Published in final edited form as:
Nat Methods. 2022 July ; 19(7): 803–811. doi:10.1038/s41592-022-01526-y.

Mass spectrometry-based draft of the mouse proteome

Piero Giansanti[#1], **Patroklos Samaras**[#1], **Yangyang Bian**[#1,2], **Chen Meng**[3], **Andrea Coluccio**[4,5,6,7], **Martin Frejno**[1], **Hannah Jakubowsky**[4,5,6,7], **Sophie Dobiasch**[8,9,10,11], **Rashmi R. Hazarika**[12,13], **Julia Rechenberger**[1], **Julia Calzada-Wack**[14], **Johannes Krumm**[1], **Sebastian Mueller**[6,7,15], **Chien-Yun Lee**[1], **Nicole Wimberger**[4,5,6,7], **Ludwig Lautenbacher**[1], **Zonera Hassan**[16], **Yun-Chien Chang**[1], **Chiara Falcomatà**[4,5,6,7], **Florian P. Bayer**[1], **Stefanie Bärthel**[4,5,6,7], **Tobias Schmidt**[1], **Roland Rad**[4,6,7,15], **Stephanie E. Combs**[8,9,10,11], **Matthew The**[1], **Frank Johannes**[12,13], **Dieter Saur**[4,5,6,7], **Martin Hrabe de Angelis**[14,17,18], **Mathias Wilhelm**[1,19], **Günter Schneider**[16,20], **Bernhard Kuster**[1,3,10,11,13,*]

[1]Chair of Proteomics and Bioanalytics, Technical University of Munich, Freising, Germany

[2]College of Life Science, Northwest University, Xi'an, P.R. China

[3]Bavarian Biomolecular Mass Spectrometry Center, Technical University of Munich, Freising, Germany

[4]Division of Translational Cancer Research, German Cancer Research Center (DKFZ) and German Cancer Consortium (DKTK), Heidelberg, Germany

[5]Chair of Translational Cancer Research and Institute for Experimental Cancer Therapy, Klinikum rechts der Isar, School of Medicine, Technical University of Munich, Munich, Germany

[6]Department of Internal Medicine II, Klinikum rechts der Isar, Technical University of Munich, Munich, Germany

[7]Center for Translational Cancer Research (TranslaTUM), School of Medicine, Technical University of Munich, Munich, Germany

[8]Department of Radiation Oncology, Klinikum rechts der Isar, Technical University of Munich, Munich, Germany

[9]Institute of Radiation Medicine, Department of Radiation Sciences, Helmholtz Zentrum München, Neuherberg, Germany

[*]Correspondence should be addressed to: kuster@tum.de.

Author contributions

P.G. performed (phosho)proteomic and transcriptomic experiments on tissue samples; Y.B., J.R, J.K., C.-Y.L., Y.-C.C., F.P.B. performed (phosho)proteomic experiments on mPDAC cell lines; P.G., P.S., Y.B., B.K. interpreted and visualized data; P.G., P.S., Y.B., C.M. generated web resource databases; P.G., P.S., Y.B. M.F., C.M., R.R.H. performed data analysis; L.L., T.S., M.T. contributed analysis tools; A.C., H.J., S.M., N.W., Z.H., C.F., S.B. isolated mPDAC cell cultures and performed the drug screen; J.C.-W., M.H.d.A. selected and provided mouse tissue samples; S.D. performed the radioresistance assay; R.R., S.E.C., F.J., D.S., M.W., G.S., B.K. supervised the study; P.G. and B.K. conceptualized the project and wrote the manuscript. All authors edited the manuscript.

Competing interests

M.W. and B.K. are founders and shareholders of OmicScouts GmbH and MSAID GmbH. They have no operational role in either company. M.F. is founder, shareholder and CEO of MSAID GmbH. T.S. is founder and shareholder of MSAID GmbH. The contents of this study are unrelated to any commercial activities. The remaining authors declare no competing interests.

[10]German Cancer Consortium (DKTK), Munich, Germany

[11]German Cancer Research Center (DKFZ), Heidelberg, Germany

[12]Population epigenetics and epigenomics, Technical University of Munich, Freising, Germany

[13]Institute of Advanced Study (IAS), Technical University of Munich, Freising, Germany

[14]Institute of Experimental Genetics, German Mouse Clinic, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany

[15]Institute of Molecular Oncology and Functional Genomics, School of Medicine, Technical University of Munich, Munich, Germany

[16]Medical Clinic and Policlinic II, Klinikum rechts der Isar, Technical University of Munich, Munich, Germany

[17]Chair of Experimental Genetics, TUM School of Life Sciences, Technical University of Munich, Freising, Germany

[18]German Center for Diabetes Research (DZD), Neuherberg, Germany

[19]Computational Mass Spectrometry, Technical University of Munich, Freising, Germany

[20]University Medical Center Göttingen, Department of General, Visceral and Pediatric Surgery, Göttingen, Germany

[#] These authors contributed equally to this work.

Abstract

The laboratory mouse ranks among the most important experimental systems for biomedical research and molecular reference maps of such models are essential informational tools. Here, we present a quantitative draft of the mouse proteome and phosphoproteome constructed from 41 healthy tissues and several lines of analyses exemplify which insights can be gleaned from the data. For instance, tissue- and cell-type resolved profiles provide protein evidence for the expression of 17,000 genes, thousands of isoforms and 50,000 phosphorylation sites *in-vivo*. Proteogenomic comparison of mouse, human and Arabidopsis reveal common and distinct mechanisms of gene expression regulation and, despite many similarities, numerous differentially abundant orthologs that likely serve species-specific functions. We leverage the mouse proteome by integrating phenotypic drug (n>400) and radiation response data with the proteomes of 66 pancreatic ductal adenocarcinoma (PDAC) cell lines to reveal molecular markers for sensitivity and resistance. This unique atlas complements other molecular resources for the mouse and can be explored online via ProteomicsDB and PACiFIC.

Keywords

Mouse; proteome; phosphoproteome; transcriptome; pharmacoproteomics

The mouse is arguably among the most important mammalian model systems for basic, translational and biomedical research[1]. The initial sequencing of the mouse genome in 2002 as well as systematic transcriptome analyses of mouse tissues in 2004 marked important milestones as these provided important informational tools to better understand the human genome and human (patho-)physiology[2,3]. As most biological processes are controlled by proteins and their post-translational modifications (PTMs), adding a proteomic dimension is a logical consequence. An initial analysis of 28 mouse tissues provided evidence for ~7,300 proteins and an earlier phosphoproteomic investigation identified ~36,000 phosphorylation sites (p-sites) in 9 tissues[4,5]. Here, we substantially extend these efforts by analyzing the proteomes and phosphoproteomes of 41 adult tissues of C57BL/6N mice covering 15 major anatomical systems and 66 *Kras*^{G12D} pancreatic ductal adenocarcinoma (PDAC) cell lines using a quantitative mass spectrometry (MS)-based approach (**Fig. 1a**).

Results

A high-quality map of the mouse proteome

Following state-of-the-art protocols for protein identification and error propagation control (see Methods), we identified 17,883 proteins (17,771 and 12,971 from tissues and cell lines, respectively) representing 16,995 of the 22,437 protein-coding genes annotated in UniProtKB (Fig. 1b, Supplementary Table 1). The phosphoproteomes (Fig. 1c, d, Supplementary Table 2) covered 40,562 and 22,113 p-sites for tissues and cell lines respectively. Re-analysis of the aforementioned previously published studies showed that, albeit still incomplete, our current draft more than doubles the coverage of both proteomic levels and encompasses nearly all of the previous data (Supplementary Fig. 1a, b). The high quality of the proteomic data is underscored by several lines of evidence: about 99% of all proteins are backed by peptide spectra that are highly similar (spectral angle, SA > 0.7) to the ones predicted by the deep learning method Prosit[6] (Supplementary Fig. 1c). Nearly 95% of all proteins are supported by at least two peptides and the median sequence coverage of proteins is 36% (Supplementary Fig. 1d). When tightening the FDR threshold from 1% to 0.1%, >80% of the proteins are retained and, at an arbitrary but high score cutoff of the search engine Andromeda of 100, less than 25% of the confidently localized p-sites are lost (Supplementary Fig. 1e). Replicate analysis showed that both the proteome and phosphoproteome measurements were generally qualitatively and quantitatively well reproducible and that differences between tissues are far larger than between replicates of the same tissue (Supplementary Fig. 2a-c).

As observed in recent maps of the human and Arabidopsis proteomes[7,8], protein and phosphorylation levels varied greatly between tissues and cell lines. The core ~5,400 proteins detected in all tissues show common expression patterns, which account for ~80% of the total protein abundance in each tissue and are functionally enriched for proteins involved in key physiological processes of any cell (Fig. 1c, d and Supplementary Fig. 3a, b). The heterogeneity of the phosphoproteome was even more pronounced such that only 2.5% of all p-sites were detected in all tissues (2.7% in all cell lines). Challenges in correctly localizing p-sites may have contributed to this low apparent consistency, but the bulk of the phosphorylation clearly reflects differences in the complex and dynamic phosphorylation signaling processes in the different tissues.

To facilitate the use of the molecular atlas by the scientific community, all data have been deployed to ProteomicsDB (https://www.proteomicsdb.org)[7]. Among many features, researchers can review, compare, validate, and download tandem mass spectra, construct or predict spectral libraries, visualize gene expression at single- or multi-omics level, evaluate the extent of PTMs on a protein of interest, model and estimate drug and radiosensitivity based on (phospho)proteome abundance or cross-interrogate hypothesis between the mouse and human proteomes. In addition, all data collected for the mouse PDAC (mPDAC) cell lines can be explored using the interactive web application PACiFIC (http://pacific.proteomics.wzw.tum.de), which allows users to interrogate the results of the phenotypic screening data, enabling the identification of protein markers of sensitivity or resistance (Fig. 1e).

Proteomic annotation of the mouse genome

Estimates for the number of protein-coding genes of the mouse genome have varied over time but currently stand at 22,000-24,000[9,10]. At the time of writing, the UniProt *Mus musculus* complete proteome set lacked experimental evidence at the protein level for 7,999 genes (~36%). Our analysis confirmed expression of 13,538 genes with prior evidence at the protein level (PE1 definition of UniProt) and identified protein products for an additional 3,457 protein-coding genes that had not yet been observed as proteins (evidence levels PE2-5; Fig. 2a). The data quality for these proteins was identical to that of PE1 proteins (Supplementary Fig. 4a-c). Exemplified by the putative adenosylhomocysteinase 3 protein (Fig. 2b, Supplementary Fig. 4d), 1,663 splice variants of 776 genes were identified from the proteomic data. By applying rigorous FDR control and computational validation (see Methods), expression of 712 distinct translation products of small open reading frames (sORFs) identified by RIBO-seq (sORFs.org) were observed as proteins. Most of these candidate sORFs were located near or within the coding region of a gene and all have an AUG or a near-cognate initiation codon. Interestingly, the detected sORF-encoded peptides (SEPs), span the entire abundance range of all detected peptides and they are translated in multiple tissues or cell lines. This may suggest that sORFs are canonical components of the mouse proteome but these need to be validated independently (Fig. 2c-e, Supplementary Table 3).

We further explored the data for unexpected proteoforms such as mutations or PTMs. Using an open-search strategy[11,12] (see Methods), we identified widespread mass deviations of peptides representing otherwise canonical protein sequences (Supplementary Fig. 5a). Many can be attributed to chemical artifacts occurring during sample preparation such as oxidation or deamidation but others are more likely to occur in cells as a result of enzymatic activity or chemical reactions of activated metabolites. We particularly investigated mass shifts possibly representing amino acid substitutions (missense mutations, Supplementary Fig. 5b). Using a series of stringent filters, as well as peptide spectra validation (see Methods), 2,418 novel candidate protein variants could be identified from the proteomic data. We note that these remain to be validated by additional means (Supplementary Fig. 5c, d, Supplementary Table 3). Future investigations can be envisaged that make use of this data *e.g.* for refining gene models or identifying new genes not present in the sequence database used in this study.

Protein expression and phosphorylation characteristics of tissues and cell lines

The phosphoproteomic data shows that more than half of all mouse proteins are phosphorylated at least once. Protein expression and phosphorylation levels show broadly similar abundance characteristics in that 90% of all proteins and p-sites are within a range of 25x around the median abundance level (Fig. 3a). A relatively small fraction of proteins makes up 90% of the total abundance, an observation that is more pronounced in tissues than in cell lines, and stronger in the proteome than the phosphoproteome (Fig. 3b). This suggests that the bulk of the phosphoproteome in a tissue is relatively stable and may contribute to the spatial and structural organization of the cell. Given that cancer cells in culture undergo cell division much more frequently than cells in tissues, it may not be surprising that p-proteomes of the cell lines appeared more dynamic.

We and others have previously observed that most proteins are expressed in most tissues but often in vastly different quantities[13] and that the latter is an important feature of the functional characteristics of a (differentiated) cell. This is also the case for the mouse proteome. Only relatively few proteins (< 4% for tissues, and < 7% for cell lines) and p-sites (< 9% for tissues, and < 6% for cell lines) were detected in a tissue- or cell-specific manner (Fig. 1c-d). Consequently, hierarchical clustering of proteome expression showed that all PDAC cell lines form one large cluster which locates close to the cluster containing the pancreas (Fig. 3c). Neuronal tissues clearly segregated from the rest but anatomically adjacent or functionally related organs clustered more closely than organs of distinct function such as the immune or female reproductive system. The functional specialization of a tissue was also clearly apparent from the list of highly abundant proteins within a tissue (Fig. 3d). For instance, protein expression in the frontal lobe (FRL) is dominated by synucleins (Snca, Sncb), the major myelin protein Plp1 and the synaptic proteins Syt1 and Syn1. The tongue proteome is dominated by proteins involved in muscle contraction and five of the 10 most abundant proteins in the eye are crystallins, essential components of the visual system.

Evolutionary conservation of factors regulating protein abundance

The amount of protein in a cell is controlled by a multitude of factors operating at the transcriptional, translational and post-translational level. Several insights have been obtained by integrating (paired) transcriptomic and proteomic data collected for tissues of the same species [8,14–16]. Here, we asked to what extent molecular determinants of protein abundance regulation are conserved between species. We collected RNA-Seq data (n=29) from the exact same tissues that underwent proteome analysis, leading to the quantification of 21,261 transcripts (genes) that were expressed in at least one tissue (Supplementary Fig. 6a, Supplementary Table 4). As observed before, protein and transcript abundance are positively correlated (Pearson correlation r>0.5 for all tissues) and with a nearly quadratic relationship indicating that high-abundant transcripts are generally more efficiently translated and/or more stable than low-abundant transcripts (Supplementary Fig. 6b, Supplementary Fig. 7, Supplementary Table 5). We also observed limited similarity in the abundance distributions of the proteomes and transcriptomes of the tissues. Proteins are more evenly distributed across the detected abundance range than transcripts, implying the contribution of factors other than mRNA levels for controlling protein abundance (Supplementary Fig. 6c).

We next investigated if molecular features previously identified for human and Arabidopsis[8,16] have predictive power for explaining protein-level variations in the mouse (Supplementary Fig. 6d). Indeed, for all three species, the analysis revealed that the largest proportion of variation in protein abundance within tissues can be explained by mRNA levels, followed by codon usage, protein-protein interactions and mRNA sequence motifs (Fig. 4a). Interestingly, codon usage appeared to be of a higher relative importance in Arabidopsis than in mouse and human and, in fact, many more codons from Arabidopsis were selected by the model than for the two mammalian species (Supplementary Fig. 6e). Similarly, mRNA motifs located in the 5'-UTR, coding region and 3'-UTR differed substantially between species indicating that regulatory mRNA sequence elements are used with variable preferences (Supplementary Fig. 6f, g). We also included phosphorylation information in building the models as this PTM has been associated with both increasing and decreasing levels of specific proteins. The analysis showed that the phosphorylation status of a protein is indeed an independent predictor of its abundance, yet, the overall contribution at steady state levels is rather small (Supplementary Fig. 6h).

Conserved and divergent features of the mouse and human proteomes

Because the mouse is an important model for human (patho-)physiology, it is particularly interesting to investigate how the proteomes of the two species are similar or different. To this end, we correlated the expression of the 10,869 orthologous protein pairs contained in this study with a human proteome recently published by the authors[17] (representing 65% of all human:mouse orthologs, Fig 4b, Supplementary Table 6). Overall, we observed a high degree of correspondence (Pearson R=0.75; Fig. 4c and Supplementary Fig. 8a, b). However, PCA analysis of protein abundance revealed a strong separation of species rather than tissues (Fig. 4d). This separation is driven by a set of 954 proteins that we define as species-variable orthologs (SVOs; see Methods) and whose abundances vary more between species than tissues (**Fig. 5e**, Supplementary Fig. 8c).

Based on transcriptome data, the mouse ENCODE Project[18] concluded that most orthologous genes show conserved co-expression between human and mouse. Neighborhood analysis of conserved co-expression (NACC) showed that this is also the case at the proteome level (Fig. 4f and Supplementary Fig. 9). Despite the general conservation of proteome expression, examples for extensive differences can be observed particularly for certain tissues (Fig. 4g). Gene ontology analysis showed that these differences can often be attributed to specific biology. For example, in the heart, many differentially abundant proteins are involved in the regulation of heart rate, possibly reflecting the 10x faster heart rate of mice compared to humans (Fig. 4h and Supplementary Fig. 10a). For the liver (Fig. 4i and Supplementary Fig. 10b), most of the strongly differentially abundant genes are cytochrome P450 enzymes (CYPs), ATP-binding cassette transporters (ABCs) and solute carrier transporters (SLCs). A few of these genes (i.e. Abca1, Abca6, Cyp1a1, Cyp2e1, and Cyp7a1) have been reported to oscillate in terms of abundance[19]. It is, therefore, possible that the source of the differences between mouse and human may be in part the result of an unsynchronous circadian rhythm at the time of the livers sampling. The differentially expressed protein families are of considerable pharmacological interest, due to their roles in drug metabolism. The large expression differences between mouse and human may, therefore, reflect the often vast differences in the pharmacokinetics, pharmacodynamics and ADME/TOX characteristics of therapeutic drugs in these two species. If not taken into account, this can limit the ability to translate results of drug studies obtained in mice to the human situation[20].

Phospho(proteome) activity landscapes reveal markers of radiation response

We and others have previously shown that the proteomic and phosphoproteomic landscapes of cancer cell lines can often explain their sensitivity or resistance to drugs based on single or groups of proteins[21–23]. Despite limitations, murine cell lines are still important tools for research and drug discovery, particularly for diseases such as pancreatic ductal adenocarcinoma (PDAC) for which only few human cell lines exist. We, therefore, extended the project to the (phospho)proteomes of 66 PDAC cell lines and measured their sensitivity to 5 doses of ionizing radiation (Fig. 5a, Supplementary Table 7).

Using machine learning (*i.e.* combining a recursive partitioning tree for feature selection and a random approach forest for predicative modelling; see Methods), we integrated the quantitative radiation and proteomic data to select 20 proteins that are associated with radiation sensitivity or resistance (Fig. 5b). The strongest sensitivity marker was Lrrfip1 (P=1.2e-9, R=-0.66, n=66, Fig. 5c), a cytosolic nucleic acid sensor that mediates the production of type I interferon, which is known to be involved in radiation response[24,25]. Conversely, high expression of Actl6a, a member of the BAF chromatin remodeling complex was strongly associated with radiation resistance (P=1.7e-7, R=0.59, n=66) confirming recent observations[26]. Training a classifier based on these 20 proteins showed reasonable accuracy in predicting radiation response (R=0.74, 0.75 and 0.66 for training, validation and hold-out data, respectively; Fig. 5b inset). Taking the same approach for the phosphorylation data also identified candidate markers that showed reasonable prediction accuracy (R=0.80, 0.81 and 0.81 for training, validation and hold-out data, respectively; Fig. 5d). While many of these p-sites had been detected before, their functions are generally not understood but can now be placed in the functional context of radiation response. For example Tcof1 pSer1227 (P=3.62e-8, R=0.62, n=66) is associated with radiation resistance. Following DNA damage, it has been shown that the nucleolar localization of NBS1[27], a major player in DNA repair, is triggered in a Tcof1 phosphorylation-dependent manner, with both CK2 and ATM kinases responsible of the phosphorylation of Tcof1[28]. Consistently, silencing of TCOF1 expression radiosensitized rat progenitor cells[29], while phosphorylation levels of Tcof1 disproportionately increased in radioresistant cells. Our analysis instead suggests that high levels of Tcof1 pSer1227 could be an indicator of low resistance to radiation treatment.

Based on the hypothesis that the response to radiation could associate proteins or p-sites involved in the same molecular pathway, we submitted all statistically highly significant candidates (R>0.5 or <-0,5 and P<0.0001) to STRING analysis[30]. One of the top hits was mitochondrial respiration and thermogenesis in which high protein expression indicated radiation resistance (Supplementary Fig. 11a). At the level of phosphorylation, a strong STRING network was detected for proteins involved in cellular response to DNA damage and chromatin modification (Fig. 5e). Among these are p-sites of Brca1 (pSer1149, P=1.4e-5, R=0.78, n=22; pSer1152, P=1.2e-4, R=0.54, n=44; pSer1154, P=2.3e-4, R=0.52, n=45), the latter two are known substrates of ATM and CDK1[31–33]. As all three sites are in close vicinity, we hypothesize that pSer1149 might also play a role in DNA damage response.

Phospho(proteome) activity landscapes reveal markers of drug response

Integrating the (phospho)proteomic and cell viability data of 36 cell lines in response to 407 drugs (Fig. 5a, Supplementary Table 7) gave rise to a number of interesting molecular pharmacodynamics biomarkers. For instance, elastic net regression (Fig. 5f and Supplementary Fig. 11b) revealed that high Sirt6 protein abundance is a sensitivity marker for the MAP2K1/2 inhibitors Trametinib (P=6.1e-8, R=-0.77, n=35; Supplementary Fig. 11c) and Combimetinib (P=8.4e-5, R=-0.62, n= 35; Supplementary Fig. 11d). While this has been noted before[34], our analysis places OTS514 (P=6.1e-4, R=-0.55, n=35; Supplementary Fig. 11e), an inhibitor of the MAPKK-like kinase TOPK, and its target into the same functional context[35,36]. Another example is Shroom2, whose abundance is correlated with KX2-391 sensitivity (P=1.3e-5, R=-0.70, n=31; Supplementary Fig. 11f), a non-ATP-competitive inhibitor of Src kinase and tubulin polymerization. Shroom2 is a key mediator of the RhoA–ROCK pathway that regulates cell motility and actin cytoskeleton organization[37] as well as regulating the accumulation of γ-tubulin and cell morphology[38]. Interestingly, our analysis showed that Shroom2 is also a sensitivity marker of three further drugs targeting tubulin polymerization (Colchicine, Plinabulin, Vincristine; Fig. 5g). Shroom2 expression also places Rigosertib (a PLK1 inhibitor) into the same group, lending support to prior suggestions that Rigosertib also acts as a (direct or indirect) tubulin polymerization inhibitor[39,40]. Candidate drug response markers were also identified at the phosphorylation level. For instance, elastic net regression revealed that the abundance of Mical2 pSer515 is a resistance marker of several drugs targeting proteins involved in the regulation of the cell cycle and the DNA damage response. Examples include the pan-CDK inhibitor AT7519 (P=1.2e-5, R=0.66, n=36), the CHEK1 inhibitor Rabusertib (P=4.2e-5, R=0.63, n=36) or the ATR inhibitor AZD6738 (P=9.6e-6, R=0.67, n=36; Fig. 5h). Mical2 is a nuclear monooxygenase that regulates transcription via depolymerization of nuclear actin which makes the above associations functionally plausible even though no function for pSer515 has been reported yet.

Discussion

Many further uses of the transcriptomic, proteomic, phosphoproteomic, as well as phenotypic radiation and drug response data reported in this study can be envisaged. Full availability of the underlying data as well as many web-based analysis tools in ProteomicsDB and PACiFIC should assist scientists in future work aiming at elucidating molecular determinants governing biological processes in the mouse. In turn, this will also lead to a better understanding of the mouse as a model for human pathology and drug discovery.

Methods

Tissue preparation

C57BL/6N mice were maintained in IVC cages with water and standard mouse chow according to the directive 2010/63/EU, German laws and GMC housing conditions (www.mouseclinic.de). Mice (16 weeks old, m/f) were euthanized with CO_2. Afterward, animal were perfused with 50 mL of precooled phosphate buffered saline, supplemented with cOmplete™ EDTA-free protease inhibitor cocktail (Roche) and a mixture of phosphatase inhibitor cocktail 1, 2 and 3 (Sigma-Aldrich). A total of 41 tissues and organs were dissected, snap-frozen in liquid nitrogen and stored at -80 °C.

Cell lines preparation

Primary murine PDAC cell lines, consisting of 66 low-passaged lines with various genetic backgrounds, were obtained from genetically engineered mice[41].

Cells were cultured in Dulbecco's Modified Eagle's Medium (DMEM) high glucose medium (Sigma) supplemented with 10% fetal calf serum (FCS), 100 U/ml penicillin and 100 μg/ml streptomycin (Invitrogen) at 37 °C in a humidified 5% CO_2 atmosphere.

For proteomics analysis cells pellets were snap-frozen in liquid nitrogen and stored at -80 °C until further use.

Radioresistance assay

Radiation was delivered at 220 kV and 15 mA with a dose rate of 0.90 Gy/min using the RS225A irradiation device (Gulmay/Xstrahl). Cell lines were screened for radiation response using the AlamarBlue proliferation assay (Thermo Fisher Scientific). Cells were either irradiated with 0 Gy, 2 Gy, 4 Gy, 8 Gy or 16 Gy 24 h after seeding. The AlamarBlue reagent was added 72 h after irradiation. After an incubation time of four hours at 37 °C in a humidified 5% CO_2 atmosphere, proliferation of cells was measured by absorbance at 570 nm and 630 nm using the ELx808 microplate reader (BioTek). Data were recorded by the Gen5 Software (BioTek).

Automated drug screen

All cell lines used for automated drug screening were routinely tested for mycoplasma contamination by PCR. Cells were cultured in DMEM high glucose (Sigma) supplemented with 10% FCS (Biochrom GmbH) and 1% Penicillin/Streptomycin (Thermo Fisher Scientific) and were digested to a single cell suspension using 1 x TrypZean® Solution (Sigma Aldrich). 1000-2000 cells per well (depending on growth rate) were seeded in 96-well plates (Corning Life Sciences) using a Multidrop™ Combi Reagent Dispenser (Thermo Fisher Scientific). After overnight incubation at 37 °C and in 5% CO_2 in a Cytomat™ 24C automated incubator (Thermo Fisher Scientific), cells were treated with the compound library using a CyBio® FeliX pipetting platform (Analytik Jena). All compounds were obtained from SelleckChem dissolved either in DMSO or water. Each drug was added in 7 concentrations (3-fold dilution series, highest concentration 10 μM). Cell viability was measured after 72 hours using CellTiter-Glo® Luminescent Cell Viability Assay (Promega).

The reagent was added using a Multidrop™ Combi Reagent Dispenser. After incubation for 10 minutes luminescence was measured in an Infinite 200 PRO microplate reader (Tecan).

Protein lysis and digest

Frozen tissues were pulverized on liquid nitrogen and resuspended in ice-cold 1% (w/v) sodium deoxycholate (SDC), 50 mM Tris-HCl, pH 8.5 buffer and subsequently lysed by 5 min boiling at 95 °C. Further homogenization was achieved using ceramic beads and an automated homogenizer (Precellys 24, Bertin) with 2 x 10 second pulses at 5000 rpm. Homogenates were clarified by centrifugation for 10 minutes at 10,000 *g* at 4 °C and protein concentration was determined by the Bradford method.

One mg of protein material was reduced with 10 mM DTT for 40 min at 37 °C and subsequently alkylated with 55 mM chloroacetamide (CAA) for 30 min at 25 °C in the dark. The solution was then diluted to a final urea concentration of 0.1% (w/v) SDC with 50 mM Tris-HCl, pH 8.0, and proteins were digested in two steps (4 h and overnight) at 37 °C with trypsin (Roche). Samples were acidified with formic acid (FA) at 1% (v/v) final concentration and centrifuged at 20,000 g for 10 min at 4°C.

The supernatant was desalted using Sep-Pak C18 cartridges (Waters), dried *in vacuo*, and stored at -80 °C until further use.

The mPDAC cells were resuspended in 2% (w/v) sodium dodecyl sulfate (SDS), 40 mM Tris-HCl, pH 7.6 buffer and subsequently lysed by heating to 95 °C for 5 min, followed by the addition of trifluoroacetic acid (TFA) to a final concentration of 1% (v/v). Subsequently N-Methylmorpholine was added to a final concentration of 2% (v/v) to adjust the pH value to ~7.5, and protein concentration was determined by the BCA method.

Protein digestion was performed via an automated SP3-based method by using the AssayMAP Bravo Platform (Agilent Technologies). For each cell lines, 100 µg of protein material in 150 µL were transferred to a 96 deep well plate, then 10 µL of SP3 beads (1:1 mix of Sera-Mag Speed Beads A and B, Fisher Scientific) were added to the lysate, and finally ethanol was added to a final concentration of 70% (v/v). The plate was incubated in a thermal shaker for 5 min at 1200 rpm, and then transferred to the magnet rack to immobilize the SP3 beads for 3 min. The supernatant was discarded, and the SP3 beads were washed 3 times with 200 µL of 80% (v/v) ethanol/water and once with 200 µL of ACN. The bound proteins were reduced by adding 100 µL of 10 mM DTT in 100 mM Tris-HCl, pH 8.5 and incubating the plate in a thermal shaker at 37 °C and 1000 rpm for 1 h. Subsequently, proteins were alkylated with 55 mM CAA for 1 h at 37 °C in the dark. Finally, 3 µg of trypsin were added to each well, and the plate was incubated overnight at 37 °C in the thermal shaker at 1000 rpm. After protein digestion, samples were acidified with FA at final concentration of 1% (v/v), desalted using the Sep-Pak tC18 96-well plate (Waters, 25 mg), dried *in vacuo*, and stored at -80 °C until further use.

Phosphopeptide enrichment and off-line fractionation

For the tissues samples, phosphopeptide enrichment was performed by loading peptides on a Fe^{3+}-IMAC column (Propac IMAC-10 4 × 50 mm column Thermo Fisher Scientific) as previously described. The enrichment was performed with buffer A (0.07% TFA, 30% acetonitrile) as loading buffer and buffer B (0.3% NH_4OH) as elution buffer. Using a UV-abs signal, the outlet of the column was monitored and two fractions were collected: one containing non-phosphorylated peptides (flow-through), the other containing phosphorylated peptide. Both factions were dried in a vacuum centrifuge and stored at -80 °C. Phosphopeptides were further fractionated into four fractions using the high pH RP fractionation[42] using self-packed StageTips, containing 5 disks of C18 material (3M Empore).

For whole proteome analysis, peptides contained in the IMAC flow-thought fraction were re-suspended in 10 mM ammonium acetate, pH 4.7, and subjected to trimodal mixed mode chromatography on an Acclaim Trinity P1 2.1 × 150 mm, 3 μm column (Thermo Fisher Scientific) for peptide fractionation[43]. A total of 32 fractions were collected, dried *in vacuo* and stored at -20 °C until LC-MS/MS analysis.

For the PDAC cells, phosphopeptides were enriched by using 5 μL Fe(III)-NTA cartridge (Agilent technologies) in an automated fashion on the AssayMAP Bravo Platform. The cartridges were primed with 200 μL of washing buffer 1 (50% ACN/0.1% TFA), and equilibrated with 200 μL of washing buffer 2 (80% ACN/0.1% TFA). The desalted peptides were dissolved in 100 μL of loading buffer (80% ACN/0.1% TFA), and loaded onto the cartridges with a flow rate of 5 μL/min, the flow-through was collected. The cartridges were washed with 50 μL of washing buffer 2, which was collected and pooled with the first flow-through. The bound phosphopeptides were eluted with 50 μL of 1% ammonia (pH 11). The eluate containing phosphorylated peptide and the flow-thought containing non-phosphorylated peptides were dried *in vacuo*, and stored at -80 °C until further use.

For whole proteome analysis, peptides contained in the Fe(III)-NTA flow-through fraction were re-suspended in 100 μL of buffer A (25 mM ammonium formate, pH 10), and subjected to high pH RP fractionation with the AssayMAP Bravo platform using RP-S cartridges (Agilent). The cartridges were primed sequentially with 150 μL of isopropanol, ACN and solvent B (80% ACN in 10 mM ammonium formate, pH 10.0), at a flow rate of 50 μL/min. The cartridges were equilibrated with 100 μL of buffer A at a flow rate of 10 μL/min, peptides were then loaded at 5 μL/min and the follow-through (FT) was collected. The cartridges were further washed with 50 μL of solvent A, and the flow-through was collected and pooled with the previous one. The peptides were eluted with 25 mM ammonium formate, pH 10, using increasing concentrations of ACN (5%, 10%, 15%, 20%, 25%, 30%, and 80%). Finally, the seven fractions with flow-through were combined into four fractions (5% + 25%, 10% + 30%, 15% + 80%, and 20% + FT), the peptides were dried *in vacuo*, and stored at -80 °C until further use.

LC-MS/MS analysis

Nano flow LC-MS/MS measurements were performed using a Dionex Ultimate 3000 UHPLC+ system coupled to a Q Exactive Orbitrap HF (Thermo Fisher Scientific) for tissues label-free. Peptides were delivered to a trap column (75 μm i.d. × 2 cm, packed in-house with 5 μm Reprosil C18 resin; Dr. Maisch) and washed using 0.1% FA at a flow rate of 5 μL/min for 10 min. Subsequently, peptides were transferred to an analytical column (75 μm i.d. × 45 cm, packed in-house with 3 μm Reprosil C18 resin, Dr. Maisch) at a flow rate of 300 nL/min. Peptides were chromatographically separated using a linear gradient of solvent B (0.1% FA, 5% DMSO in ACN) and solvent A (0.1% FA, 5% DMSO in water). Linear gradients were as follows: from 4% to 32% of B in 82 min for full proteome analysis, and from 4% to 27 % of B in 80 min and a two-step increase for phosphoproteome analysis. The total measurement time for each sample was 90 min.

The full proteome analysis of the PDAC cells was performed on a micro-flow LC-MS/MS system using a modified Vanquish pump (Thermo Fisher Scientific) coupled to a Q Exactive Orbitrap HF-X (Thermo Fisher Scientific) mass spectrometer. Chromatographic separation was performed via direct injection on a 15 cm Acclaim PepMap 100 C18 column (2 μm, 1 mm ID, Thermo Fisher Scientific) at a flow rate of 50 μL/min, using a 60 min linear gradient (3% to 28%) of solvent B (0.1% FA, 3% DMSO in ACN) and solvent A (0.1% FA, 3% DMSO in water). The total measurement time for each sample was 65 min.

For the PDAC phosphoproteome analysis, the Dionex Ultimate 3000 UHPLC+ system described above, coupled to an Orbitrap Exploris 480 (with a FAIMS device, Thermo Fisher Scientific) mass spectrometer was used.

Phosphopeptides were chromatographically separated using a two-step increase gradient of solvent B (0.1% FA, 5% DMSO in ACN) and solvent A (0.1% FA, 5% DMSO in water), from 4% to15% of B in 60 min, followed by 15% to 27% of B in 34 min. The total measurement time for each sample was 120 min.

The mass spectrometers were operated in a data-dependent acquisition (DDA) to automatically switch between MS and MS/MS.

For label-free samples full scan MS spectra were recorded in the Orbitrap from 360 to 1,300 m/z at a resolution of 60,000 using an automatic gain control (AGC) target value of 3e6 charges and maximum injection time (maxIT) of 10 ms (50 ms for HF-X and 45 ms for Exploris 480). After the survey scan, the 20 most intense precursors were isolated with an isolation window of 1.7 m/z (1.3 m/z for HFX and 1.2 m/z for Exploris 480) for HCD fragmentation, and fragments ions were recorded in the Orbitrap at a resolution of 15,000 (30,000 for Exploris 480), AGC of 1e5 and maxIT of 25 ms (22 for HF-X and 54 ms for Exploris 480). For phosphoproteome analysis of tissues samples, the 15 most intense precursor were selected instead, while AGC and maxIT were set to 2e5 and 50ms, respectively. For the Exploris 480, the data dependent mode was set to cycle time, the cycle time was set to 0.8 s, and an internal compensate voltages (CV) stepping experiments with CV values of -35 V, -45 V and -60 V was applied.

Normalized collision energy (NCE) was set to 25% for the HF and 28% for the HF-X and Exploris 480. For all methods, charge state screening was enabled and only precursors with charge states between 2 and 6 were selected for fragmentation. Dynamic exclusion was set to 25 s and27 s for the measurements on the HF and HF-X, respectively. For the Exploris 480, the precursors including isotopes were excluded if they appeared twice within 90 s.

Identification and quantitation of peptides and proteins

Raw mass spectrometry data (1,970 runs) were processed using MaxQuant[44] (version 1.5.3.8), using default settings with the following minor changes: oxidized methionine (M) and acetylation (protein N-term) were selected as variable modifications, carbamidomethyl (C) as fixed modification, trypsin was selected as the proteolytic enzyme, and up to two missed cleavages were allowed, minimum peptide length of seven amino acids. Searches were performed against the *Mus musculus* UniProt FASTA database (UP000000589, March 2018, 61,307 entries), containing canonical and isoform sequences, as well as against a common contaminants database.

The false-discovery rate (FDR) was set to 1, meaning 100%, to allow for a subsequent PSMs rescoring via Prosit and Percolator[6], implemented in ProteomicsDB. After rescoring, PSMs and peptides were filtered at an FDR of 1% (calculated on sample level) and protein-level and gene-level FDR estimates were calculated using the picked target-decoy approach[45] (on sample and whole dataset levels). For the protein identification and quantification ProteomicsDB used only protein or gene specific peptides with a q-value lower or equal to 0.01. Consequently, unambiguously identified proteins with a q-value lower or equal to 0.01 formed their own protein group, while for the rest of the proteins, which cannot be unambiguously identified by unique peptides were grouped in one gene group and quantified together, only if the corresponding gene q-value was lower or equal to 0.01. Within a gene group, the protein with the most identified peptides was selected as representative of the group.

Intensity-based absolute quantification (iBAQ)[46] was used as protein abundance estimator, and calculated as previously described. Unless otherwise stated, iBAQ values were normalized based on the total sum of all protein intensities, log10 transformed and right-shifted by 10 log10 units into positive numerical space. Further bioinformatics analysis was performed with Perseus[47] (version 1.6.5.0), Microsoft Excel (version 2013) and R (version 3.6.3)[48]. Protein and gene annotations were extracted from the UniProt, Ensembl, and MGI databases.

Visualization and comparison of peptide spectra was performed with the web-tool Universal Spectrum Explorer[49], using default settings. For the generation of mirror plots, fragment ions in the predicted spectrum were annotated only if they were not found in the experimental spectrum, and their intensity was higher than 10%.

Identification and quantitation of phosphorylation sites

Raw mass spectrometry data were processed with MaxQuant (version 1.6.0.13), as described above, with the following changes: phosphorylation on serine, threonine and tyrosine (Phospho STY) was selected as variable modification, and results were adjusted to 1 % FDR at protein, peptide, and site levels within the MaxQuant pipeline. A site localization probability of at least 0.75 (class I) was used as thresholds for the localization of phosphorylated residues.

Phosphorylation site intensities were normalized by the total sum approach, log10 transformed and right-shifted by 10 log10 units into positive numerical space.

Analysis of protein expression

Protein expression analysis including normalization, hierarchical cluster analysis and principal component analysis (PCA) were performed using R (version 3.6.3) on log transformed and normalized iBAQ intensities. For the hierarchical clustering euclidean distance and complete linkage were used.

Gene Ontology Analysis

Gene Ontology (GO) analysis of was performed either in Perseus (version 1.6.15.0) or with the R package topGO, using annotations from UniProt and MGI, respectively. Fisher exact test was performed with a FDR < 0.05. Only long lists of significantly enriched GO terms were further uploaded to REViGO[50] for summarization, using default parameters, restricting the database to *Mus musculus*.

Consolidation and extension of the mouse proteome

A great challenge in MS-based proteomics is the choice of an appropriate sequence database, which serves as reference for the peptides as proteins identification step. The UniProt is the most used protein sequence and functional annotation catalogue for human and mouse analysis[51], as it consists of manually annotated and reviewed entries. However, it is still not clear which of the mouse genes annotated in the database (22,437, at the time of writing) are translated into proteins, therefore each entry is assigned with a level that indicates the type of evidence that supports its existence (www.uniprot.org/help/protein-existence).

Because in shotgun proteomics experiments it is common to obtain inferred protein groups rather than unambiguously identified proteins, to re-annotate the UniProt existence level on the basis of the here reported identifications we defined the leading protein as the first protein within each gene group, as this would best represent the identification, being the one with the higher number of peptide identifications. All other proteins that were not selected as “leading”, were excluded from the analysis and retained the original UniProt level.

For sORF-encoded peptides (SEPs) identification, full proteome MS/MS spectra were searched and rescored as described above, against a custom database containing murine short open reading frame-encoded polypeptide sequences (downloaded from www.sorfs.org, September 2018, 503,779 entries) together with protein sequences from UniProt (UP000000589). After percolator, data were filtered for 1% FDR (PSMs and peptides level), peptides mapped to any annotated UniProt, Ensembl, or NCBI gene model were eliminated from the list, along with any candidate SEPs with an SA < 0.7. Furthermore, peptides that could be mapped to an annotated gene model by a single amino acid substitution, deletion or insertion were discarded.

The final list of peptides was then queried in the sORFs.org database to retrieve sequence annotations: location relative the transcript, biotype, and whether the sORF translation is already supported by mass spectrometry data.

Candidate SEPs were also synthetized at JPT Peptide Technologies using Fmoc-based SPOT synthesis on membranes and measured on the same LC-MS system that was used for the data acquisition of the tissue full proteome samples (see ‘LC–MS/MS analysis’ section). Experimental and synthetic peptide spectra were extracted from the raw files and used for similarity calculation without any prior spectrum processing. SA comparison between spectra of the tissue samples and synthetic peptides was performed using in-house R scripts, and a SA score of at least 0.7 was used as thresholds for candidate validation. For variant peptides identification, the full proteome dataset was additionally processed applying the so-called “open-search” strategy[11], overcoming the limitations of obtaining a comprehensive protein database for each sample from exome or RNA sequencing[52]. To this end, raw mass spectrometry data were converted to mzML format using msconvert in the ProteoWizard Toolkit[53]. Peak lists were searched with MSFragger[12] (version 2.4) using the same UniProt database as mentioned above, but in a concatenated forward-decoy version. Precursor mass tolerance was set to 500 Da and remaining settings were used at their default value. The search results were processed with PeptideProphet[54] and ProteinProphet[55] as part of the MSFragger suite, and identification were flittered to retain only hits with a probability > 0.9.

About 50% of all the peptide-spectrum matches (PSMs) were identified outside the -0.01 to 0.01 Δmass bin, representing the modified peptides. For ~4 million PSMs (~25% of the total identification hits) we could assign at least one mass shift reported in the Unimod database (www.unimod.org). We then specifically looked for mass shifts resembling an amino acid substitution (AAsub), with the exception for those involving lysine or arginine at the peptide c-terminus, in order to maintain trypsin specificity. Similarly, modification involving amino acids not present in the identified peptide sequence were discarded.

Because not all mass shifts can be unambiguously assigned to a single modification, either being translational, chemical or post-translational (e.g. Ala -> Ser, Oxidation, and Phe -> Tyr result all in the same shift of + 15.9949 Da), we excluded from the analysis amino acid substitution that could be explained by any of the Unimod annotated modifications (representing potential artefact or known modifications). We then defined as illogical, all those substitutions involving a non-alkylated cysteine residue, as well as those requiring more than 1 nucleotide substitutions.

To further validate the existence of a peptide variant, for any given PSMs, each potential modification was computed individually, and obtained sequences were subjected to spectra prediction by Prosit, using the same charge state detected for the endogenous peptide. To this end, spectra of endogenous peptides (i.e. those identified in the full proteome and phosphoproteome datasets) were compared against those obtained by Prosit. SA score was calculated using an in-house R script, and a SA score of at least 0.7 was used as thresholds for effective peptide identification.

AAsubs retaining a SA of at least 0.7 were further filtered to remove any peptide found to be identical and to overlap with part of annotated proteins in the UniProt, Ensembl, or NCBI mouse databases.

Because deamination of glutamine and asparagine is known to be a prevalent artefact during sample preparation, Prosit-validated substitutions to aspartic acid and glutamic acid whose non-deamidated peptide sequence was found to be identical and to overlap with part of an annotated protein were considered ambiguous and therefore discarded.

For all spectra comparisons, only those peaks which are either shared between spectra or exclusive to the predicted or synthetic peptide spectra were taken into account for the calculation of the normalized spectral contrast angle. Peaks exclusive to experimental spectra (tissues or PDACs samples) were ignored.

PACiFIC

PDACs raw mass spectrometry data were processed using MaxQuant (version 1.6.0.13), and results filtered for an FDR < 0.01 at the level of proteins, peptides and modifications, using the settings described above, but enabling the match between runs feature and phospho (STY) as variable modification. Full and phosphoproteome data were processed together as two separate groups. The two datasets were concatenated and used as predictors together. First, reverse and contaminates sequences in both datasets were excluded, then, unless otherwise stated, missing values in full proteome and phosphoproteome data were imputed using the protein-wise half-lowest method (analogous to the LOD2 method[56]) based on the rationale that the missing values in mass spectrometry experiments tend to accumulate at the lower end of the overall intensity distribution. Subsequently, the data were log10 transformed.

For each dose-response dataset (radiation or drugs), data were normalized into a range between 1 (no response or full viability) and 0 (full response or no viability). Afterwards, the classical symmetric four-parameter log-logistic model was fitted to each combination of drugs or radiation and cell lines in each dataset. We also calculated the standardized area under the dose-response curve (AUC) for each fitted model across the tested drug concentration or absorbed radiation range. Here, the AUC was defined as the area under the dose-response curve between zero and the Upper Bound or maximum predicted viability (ymax), divided by the area under y=ymax from the lowest to the highest concentration or absorbed radiation tested.

Radiation sensitivity prediction

To test whether we can use proteins or p-sites (predictors) to predict the radiation sensitivity, the protein and p-site intensities were used to fit statistical models using both linear (correlation analysis and elastic regression) and nonlinear methods (recursive partitioning tree and random forest). In the nonlinear model analysis, we aim to select a small number of the most relevant proteins and p-sites to the radiation sensitivity (feature selection) so the predictors in the models can be evaluated from a biological point of view. This is achieved by a two-step procedure. The first step is feature selection using recursive partitioning tree, which selects a small number of proteins or p-site to build predictive models. To identify robust predictors, the 66 PDAC cell lines were randomly split into a training/validation set of size 49 and a hold-out set of size 17. In the model fitting, the training/validation set were further randomly split into training (32 cell lines) and validation (17 cell lines) set. Recursive partitioning tree were trained using the training set, then the fitted models were applied to predict the radiation sensitivity of validating set. The Pearson correlation between predicted and measured radiation sensitivity were computed to evaluate the performance of the model. This procedure was repeated 100 times using proteins and p-site as predictors, separately. Next, to select the most robust predictors, the proteins and sites were ranked by the score calculated as the weighted sum of the predictor importance in both validation and training set:

$$s = \sum_{i=1}^{100} \left(R_i^{train} \cdot V_i^{train} + R_i^{test} \cdot V_i^{test}\right)$$

where R is the correlation coefficient between predicted and measure response, V means importance of a predictor measured by Gini importance index. In a partitioning tree, if a node contains samples belonging to J classes, the Gini purity g of a node is

$$g = \sum_{1}^{J} P_j\left(1 - P_j\right)$$

where P_j is the proportion of samples from class j ($j = 1, \ldots, J$) in the total number of samples. A parent node is divided into two child nodes (g^L and g^R) in the partitioning tree. Assuming a node is split based on the expression of protein x, the node consists of n samples and m samples fall into the node g^R in the branch, the Gini importance index for the node I_x is calculated as

$$I_x = g_x - \frac{m}{n} g_x^R - \frac{n-m}{n} g_x^L$$

When a protein is selected multiple times over a tree, the sum of all Gini importance indices of nodes using that protein is the Gini importance index for the protein. Therefore, the Gini importance index depicts how prediction accuracy is lowered when the corresponding predictor is removed from the model. The superscript *train* and *test* indicate the training and validation set, respectively. The subscript i indicates the ith sampling. The performance of models from training set is included to lower the weight of deficient models who give inaccurate prediction even in the presence of overfitting.

In the second step, we test whether a small number of proteins or p-sites is enough to predict radiation sensitivity. To do so, we selected the 20 top-ranked predictors and used them as predictors to train predictive models using the random forest methods (using R package randomForest[57]) with the same procedure as above. Finally, the performance of the random forest models was evaluated using the hold-out set. In each random forest model, 500 trees were trained. The chosen number of 20 predictors (*i.e.* proteins of p-sites) is somewhat arbitrary but yielded reasonable predictions in both training and validation sets. In addition, the number is small enough to enable checking their functions via thorough literature mining. At the same time, considering 20 predictors provides for a higher chance to contain true positive results compared to selecting 1 or 2 predictors only. The functional interaction among selected predictors was derived from the STRING database[30].

Identifying drug response markers using correlation and elastic net analyses

Due to the large number of drugs (407 drugs) but smaller number of cell lines (on average 36 cell lines have drug sensitivity data), a rigorous examination of nonlinear models is challenging. Therefore, we decided to use correlation analysis and elastic net regression to discover potential protein and p-site markers predicting drug sensitivity.

In the correlation analysis, we calculated correlation coefficients between the drug responses (measured as the area under the dose-response curve; AUC) and protein or p-site intensities. To reduce the spurious correlations, we refrained from imputing missing values in both drug response and proteomic data and excluded pair-wise incomplete observations. Only drug-protein/p-site pairs with at least 10 pair-wise complete observations were included in the analysis. In our analysis, the P-value was used as the ranking statistic rather than hard cutoffs denoting statistical significance, therefore, we did not correct the P-values of these correlations for multiple testing.

To identify linear combinations of multiple proteins/p-sites predicting drug response, we applied elastic net regression[22]. Elastic net regression takes advantage of both LASSO (L1) and RIDGE (L2) penalties, therefore, forces most of the coefficients to zero (property of LASSO penalty) and, at the same time, selects only a subset of predictors showing high correlation to the dependent variable (drug response; owing to the RIDGE penalty). Double shrinkage of coefficients is prevented by a scaling factor while the hyperparameter α? [0,1] is used to control the balance between the L2-penalty (α=0) and the L1-penalty (α=1). In addition, a second hyperparameter λ controls the degree of regularization[58]. In our analysis, α (three options are 0.01, 0.05, 0.1) and λ were optimized using 10-fold cross-validation with mean-squared-error (MSE) as the loss function. We used 100 bootstrap samples of cell lines to select robust protein and site markers predicting drug response. Elastic net models were fitted as described above in every bootstrap sample. The results were summarized using two statistics: the regression coefficient and the selection frequency, which is essentially the number of times a protein or p-site was selected out of 100 bootstrap samples. Since these models were trained independently on each drug, we z-transformed the coefficients in order to facilitate the comparison of effect sizes between different models (referred as normalized effect size). Elastic net regression models were fit using the R package glmnet[58]. Cell lines that do not have drug response data were excluded in the analysis, resulting in 36 cell lines.

RNA sequencing

Total RNA was isolated using the RNeasy Mini Plus Kit (Qiagen), according to the manufacturer instructions. RNA was quantified (Nanodrop, Thermo Fisher Scientific) and quality checked with a Tape Station D1000 (Agilent Technologies). RIN values between 7 and 10 were accepted for further analysis. cDNA libraries were prepared using the TruSeq Stranded mRNA Sample Preparation kit (Illumina) according to the manufacturer instructions. Clusters were generated and sequenced with the HiSeq 4000 platform (Illumina) to a depth of 50 million reads per sample.

Raw read files were quality-checked with FastQC software (version 0.11.8; www.bioinformatics.babraham.ac.uk/projects/fastqc), and RNA-seq data were trimmed to remove adaptor contaminations and poor-quality base calls using Trim galore (version 0.5.0; www.bioinformatics.babraham.ac.uk/projects/trim_galore). After that, resulting read files were checked again with FastQC and mapped with Kallisto[59] (version 0.44.0) to the mouse GENCODE M19 transcriptome, using default settings.

Gene-level summarization of transcripts per million (TPM) values were computed in R by the tximport package[60], and a cutoff of -3 zTPM[61] was used as lower limit for detection across samples. Unless otherwise stated, genes TPMs were normalized based using the total sum approach, as for the proteome dataset.

Protein-to-mRNA ratio

The Pearson correlation coefficient was used for correlating proteome and transcriptome gene abundances in single tissues. For the proteome dataset, the protein with the largest iBAQ value was selected as representative of a given gene. The slopes were estimated by ranged major-axis (RMA) regression using the lmodel2 R package[62], as independent and dependent variables are not expressed in the same units and the error variances of the two variable differ along their value.

Features for protein level prediction models

Predictors selected for this analysis were: mRNA levels, codon usage, non-synonymous-to-synonymous substitution (dN/dS) ratios, which are a measure of evolutionary conservation, gene/coding sequence (CDS) length, exon number, the number of putative protein interactions and mRNA sequence motifs (k-mers of size 3 to 7 nucleotides).

De novo motif identification

Raw mass spectrometry data were processed using MaxQuant (version 1.6.0.13), with a false-discovery rater (FDR) < 0.01 at the level of proteins, peptides and modifications, using the settings described above, and enabling the match between runs feature. Full and phosphoproteome data were processed together as two separate groups, and only full proteome raw data were used for protein quantification. Before any subsequent analysis, MaxQuant output tables were filtered for contaminants, reverse and proteins only identified by a modified peptides. Intensity-based absolute quantification (iBAQ) was used as protein abundance estimator. The iBAQ protein intensities were log10-transformed and median-centered. Furthermore, normalized iBAQs were shifted into positive numerical space by the overall median of the raw iBAQ values. For each protein group, we then selected the first UniProt entry in the “Proteins” column to represent this protein group and subsequently assigned it the corresponding Ensembl transcript ID.

Trimmed RNA-seq reads were mapped with the STAR alignment software[63] (version 2.6.1c) to mouse genome annotation GRCm38, with the parameter of maximum number of multiple alignments allowed for a read to be equal to 1 (outFilterMultimapNmax). To estimate the mature mRNA levels, for each sample the number of reads that map to exonic and intronic regions of the transcript was counted separately and then normalized by the total exonic and intronic region lengths, respectively. Normalized intronic counts were subtracted from normalized exonic counts to obtain the mature mRNA counts. The resulting counts per sample were corrected by the library size factor obtained with the Bioconductor package DESeq2[64], log10-transformed and median centered as described for the proteome data. Transcripts with 10 reads per 1 kb were treated as transcribed.

Tissue-specific PTRs were computed using the normalized protein and transcript levels. For Human and A. thaliana, PTRs were obtained from to Eraslan *et al.*[16] and to Mergner *et al.*[8], respectively.

GEMMA software (version 0.94.1)[65] was used to identify *de novo* motifs in 5' UTR, CDS and 3' UTR regions similar to Eraslan *et al.*[16], using the tissue-specific PTRs as response variables. GEMMA uses a linear mixed model, in which the effect of each individual k-mer on the median PTR across tissues is assessed while controlling for the effect of other k-mers (random effects), region length and GC percentage (fixed effects). The motif search was performed for k-mers ranging from 3 to 7 nucleotides. Obtained P values were adjusted for multiple testing with Benjamini–Hochberg's FDR and jointly computed across the P values of all tissues. Gemma was run using the median PTR with FDR < 0.1 and covariates set to 'false'.

In total, 34/50/80 (Mouse/Human/A. thaliana) significant putative motifs were obtained based on their sequence (5' UTR n = 26/20/28; 3' UTR n = 6/25/41; CDS n = 2/5/11) and sub-sequence (initial, all, end) region. The presence or absence of each enriched motif with respect to each gene was extracted in form of a binary matrix and used for downstream multivariate feature selection analysis.

Codon usage

Codon usage statistics for the Mouse, Human, and A. thaliana genome were obtained from Kazusa (www.kazusa.or.jp/codon) and parsed to extract GI sequence identifiers. These identifiers were mapped to corresponding UniProt entries and Ensembl/TAIR10 annotation using the UniProt Retrieve/ID mapping tool. The extracted TAIR10 annotation was merged with the Kazusa codon usage dataset. Codon frequencies were calculated for each gene by dividing the count (x 3) of a given codon by the full length of the coding sequence.

Synonymous and non-synonymous substitution rates

The dN/dS substitution rates were calculated from CDS pairs of closest species between *Mus musculus* and *Mus spretus*, *Homo sapiens* and *Pan troglodytes*, and *A. thaliana* and *A. lyrata*. Reciprocal best BLAST hits with an e-value cutoff of =1e-08 were used to identify orthologs between each pair. Individual CDS pairs were aligned using PRANK[66] and Gblocks[67] was applied to eliminate poorly aligned positions in an alignment with a cutoff of 8 contiguous nonconserved positions and none gap position allowed. The codeml package from the program PAML[68] for pairwise sequence comparison was used to estimate synonymous substitution and non-synonymous substitution rates (dS and dN, respectively).

Exon number and gene length

The total number of exons and the total gene lengths for Mouse *and* Human was retrieved from Ensembl and for A. thaliana was obtained from the Araport11.

Protein-Protein interactions and phosphorylation status

Protein–protein interactions were downloaded from STRING[30] (https://string-db.org/). The number of protein interaction partners was extracted for each gene. No scoring filter was applied.

The phosphorylation status was extracted from the MaxQuant output tables obtained for the *de novo* motif identification, in form of a binary matrix (i.e. if any of the protein-product of a given gene has been detected as phosphorylated or not).

Model-based feature selection

We used the above features as predictors for gene-level protein abundance. To do this, we merged all predictors into a single feature matrix for each tissue. Preliminary pairwise correlation analysis showed only weak to moderate correlations among individual features, suggesting that multicollinearity was not an issue. The exception were high correlations among several of the sequence motifs. Such high correlations were typically seen in situations where one of the motifs was a substring of the other. To filter these out, we identified motifs that correlated > 0.6, and retained only the longer of the two for subsequent model selection. To select the most predictive features for protein abundance in each tissue, we used a forward–backward model selection approach in a multiple regression framework. The method was implemented using *stepwiseAIC()* function in R, which compares the fit of nested models. To ensure that the comparison of model AICs was not affected by unequal sample size, missing data were removed before the analysis. For each tissue, features from the best fitting models were summarized in an incidence matrix along with the effect direction (positive or negative effect on protein levels). Our previous work[8] showed that the forward-backward model selection approach in this setting does not lead to overfitting when compared to a Lasso approach followed by cross-validation. To determine the importance of each feature to the overall model fits, R^2 variance decomposition was performed using the ‘genici’ metric which is implemented in the relaimpo R package[69]. Relative feature contributions were calculated for each tissues as well as averaged across all tissues.

Orthologous genes quantification

The recently published proteome atlas by Wang *et al.*[17], containing 21 matching tissues, was selected as reference human dataset for orthologous gene analysis at the proteome level. Because of the high redundancy (protein isoforms) and high similarity (protein families) within any given proteome, combined with the impossibility to maintain the same peptide-to-protein parsimony scheme across two independent datasets, homologous proteins calling is rather challenging.

To overcome these limitations, we decided to carry out homology comparison at the gene level (DNA homology) and used the largest iBAQ value of each gene to estimate its absolute quantities. This heuristic approach was repeated across the 21 matching tissues between the two datasets. Because of the exact origin of the human brain tissue was unknown, mouse ‘brain’ gene iBAQ values were obtained by calculating the median across the 6 mouse nervous system tissues. Likewise, the mouse ‘adipose tissue’ iBAQ values were obtained by calculating the mean between the brown and white adipose tissues.

The analysis was restricted to only orthologous genes with a 1:1 correspondence between the two species, and the list of mouse-to-human orthologues was obtained from the Ensembl BioMart (https://www.ensembl.org/biomart/martview).

Before variance decomposition and neighborhood analysis of conserved co-expression (see below) the proteome dataset was filtered for those orthologous genes quantified in at least 6 mouse and human tissues pairs. Missing values were then replaced by imputation with values representing a normal distribution around the detection limit[47], for each tissues and species, separately.

Variance decomposition

The abundance of a gene in any given sample is dependent on biological sample origin, meaning by the tissue and the species the sample comes from. To assess the contribution of these two factors to gene abundance variation, we used a linear mixed model[18], implemented in the R package lme4[70]. The iBAQ ortholog values were modeled as a function of tissue and species, both considered as random factors. The restricted maximum likelihood estimators for the random effects of tissue, species and residual variance were normalized by their sum to obtain the variance contributions. We then classified orthologs mouse-human gene pair whose fraction of variance explained by tissues or species was at least 50% of the total variance and two-fold greater than the other fraction, in tissue-variable genes (TVOs) and species-variable genes (SVOs), respectively. SVOs are genes whose abundance varies much between species but little across tissues (e.g. high abundant in all mouse tissues and low abundant in all human tissues, or vice versa). In contrast, TVOs represent genes that have a much more similar abundance pattern between human and mouse, therefore potentially better suited to model human biology. Remaining genes were classified as mixed-variable orthologs.

Neighborhood analysis of conserved co-expression (NACC)

To evaluate the degree of conservation of orthologous genes abundance between mouse and human, at the proteome level, NACC analysis was performed as described previously[18] with some adjustments. The NACC value for any given orthologous gene pair, was calculated as follows: first, we retrieved the 20 orthologs (neighbors) with the smallest Euclidean abundance distance to the test gene, within the mouse dataset, and calculated the mean distance test gene-neighbor genes (M). Next, we retrieved the human orthologous of these neighbors, and calculate the mean Euclidean distance to the mouse test gene (Mh). The same procedure was repeated within the human dataset, obtaining H and Hm.

The NACC value for each orthologs mouse-human gene pair was calculated as follow:

$$\mathrm{NACC} = \frac{(\mathrm{Mh} - \mathrm{M}) + (\mathrm{Hm} - \mathrm{H})}{2}$$

A random set of NACC values was also calculated using the same procedure, but randomizing the orthology between the two species (i.e. selecting 20 random non-orthologous genes to calculate Mh and Hm). The analysis was performed using either the iBAQ (proteome level) or TPM (transcriptome level) ortholog values. To this end, as for the mouse dataset, human RNA-seq data of 15 matching tissues from Wang *et al.* were mapped with Kallisto (version 0.44.0) to the human GENCODE 29 transcriptome, gene-level summarized, and filtered for a zTPM value > -3, before being merged to the mouse data.

The lower the NACC value, the stronger is the tendency of a gene to be co-expressed with orthologs of similarly expressed genes between the two species.

Benchmarks against previously published datasets

Protein identification as well as phosphorylation sites and phosphoprotein identifications were compared against previously published mouse (phospho)proteomes[4,5]. The proteinGroups.txt output of the MaxQuant 1% FDR search was used to compare protein identifications. For each protein group entry, the first protein in the “Proteins” column and the first genes in the “Gene names” column were selected to represent the protein identification and compared to the corresponding entries reported by Geiger *et al.* in the analogous file. For the phosphoproteome comparison, raw mass spectrometric data from Huttlin *et al.* were downloaded (https://phosphomouse.hms.harvard.edu/data/Phosphorylation) and re-processed together with our data as described above. Only class I sites were considered for the comparison.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The authors are grateful to all members of the Kuster lab for fruitful discussions and creating reagents and tools used in this study. We thank the sample processing laboratory and the genomics and proteomics core facility at the German Cancer Center (DKFZ) in Heidelberg for RNA sequencing, Markus Raspe for help with the drug screen, Frauke Neff, Valerie Gailus-Durner and Helmut Fuchs for their contribution to early discussions of the project, Basak Eraslan and Julien Gagneur for providing access to the source code of the *de novo* mRNA motif analysis scripts, Andrea Hubauer,Michaela Krötz-Fahning, Magdalena Zukowska, Juliane Götzfried and Julia Manolow for expert laboratory assistance.

This work was supported by the Deutsche Forschungsgemeinschaft (SFB1321: P06 to D.S., P13 to G.S., P18 to B.K., S01 to D.S., R.R., G.S.), the German Federal Ministry of Education and Research (Infrafrontier grant 01KX1012 to M.H.d.A., ProteomeTools grant 031L0008A to B.K., DIAS grant 031L0168 to B.K.), the German Center for Diabetes Research (DZD, to M.H.d.A.) and the European Research Council (ERC AdG No. 833710 to B.K., ERC CoG No. 648521 to D.S.). Y.B. is grateful for postdoctoral fellowships from the Alexander von Humboldt Foundation, the Carl Friedrich von Siemens Foundation, and the National Natural Science Foundation of China (81600046). B.K., F.J. and R.R.H. acknowledge support from the Technical University of Munich Institute for Advanced Study, funded by the German Excellent Initiative and the European Seventh Framework Programme under grant agreement no. 291763. “The IBM infrastructure hosting ProteomicsDB and Prosit is operated and maintained by the SAP University Competence Center (UCC) of the Technical University Munich.”

Data availability

The data supporting the findings of this study are available within the paper, the supplementary information and the public repositories. The Mus musculus UniProt FASTA database (UP000000589) was downloaded from the UniProt website (https://www.uniprot.org/). The sORF database was downloaded from www.sorfs.org. The Mus musculus Ensembl, MGI, or NCBI databases, along with their annotation were downloaded from www.ensembl.org, www.informatics.jax.org, or www.ncbi.nlm.nih.gov, respectively. The the Unimod database was downloaded from www.unimod.org. The GENCODE M19 transcriptome was downloaded from www.gencodegenes.org.

Transcriptome sequencing and quantification data are available at ArrayExpress (www.ebi.ac.uk/arrayexpress) under the identifier E-MTAB-10276.

The mass spectrometry proteomics data have been deposited in the ProteomeXchange Consortium via the PRIDE partner repository[71] with the dataset identifier PXD030983.

References

1. the mouse genome. Nature. 2002; 420: 510.
2. Consortium MGS, et al. Initial sequencing and comparative analysis of the mouse genome. Nature. 2002.
3. Su AI, et al. A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Natl Acad Sci U S A. 2004; doi: 10.1073/pnas.0400782101
4. Geiger T, et al. Initial Quantitative Proteomic Map of 28 Mouse Tissues Using the SILAC Mouse. Mol Cell Proteomics. 2013; 12: 1709–1722. [PubMed: 23436904]
5. Huttlin EL, et al. A Tissue-Specific Atlas of Mouse Protein Phosphorylation and Expression. Cell. 2010; 143: 1174–1189. [PubMed: 21183079]
6. Gessulat S, et al. Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. Nat Methods. 2019; 16: 509–518. [PubMed: 31133760]
7. Wilhelm M, et al. Mass-spectrometry-based draft of the human proteome. Nature. 2014; doi: 10.1038/nature13319
8. Mergner J, et al. Mass-spectrometry-based draft of the Arabidopsis proteome. Nature. 2020; doi: 10.1038/s41586-020-2094-2
9. Bult CJ, et al. Mouse Genome Database (MGD) 2019. Nucleic Acids Res. 2019; doi: 10.1093/nar/gky1056
10. The UniProt Consortium. UniProt: a worldwide hub of protein knowledge The UniProt Consortium. Nucleic Acids Res. 2019; doi: 10.1093/nar/gky1049
11. Chick JM, et al. A mass-tolerant database search identifies a large proportion of unassigned spectra in shotgun proteomics as modified peptides. Nat Biotechnol. 2015; 33: 743–749. [PubMed: 26076430]
12. Kong AT, Leprevost FV, Avtonomov DM, Mellacheruvu D, Nesvizhskii AI. MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. Nat Methods. 2017; 14: 513–520. [PubMed: 28394336]
13. Uhlén M, et al. Tissue-based map of the human proteome. Science (80-). 2015; doi: 10.1126/science.1260419
14. Edfors F, et al. Gene-specific correlation of RNA and protein levels in human cells and tissues. Mol Syst Biol. 2016; doi: 10.15252/msb.20167144
15. Franks A, Airoldi E, Slavov N. Post-transcriptional regulation across human tissues. PLoS Comput Biol. 2017; doi: 10.1371/journal.pcbi.1005535

16. Eraslan B, et al. Quantification and discovery of sequence determinants of protein-per-mRNA amount in 29 human tissues. Mol Syst Biol. 2019; 15 e8513 [PubMed: 30777893]

17. Wang D, et al. A deep proteome and transcriptome abundance atlas of 29 healthy human tissues. Mol Syst Biol. 2019; 15 e8503 [PubMed: 30777892]

18. Yue F, et al. A comparative encyclopedia of DNA elements in the mouse genome. Nature. 2014; 515: 355–64. [PubMed: 25409824]

19. Robles MS, Cox J, Mann M. In-vivo quantitative proteomics reveals a key contribution of post-transcriptional mechanisms to the circadian regulation of liver metabolism. PLoS Genet. 2014; 10

20. Hackam DG, Redelmeier DA. Translation of research evidence from animals to humans. Journal of the American Medical Association. 2006; doi: 10.1001/jama.296.14.1731

21. Frejno M, et al. Pharmacoproteomic characterisation of human colon and rectal cancer. Mol Syst Biol. 2017; doi: 10.15252/msb.20177701

22. Frejno M, et al. Proteome activity landscapes of tumor cell lines determine drug responses. Nat Commun. 2020; doi: 10.1038/s41467-020-17336-9

23. Nusinow DP, et al. Quantitative Proteomics of the Cancer Cell Line Encyclopedia. Cell. 2020; doi: 10.1016/j.cell.2019.12.023

24. Yang P, et al. The cytosolic nucleic acid sensor LRRFIP1 mediates the production of type i interferon via a B-catenin-dependent pathway. Nat Immunol. 2010; doi: 10.1038/ni.1876

25. Weichselbaum RR, et al. An interferon-related gene signature for DNA damage resistance is a predictive marker for chemotherapy and radiation for breast cancer. Proc Natl Acad Sci U S A. 2008; doi: 10.1073/pnas.0809242105

26. Warmerdam DO, et al. PHF6 promotes non-homologous end joining and G2 checkpoint recovery. EMBO Rep. 2020; doi: 10.15252/embr.201948460

27. Larsen DH, et al. The NBS1-Treacle complex controls ribosomal RNA transcription in response to DNA damage. Nat Cell Biol. 2014; doi: 10.1038/ncb3007

28. Ciccia A, et al. Treacher Collins syndrome TCOF1 protein cooperates with NBS1 in the DNA damage response. Proc Natl Acad Sci U S A. 2014; doi: 10.1073/pnas.1422488112

29. Weber TJ, et al. Stable Acinar Progenitor Cell Model Identifies Treacle-Dependent Radioresistance. Radiat Res. 2019; doi: 10.1667/RR15342.1

30. Szklarczyk D, et al. STRING v11: Protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. 2019; doi: 10.1093/nar/gky1131

31. Johnson N, et al. Compromised CDK1 activity sensitizes BRCA-proficient cancers to PARP inhibition. Nat Med. 2011; doi: 10.1038/nm.2377

32. Kim SS, et al. Impaired skin and mammary gland development and increased γ-irradiation-induced tumorigenesis in mice carrying a mutation of S1152-ATM phosphorylation site in Brca1. Cancer Res. 2009; doi: 10.1158/0008-5472.CAN-09-2418

33. Cortez D, Wang Y, Qin J, Elledge SJ. Requirement of ATM-dependent phosphorylation of Brca1 in the DNA damage response to double-strand breaks. Science (80-). 1999; doi: 10.1126/science.286.5442.1162

34. Strub T, et al. SIRT6 haploinsufficiency induces BRAFV600E melanoma cell resistance to MAPK inhibitors via IGF signalling. Nat Commun. 2018; doi: 10.1038/s41467-018-05966-z

35. Kim EJ, Juhnn YS. Cyclic AMP signaling reduces sirtuin 6 expression in non-small cell lung cancer cells by promoting ubiquitin-proteasomal degradation via inhibition of the Raf-MEK-ERK (raf/mitogen-activated extracellular signal-regulated kinase/ extracellular signal-regul. J Biol Chem. 2015; doi: 10.1074/jbc.M114.633198

36. Cea M, et al. Evidence for a role of the histone deacetylase SIRT6 in DNA damage response of multiple myeloma cells. Blood. 2016; doi: 10.1182/blood-2015-06-649970

37. Yuan J, et al. SHROOM2 inhibits tumor metastasis through RhoA–ROCK pathway-dependent and-independent mechanisms in nasopharyngeal carcinoma. Cell Death Dis. 2019; doi: 10.1038/s41419-019-1325-7

38. Lee C, Scherr HM, Wallingford JB. Shroom family proteins regulate γ-tubulin distribution and microtubule architecture during epithelial cell shape change. Development. 2007; doi: 10.1242/dev.02828

39. Baker SJ, et al. A Contaminant Impurity, Not Rigosertib, Is a Tubulin Binding Agent. Mol Cell. 2020; doi: 10.1016/j.molcel.2020.05.024

40. Jost M, et al. Pharmaceutical-Grade Rigosertib Is a Microtubule-Destabilizing Agent. Mol Cell. 2020; doi: 10.1016/j.molcel.2020.06.008

41. Mueller S, et al. Evolutionary routes and KRAS dosage define pancreatic cancer phenotypes. Nature. 2018; doi: 10.1038/nature25459

42. Ruprecht B, Zecha J, Zolg DP, Kuster B. High pH Reversed-Phase Micro-Columns for Simple, Sensitive, and Efficient Fractionation of Proteome and (TMT labeled) Phosphoproteome Digests. Methods in molecular biology (Clifton, N.J.). 2017; 1550: 83–98.

43. Yu P, et al. Trimodal Mixed Mode Chromatography That Enables Efficient Offline Two-Dimensional Peptide Fractionation for Proteome Analysis. Anal Chem. 2017; 89: 8884–8891. [PubMed: 28759994]

44. Cox J, Mann M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat Biotechnol. 2008; 26: 1367–1372. [PubMed: 19029910]

45. Savitski MM, Wilhelm M, Hahne H, Kuster B, Bantscheff MA. scalable approach for protein false discovery rate estimation in large proteomic data sets. Mol Cell Proteomics. 2015; doi: 10.1074/mcp.M114.046995

46. Schwanhäusser B, et al. Global quantification of mammalian gene expression control. Nature. 2011; 473: 337–342. [PubMed: 21593866]

47. Tyanova S, et al. The Perseus computational platform for comprehensive analysis of (prote)omics data. Nat Methods. 2016; 13: 731–740. [PubMed: 27348712]

48. R Development Core Team R. F. F. S. C. R: A Language and Environment for Statistical Computing. Vienna Austria R Found Stat Comput. 2008; 1

49. Schmidt T, et al. Universal Spectrum Explorer: A Standalone (Web-)Application for Cross-Resource Spectrum Comparison. J Proteome Res. 2021; doi: 10.1021/acs.jproteome.1c00096

50. Supek F, Bošnjak M, Škunca N, Šmuc T. Revigo summarizes and visualizes long lists of gene ontology terms. PLoS One. 2011; doi: 10.1371/journal.pone.0021800

51. Apweiler R, et al. Update on activities at the Universal Protein Resource (UniProt) in 2013. Nucleic Acids Res. 2013; doi: 10.1093/nar/gks1068

52. Menschaert G, Fenyö D. Proteogenomics from a bioinformatics angle: A growing field. Mass Spectrom. Rev. 2017; 36: 584–599. [PubMed: 26670565]

53. Chambers MC, et al. A cross-platform toolkit for mass spectrometry and proteomics. Nat Biotechnol. 2012; 30: 918–920. [PubMed: 23051804]

54. Keller A, Nesvizhskii AI, Kolker E, Aebersold R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Anal Chem. 2002; 74: 5383–92. [PubMed: 12403597]

55. Nesvizhskii AI, Keller A, Kolker E, Aebersold R. A statistical model for identifying proteins by tandem mass spectrometry. Anal. Chem. 2003; 75: 4646–58. [PubMed: 14632076]

56. Webb-Robertson BJM, et al. Review, evaluation, and discussion of the challenges of missing value imputation for mass spectrometry-based label-free global proteomics. Journal of Proteome Research. 2015; doi: 10.1021/pr501138h

57. Liaw A, Wiener M. Classification and Regression by randomForest. R News. 2002.

58. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw. 2010; doi: 10.18637/jss.v033.i01

59. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol. 2016; 34: 525–527. [PubMed: 27043002]

60. Soneson C, Love MI, Robinson MD. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. F1000Research. 2015; 4 1521 [PubMed: 26925227]

61. Hart T, Komori H, LaMere S, Podshivalova K, Salomon DR. Finding the active genes in deep RNA-seq gene expression studies. BMC Genomics. 2013; 14: 778. [PubMed: 24215113]

62. Csárdi G, Franks A, Choi DS, Airoldi EM, Drummond DA. Accounting for Experimental Noise Reveals That mRNA Levels, Amplified by Post-Transcriptional Processes, Largely Determine Steady-State Protein Levels in Yeast. PLoS Genet. 2015; doi: 10.1371/journal.pgen.1005206

63. Dobin A, et al. STAR: Ultrafast universal RNA-seq aligner. Bioinformatics. 2013; doi: 10.1093/bioinformatics/bts635

64. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014; doi: 10.1186/s13059-014-0550-8

65. Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. Nat Genet. 2012; doi: 10.1038/ng.2310

66. Löytynoja A. Phylogeny-aware alignment with PRANK. Methods Mol Biol. 2014; doi: 10.1007/978-1-62703-646-7_10

67. Castresana J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol. 2000; doi: 10.1093/oxfordjournals.molbev.a026334

68. Yang Z. PAML 4: Phylogenetic analysis by maximum likelihood. Mol Biol Evol. 2007; doi: 10.1093/molbev/msm088

69. Grömping U. Relative importance for linear regression in R: The package relaimpo. J Stat Softw. 2006; doi: 10.18637/jss.v017.i01

70. Bates D, Mächler M, Bolker B, Walker S. Fitting Linear Mixed-Effects Models Using lme4 | Bates | Journal of Statistical Software. J Stat Softw. 2015.

71. Vizcaino JA, et al. The PRoteomics IDEntifications (PRIDE) database and associated tools: status in 2013. Nucleic Acids Res. 2013; 41: D1063–9. [PubMed: 23203882]

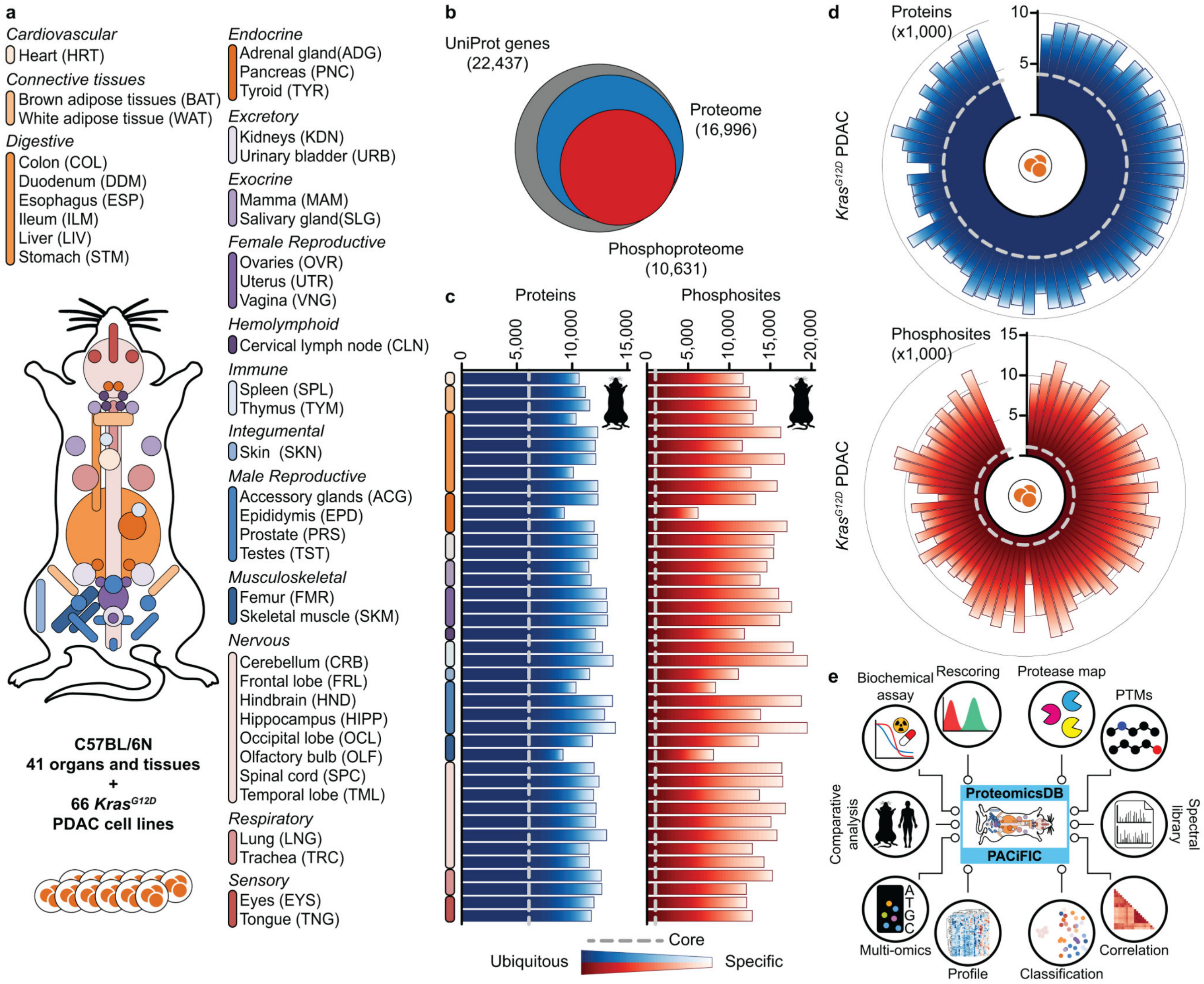

Figure 1. Proteomic map of mouse tissues.

a, Illustration of the 41 tissues (covering 15 systems) and 66 PDAC cell lines subjected to proteome analysis. Each organ system is represented by a unique color code, and each tissue has a unique abbreviation, both are kept consistent throughout the figures. **b,** Number and overlap of identified protein-coding genes in the proteome and phosphoproteome datasets compared to the UniProt database. **c, d,** The number of protein and class I p-site (localization probability > 0.75) identifications for each tissue and cell line is displayed by heatmap bars. The color gradient within each bar reflects the number of samples each protein or p-site was identified in, where the darkest color regions represent the ubiquitous proteomes and phosphoproteomes. Dashed lines indicate proteins and p-sites identified and quantified in all tissues or cell lines. **e,** Schematic representation of the data and analysis workflows available in ProteomicsDB and PACiFIC.

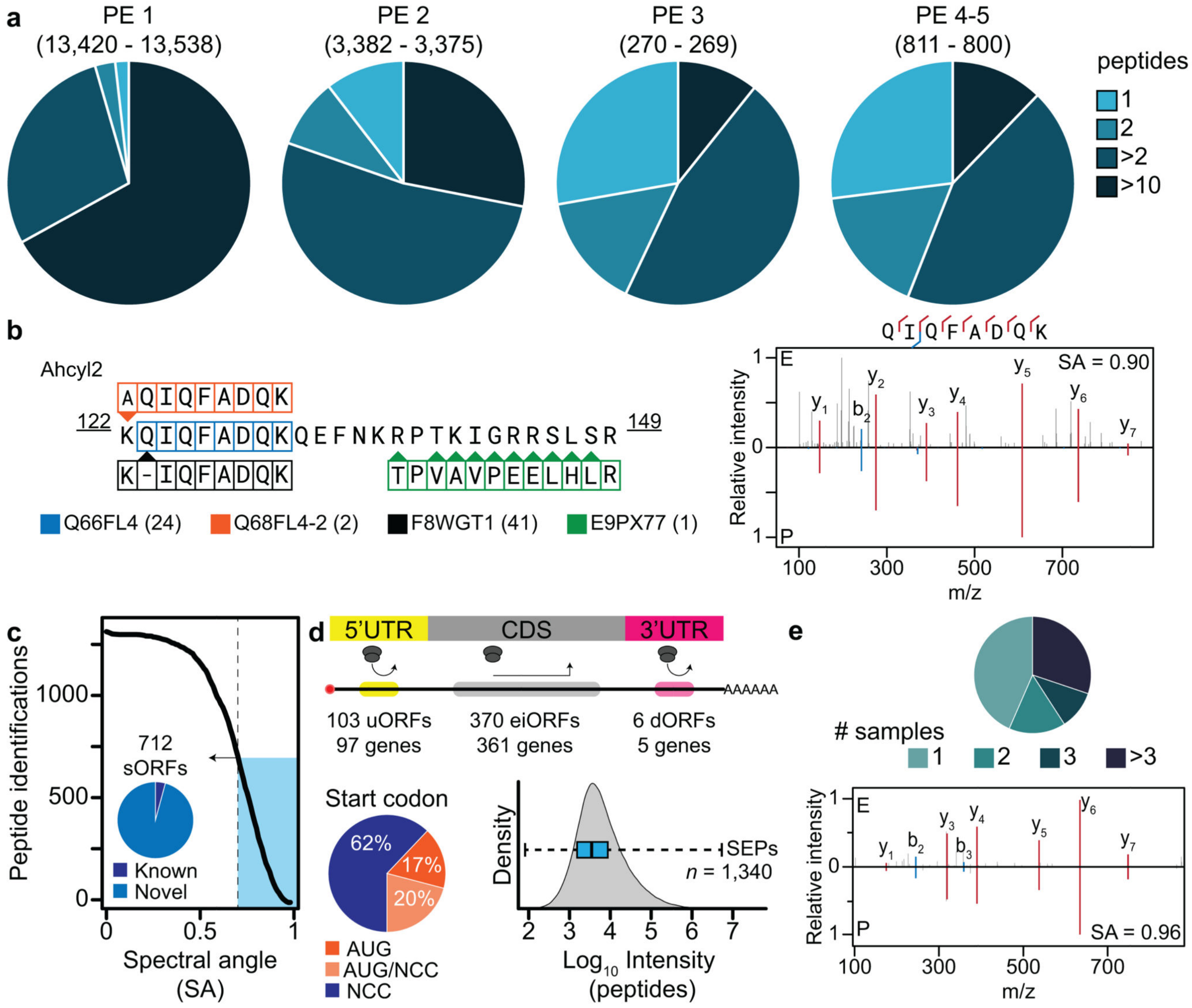

Figure 2. Consolidation of the mouse proteome.

a, Pie charts showing the percentage of proteins identified by one or multiple peptides and grouped by UniProt protein evidence annotations (PE 1-5)). Numbers in brackets refer to the number of identified proteins, along with the number of unique genes they represent. **b,** Spectrum validation of four protein products for the gene Ahcyl2. In the left panel, the amino acid sequence of the canonical protein (Q68FL4) is shown, along with the three alternative products. Portions of the sequences identified in our dataset and which discriminate between the 4 isoforms are highlighted. In the right panel, a mirror plot of the experimental (E, top) and predicted (P, bottom) tandem mass spectra are shown for a representative peptide. Red and blue signals indicate y- and b-type fragment ions, respectively. Calculated spectral angle (SA) of 0.9 indicates near identical spectra. **c,** Number of observed sORF-encoded peptides (SEPs) as a function of the SA comparing measured and predicted reference spectra. SA values of >0.7 (dotted line) indicate near perfect agreement. At this cutoff, our dataset retains 719 SEPs, mapping to 712 unique sORFs (blue area). The inserted pie chart shows the proportion of sORFs with or without MS-based supporting evidence in the sORFs.org database. **d,** Classification and characterization of the validated (SA > 0.7) sORFs, in terms of genetic coordinates (top), initiation codon usage (bottom-left), and intensity distribution (bottom-right). The box indicate the IQR, the black vertical line indicate median value, and whiskers extend to the maximum and minimum values **e,** Identification frequency of the validated SEPs across all tissues and all cell lines. Bottom panel, mirror plot of the experimental (E, top) and predicted peptide (P, bottom) tandem mass spectra of an identified SEP (EDNPFAGSR) without previous MS-based supporting evidence, representing the Rbakdn gene.

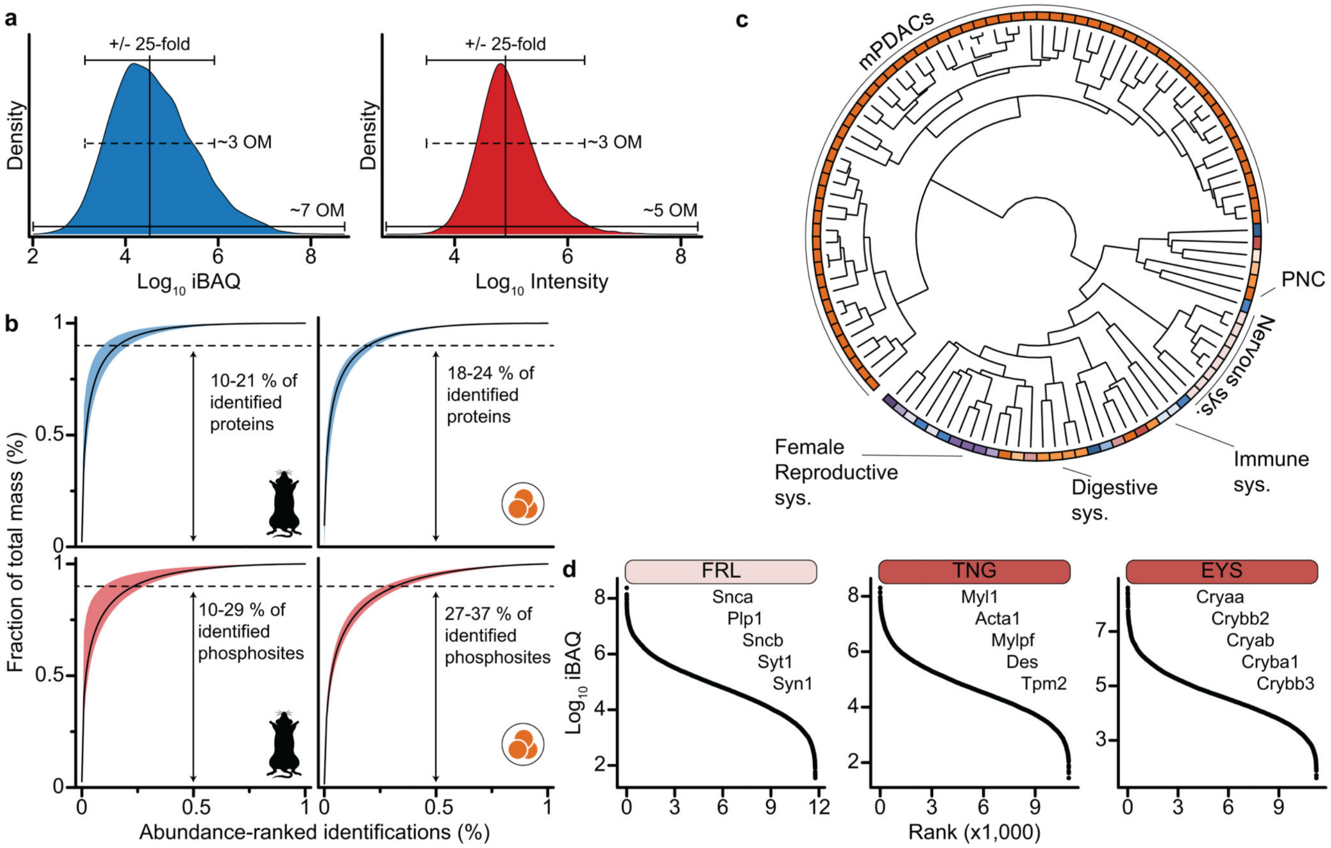

Figure 3. Proteomic expression landscapes in the mouse.

a, Dynamic range of protein abundance (blue) and p-sites (red). Protein abundance spans ~7 orders of magnitude (OM), whereas p-sites abundance only spans ~5. In both cases, ~90% of the proteome or phosphoproteome is confined to within ~3 OM around the median value. **b,** Cumulative protein (top) and p-site (bottom) intensities (ranked by abundance; x axis) and their contribution to total proteome and phosphoproteome mass (y axis), respectively across all tissues or PDAC cell lines. The black solid line indicates the median, the filled area corresponds to the minimum and maximum across tissues or cell lines. **c,** Unsupervised clustering of mouse tissues and mPDAC proteomes, showing that strong qualitative and quantitative expression differences exist between the different proteomes. The clustering separates tissues from mPDACs, but also distinguishes the nervous system tissues, the female reproductive system tissues, the immune system tissues, and, to a lesser extent, the digestive system tissues. **d,** Dynamic range of the intensity-ranked proteomes of three representative tissues. Five of the most abundant genes which relate to the functional specialization of the respective tissue are listed in descending order.

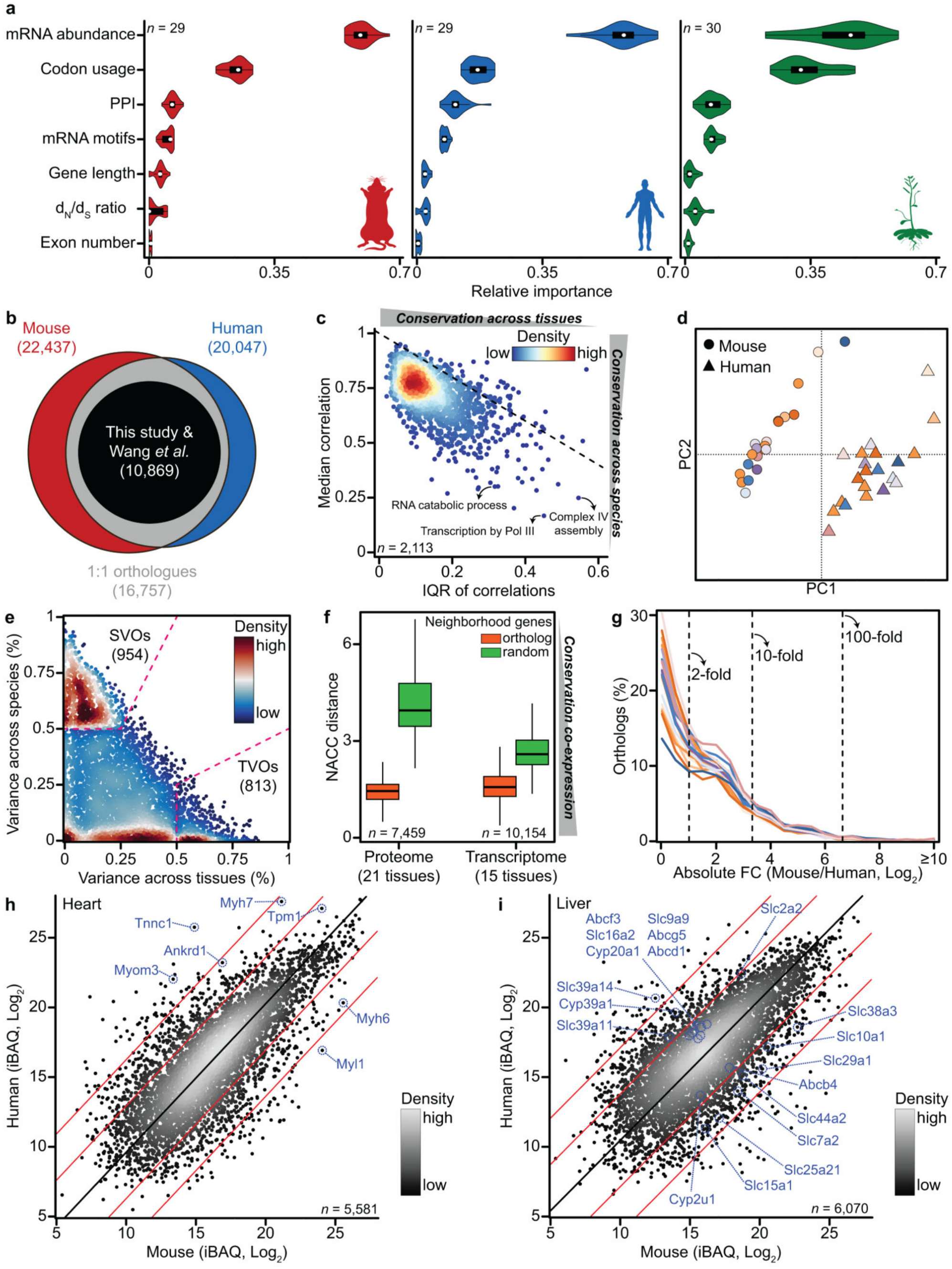

Figure 4. Proteome comparative analysis across tissues and species.

a, Violin plots (n = 29 tissues) depicting the spread in relative contribution of the selected molecular features that can predict gene-level protein abundance using our model across tissues and species. The white dot denotes the median, while box borders indicate the first and third quartiles. Whiskers extend to the maximum and minimum values. **b,** Venn diagram of the relationship between orthologs and identified genes in the two species. **c,** Scatter plot of Pearson correlation coefficients as a measure for co-expression conservation. Each dot represents a gene annotation category (molecular functions, biological processes, or cellular components). Across each tissues pairs, when restricted to only the members of a given category, the proteome expression is highly correlated between mouse and human for the majority of the tested ontologies. However, for a small fraction of functional categories, their members are far less well conserved (higher variability of the person correlation across tissues, x-axis), suggesting different functional remodeling of the mouse and human proteomes during evolution. The dashed line marks the diagonal. **d,** PCA analysis of the 21 mouse and human matching tissues showing a predominant clustering of the proteomes by species. Each tissue is represented by a color matching the ones used in Figure 1 to represent the different anatomical systems. **e,** Proportion of gene expression variance explained by tissues (x-axis) and by species (y-axis) for each orthologous mouse-human gene pair (*n* = 7,459). The proteome abundance variations between mouse and human can be modeled considering two contributing factors: the species of origin and the type of tissues. Variance decomposition identified a large set of species-variable orthologs (SVOs) and tissue-variable orthologs (TVOs). The density estimation is calculated independently for each of the 3 sections of the plot, denoted by the dashed lines. **f,** Neighbourhood analysis of conserved co-expression (NACC) between mouse and human matching tissues at the proteome and transcriptome level. The distribution of NACC distances for each gene is shown, which represents the tendency of a gene to be co-expressed with the same set of orthologs in both species. The boxes indicate the interquartile range (IQR), the black horizontal lines indicate median values, and whiskers extend to +/- 1.5×IQR; no outliers are shown. **g,** Percentage of orthologs having a certain fold change when comparing each tissues pair. Between the two species, orthologs can differ as much as 100-fold. The colored lines indicate the different tissues. **h, i,** Scatter plot depicting proteome-based expression levels of mouse and human genes with 1:1 orthologs, highlighting differentially expressed genes in heart (h), and liver (i). The solid black line indicates the linear model estimated by reduced major-axis regression, other lines indicate absolute fold changes from the regression line of log2(10) and log2(100).

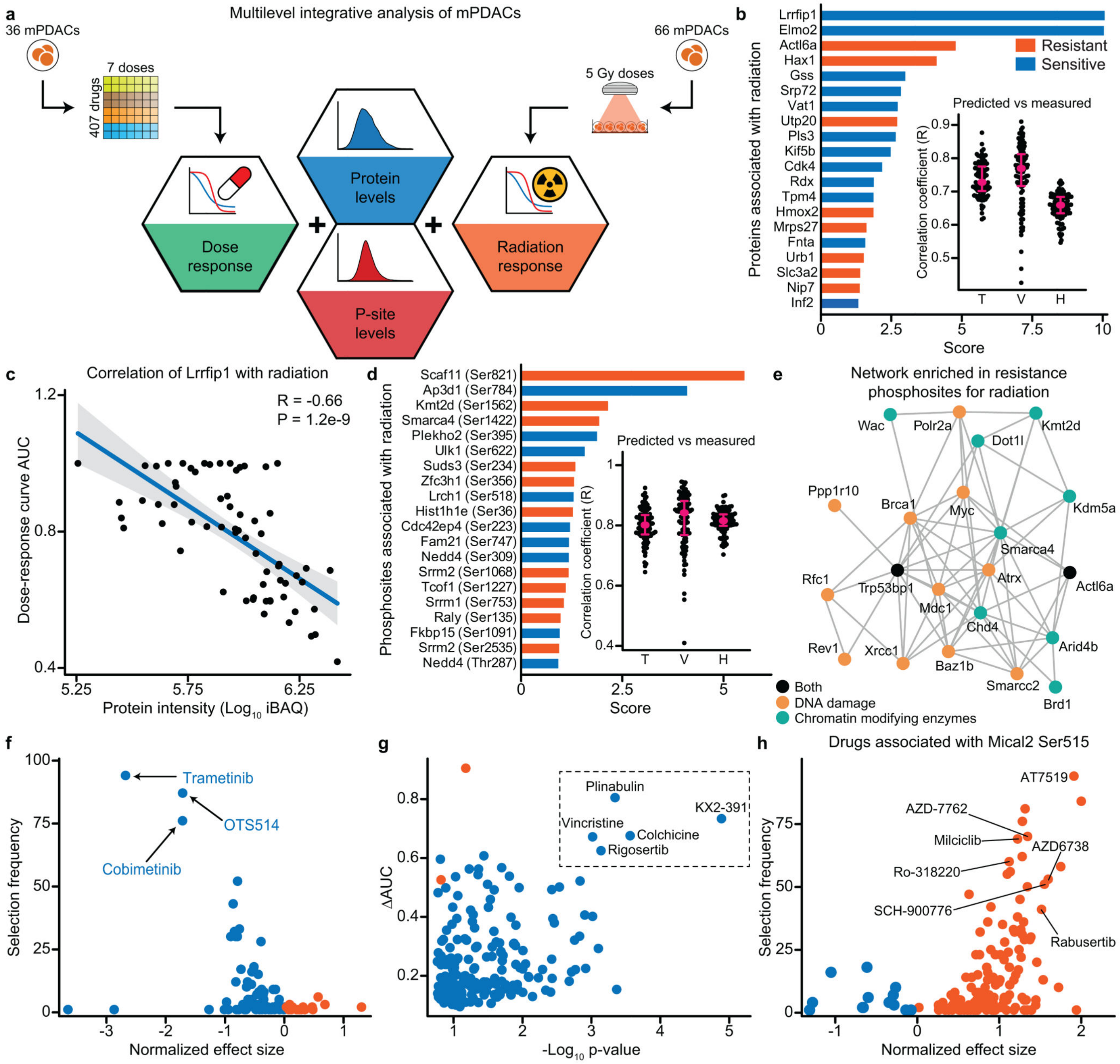

Figure 5. Linking large proteomic data collection with phenotypic drug and radiation response data.

a, Schematic representation of the multilevel integrative analysis workflow performed in this study to identify protein or p-site signatures associated with sensitivity or resistance. **b,** General selection at protein level by the partitioning tree method of the mPDACs panel in the radiation response dataset. The inset shows the prediction accuracy (Pearson correlation, *n* = 100 predictive models) between the predicted and measured radiation activity of random forest models combining the selected 20 proteins (see Methods). The median value and the IQR are indicated in purple. T, V, and H indicate the training, the validation and the hold-out data, respectively. Markers for resistance and sensitivity are colored in orange and blue, respectively. This color scheme is consistently used throughout the other panels of the figure. **c,** Lrrfip1 is a sensitive marker for radiation response (n = 66 cell lines, Pearson correlation, two-sided Pearson correlation test P < 0.05). The filled area indicates the 95% confidence interval, in blue the regression line. **d,** Same as Fig. 5b, but for p-sites. **e,** STRING-based interaction networks as in (e).DNA damage and chromatin modifying enzyme networks are highly enriched in p-sites positively correlated with radiation activity. **f,** Scatter plot from elastic net regression analysis showing that Sirt6 is a sensitivity marker for multiple inhibitors targeting Mek1/2. **g,** Scatter plot showing that Shroom2 is a sensitivity marker for five drugs targeting tubulin. ΔAUC indicates the difference between the maximum and minimum value of the standardized area under the dose-response curve (AUC) across the tested cell lines, plotted against the p-values of the Pearson correlation between Shroom2 abundance and drug sensitivity. **h,** Scatter plot showing that Mical2 Ser515 is a resistant marker for multiple inhibitors targeting CDK, CHK1, or ATR.