set of interactions is manually curated and of high confidence.

Combining information from the different databases, a list of 35,693 unique interactions (encompassing 1,766 proteins) was generated. The Spearman coefficient of correlation of protein abundances (Spearman's r) and the associated p-value were calculated for all the 5,026,035 protein pairs that can be formed from the 3171 proteins measured in the proteomics dataset. For this, the cor.mtest function from the package corrplot was applied with the Benjamini-Hochberg correction for multiple testing. Distribution of pairwise correlation values for three different sets was visualized with the density plots. The sets represented pairs found to interact in the respective database, all background NCI60 pairs (common to all analysis) and protein pairs that were both measured by NCI60 and present in the respective database, but not reported as interacting. The mean correlation values between the datasets were compared with the Wilcoxon test in R.

Pairwise correlation analysis of the mRNA levels was based on the expression data retrieved from the CellMiner. Cell lines with missing values (CNS.SF_539, ME.MDA_N and LC.NCI_H23) and transcripts for which the matching proteins were not measured were excluded from the analysis. Therefore, the final analysis was performed on a complete matrix with 57 cell lines and 2,835 transcripts. The Spearman's r and associated p-values were calculated as above for the 4,017,195 mRNA pairs that can be formed from the 2835

measured transcripts. Distribution of correlation values was compared between the set of true interaction partners and the corresponding background sets as described above.

Pathway activity analysis

The activity of pathways, as they are described in ACSN, has been computed using ROMA (Martignetti et al., 2016). Among all the modules defined in ACSN, only 11 show a significant dispersion over the data set: AKT_MTOR, HR (Homologous Recombination), NER (nucleotide Excision Repair), TNF response, Death Receptors regulators, Apoptosis, caspases, E2F3 and E2F4 targets, HIF1 and cytoskeleton polarity. For these modules, the mean activity score for each type of cancer cell lines was computed and mapped onto the atlas (from bright green for low values to bright red for high values). To assess module differential activity between proteotypes, we computed a *t*-test on the activity scores in cell lines of a cancer type versus the activity of all other cancer cell lines. The definition of genes composing each module can be found in http://acsn.curie.fr

Drug sensitivity prediction using elastic net

The elastic net regularized regression algorithm was applied to predict drug response for 240 FDA-approved or investigational NSC-designated compounds. Some widely studied drugs are represented by more than one NSC identifier, with each identifier associated with a distinct compound sample and series of NCI-60 drug activity assays. For each compound, two sets of input data were evaluated. These included NCI-60 mRNA expression, gene-level mutation alone and in combination with SWATH-MS protein expression. mRNA expression data was available for 25,040 genes, and derived from CellMiner (discover.nci.nih.gov/cellminer and discover.nci.nih.gov/cellminecdb) (Rajapakse et al., 2018; Reinhold et al., 2012; Reinhold et al., 2015; Reinhold et al., 2017), with missing values imputed using the impute.knn function (with default parameters) of the Bioconductor impute package. Gene-level mutation profiles were available for 9,307 genes, and were obtained from CellMiner exome sequencing data, with values indicating the percent conversion to a variant form for the case of expected function-impacting alterations (frameshift, nonsense, splice-sense, missense mutations by SIFT/PolyPhen2 analysis). SWATH-MS based protein expression data from the current study was also included.

Elastic net analysis was done using the glmnet R package (Friedman et al., 2010). The elastic net analysis was conducted using a multi-step pipeline involving cross-validations performed in a nested manner. The “outer” cross-validation is a leave-one-out cross validation that is conducted over all computational steps present in the “inner” pipeline, and it is used to

validate model performance. The “inner” cross-validation are conducted to select elastic net hyperparameters (alpha and lambda) and for predictor set trimming, using data from a set of ~59 cell lines.

The elastic net parameters alpha and lambda were selected by minimizing the cross-validation error (average of 10 replicates of 10-fold cross-validation) within the “inner” pipeline. The selected alpha and lambda parameters were then applied to 200 runs of the elastic net algorithm, each using a random data subset derived from 90% of the available cell lines. The 200 resulting coefficient vectors were then averaged, and predictors were ranked by the magnitude of their average coefficient weight. To select a limited number of predictors with potential to generalize to new data, top k-element predictor sets (by average coefficient weight magnitude) were evaluated using standard linear regression and 10-fold cross-validation. The appropriate k was set to the smallest value yielding a cross-validation error within one standard deviation of the minimum cross-validation error.

To obtain a robust estimate of performance on unseen data, leave-one-out cross-validation was applied to the overall procedure as part of the “outer” pipeline. Specifically, drug response for each cell line was predicted using an elastic net model derived using the remaining held out data (and the steps outlined above). The vector of predicted response values was then correlated with the actual response values, with the Pearson’s correlation coefficient providing an estimate of the predictive value of the applied input data combination. More details of the elastic net algorithm are provided in Supplementary Note 6.

Elastic net analysis was done using the rcellminerElasticNet R package (https://bitbucket.org/cbio_mskcc/rcellminerelasticnet), which facilitates the application of the glmnet R package (which provides the elastic net algorithm code) to data from the rcellminer and rcellminerData packages (Luna et al., 2016). rcellminerElasticNet also provides utility functions for summarizing and visualizing elastic net results.

Results for the elastic net analysis are available from this URL: https://discover.nci.nih.gov/cellminerreviewdata/swath_analysis/swathOutput_062316_all.tar.gz. This compressed file contains results for the analysis run with all features and selected common features. Each drug compound has three files for each combination of molecular features used in a particular run of the elastic net algorithm: 1) a knitr report R Markdown (.Rmd) file containing the code that was run, 2) an RData (.Rdata) file containing the results of each elastic net run (see elasticNet() documentation in the rcellminerElasticNet package), 3) the rendered knitr report as a webpage (.html).

Beyond the knitr report containing code, the elastic net pipeline is made reproducible using a Docker image. Docker (www.docker.com) is an emerging platform for conducting reproducible research in the biomedical research community. All necessary software and

dependencies to run the described analysis have been embedded in the available Docker container to provide readers an environment that runs on all major operating systems (including Windows, OSX, and Linux), making Docker containers self-contained, portable, and capable of performing at levels similar to the host system.

The Docker container is available at the Docker Hub repository: cannin/swath (https://hub.docker.com/r/cannin/swath/). Key dependencies installed, include: RStudio Server (https://www.rstudio.com/), rcellminer/rcellminerData (Luna et al., 2016), and rcellminerElasticNet. With these installed dependencies, readers have the opportunity to 1) re-run analysis for specific drug compounds and modify the code in order to extend the analysis using RStudio Server, a web-based version of the RStudio R editor, and 2) use an R Shiny app web-based data explorer to further understand described results. Instructions on the usage of the Docker container are located at the rcellminerElasticNet project page (https://bitbucket.org/cbio_mskcc/rcellminerelasticnet).

Supplementary References

Chatr-Aryamontri, A., Oughtred, R., Boucher, L., Rust, J., Chang, C., Kolas, N.K., O'Donnell, L., Oster, S., Theesfeld, C., Sellam, A.*, et al.* (2017). The BioGRID interaction database: 2017 update. Nucleic acids research *45*, D369-D379.

Craig, R., and Beavis, R.C. (2003). A method for reducing the time required to match protein sequences with tandem mass spectra. Rapid Commun Mass Spectrom *17*, 2310-2316.

Eng, J.K., Jahan, T.A., and Hoopmann, M.R. (2013). Comet: an open-source MS/MS sequence database search tool. Proteomics *13*, 22-24.

Escher, C., Reiter, L., MacLean, B., Ossola, R., Herzog, F., Chilton, J., MacCoss, M.J., and Rinner, O. (2012). Using iRT, a normalized retention time for more targeted measurement of peptides. Proteomics *12*, 1111-1121.

Fabregat, A., Korninger, F., Viteri, G., Sidiropoulos, K., Marin-Garcia, P., Ping, P., Wu, G., Stein, L., D'Eustachio, P., and Hermjakob, H. (2018). Reactome graph database: Efficient access to complex pathway data. PLoS Comput Biol *14*, e1005968.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw *33*, 1-22.

Geer, L.Y., Markey, S.P., Kowalak, J.A., Wagner, L., Xu, M., Maynard, D.M., Yang, X., Shi, W., and Bryant, S.H. (2004). Open mass spectrometry search algorithm. J Proteome Res *3*, 958-964.

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

Gholami, A.M., Hahne, H., Wu, Z.X., Auer, F.J., Meng, C., Wilhelm, M., and Kuster, B. (2013b). Global Proteome Analysis of the NCI-60 Cell Line Panel. Cell Rep *4*, 609-620.

Gillet, L.C., Navarro, P., Tate, S., Rost, 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. Molecular & cellular proteomics : MCP *11*, O111 016717.

Guo, T., Kouvonen, P., Koh, C.C., Gillet, L.C., Wolski, W.E., Rost, 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. Nature medicine.

Hauri, S., Comoglio, F., Seimiya, M., Gerstung, M., Glatter, T., Hansen, K., Aebersold, R., Paro, R., Gstaiger, M., and Beisel, C. (2016). A High-Density Map for Navigating the Human Polycomb Complexome. Cell Rep *17*, 583-595.

Huttlin, E.L., Ting, L., Bruckner, R.J., Gebreab, F., Gygi, M.P., Szpyt, J., Tam, S., Zarraga, G., Colby, G., Baltier, K.*, et al.* (2015). The BioPlex Network: A Systematic Exploration of the Human Interactome. Cell *162*, 425-440.

Keller, A., Eng, J., Zhang, N., Li, X.J., and Aebersold, R. (2005). A uniform proteomics MS/MS analysis platform utilizing open XML file formats. Mol Syst Biol *1*, 2005 0017.

Keller, A., Nesvizhskii, A.I., Kolker, E., and Aebersold, R. (2002). Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Anal Chem *74*, 5383-5392.

Kuperstein, I., Bonnet, E., Nguyen, H.A., Cohen, D., Viara, E., Grieco, L., Fourquet, S., Calzone, L., Russo, C., Kondratova, M*., et al.* (2015). Atlas of Cancer Signalling Network: a systems biology resource for integrative analysis of cancer data with Google Maps. Oncogenesis *4*, e160.
Luna, A., Rajapakse, V.N., Sousa, F.G., Gao, J., Schultz, N., Varma, S., Reinhold, W., Sander, C., and Pommier, Y. (2016). rcellminer: exploring molecular profiles and drug response of the NCI-60 cell lines in R. Bioinformatics *32*, 1272-1274.
MacLean, B., Eng, J.K., Beavis, R.C., and McIntosh, M. (2006). General framework for developing and evaluating database scoring algorithms using the TANDEM search engine. Bioinformatics *22*, 2830-2832.
Martignetti, L., Calzone, L., Bonnet, E., Barillot, E., and Zinovyev, A. (2016). ROMA: Representation and Quantification of Module Activity from Target Expression Data. Front Genet *7*, 18.
Mosca, R., Ceol, A., and Aloy, P. (2013). Interactome3D: adding structural details to protein networks. Nat Methods *10*, 47-53.
Ori, A., Banterle, N., Iskar, M., Andres-Pons, A., Escher, C., Khanh Bui, H., Sparks, L., Solis-Mezarino, V., Rinner, O., Bork, P*., et al.* (2013). Cell type-specific nuclear pores: a case in point for context-dependent stoichiometry of molecular machines. Mol Syst Biol *9*, 648.
Ori, A., Iskar, M., Buczak, K., Kastritis, P., Parca, L., Andres-Pons, A., Singer, S., Bork, P., and Beck, M. (2016). Spatiotemporal variation of mammalian protein complex stoichiometries. Genome Biol *17*, 47.
Park, Y.M., Squizzato, S., Buso, N., Gur, T., and Lopez, R. (2017). The EBI search engine: EBI search as a service-making biological data accessible for all. Nucleic acids research *45*, W545-W549.
Rajapakse, V.N., Luna, A., Yamade, M., Loman, L., Varma, S., Sunshine, M., Iorio, F., Sousa, F.G., Elloumi, F., Aladjem, M.I*., et al.* (2018). CellMinerCDB for Integrative Cross-Database Genomics and Pharmacogenomics Analyses of Cancer Cell Lines. iScience *10*, 247-264.
Reinhold, W.C., Sunshine, M., Liu, H.F., Varma, S., Kohn, K.W., Morris, J., Doroshow, J., and Pommier, Y. (2012). CellMiner: A Web-Based Suite of Genomic and Pharmacologic Tools to Explore Transcript and Drug Patterns in the NCI-60 Cell Line Set. Cancer Res *72*, 3499-3511.
Reinhold, W.C., Sunshine, M., Varma, S., Doroshow, J.H., and Pommier, Y. (2015). Using CellMiner 1.6 for Systems Pharmacology and Genomic Analysis of the NCI-60. Clinical cancer research : an official journal of the American Association for Cancer Research *21*, 3841-3852.
Reinhold, W.C., Varma, S., Sunshine, M., Rajapakse, V., Luna, A., Kohn, K.W., Stevenson, H., Wang, Y., Heyn, H., Nogales, V*., et al.* (2017). The NCI-60 Methylome and Its Integration into CellMiner. Cancer Res *77*, 601-612.
Reiter, L., Claassen, M., Schrimpf, S.P., Jovanovic, M., Schmidt, A., Buhmann, J.M., Hengartner, M.O., and Aebersold, R. (2009). Protein identification false discovery rates for very large proteomics data sets generated by tandem mass spectrometry. Molecular & cellular proteomics : MCP *8*, 2405-2417.

Rosenberger, G., Koh, C.C., Guo, T., Röst, H.L., Kouvonen, P., Collins, B.C., Heusel, M., Liu, Y., Caron, E., and Vichalkovski, A. (2014). A repository of assays to quantify 10,000 human proteins by SWATH-MS. Scientific data *1*, 140031.
Rost, H.L., Rosenberger, G., Navarro, P., Gillet, L., Miladinovic, S.M., Schubert, O.T., Wolskit, W., Collins, B.C., Malmstrom, J., Malmstrom, L.*, et al.* (2014). OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. Nature Biotechnology *32*, 219-223.
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 research *38*, D497-501.
Schubert, O.T., Gillet, L.C., Collins, B.C., Navarro, P., Rosenberger, G., Wolski, W.E., Lam, H., Amodei, D., Mallick, P., MacLean, B.*, et al.* (2015). Building high-quality assay libraries for targeted analysis of SWATH MS data. Nat Protoc *10*, 426-441.
Shteynberg, D., Deutsch, E.W., Lam, H., Eng, J.K., Sun, Z., Tasman, N., Mendoza, L., Moritz, R.L., Aebersold, R., and Nesvizhskii, A.I. (2011). iProphet: multi-level integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates. Molecular & cellular proteomics : MCP *10*, M111 007690.
Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K.P.*, et al.* (2015). STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic acids research *43*, D447-452.
Tabb, D.L., Fernando, C.G., and Chambers, M.C. (2007). MyriMatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. J Proteome Res *6*, 654-661.
Vinayagam, A., Hu, Y., Kulkarni, M., Roesel, C., Sopko, R., Mohr, S.E., and Perrimon, N. (2013). Protein complex-based analysis framework for high-throughput data sets. Science signaling *6*, rs5.