Mass spectrometry-based draft of the mouse proteome

Piero Giansanti,1,21 Patroklos Samaras,1,21 Yangyang Bian,1,21, Chen Meng,2 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,6,7,13, Chien-Yun Lee,1 Nicole Wimberger,4,5,6,7, Ludwig Lautenbacher,3, 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,4,6,7,35, 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,4,5,6,7, and Bernhard Kuster,1,3,10,11,13

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.

The mouse is arguably among the most important mammalian model systems for basic, translational and biomedical research. 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. As most biological processes are controlled by proteins and their posttranslational modifications (PTMs), adding a proteome dimension is a logical consequence. An initial analysis of 28 mouse tissues provided evidence for roughly 7,300 proteins and an earlier phosphoproteome investigation identified roughly 36,000 phosphorylation sites (p-sites) in nine tissues. 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 KrasG12D 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 (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 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.
p-sites for tissues and cell lines, respectively. Reanalysis 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 supported by at least two peptides and the median sequence coverage of proteins is 36% (Supplementary Fig. 1d). When tightening the false-discovery rate (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 of roughly 5,400 proteins detected in all tissues show common expression patterns, which account for around 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.
Fig. 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 (PEI–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 four 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 SA of 0.9 indicates near identical spectra. c, Number of observed 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 indicates the interquartile range (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 Rbokdh gene.

To facilitate the use of the molecular atlas by the scientific community, all data have been deployed to ProteomicsDB (https://www.proteomicsdb.org). 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 levels, evaluate the extent of PTMs on a protein of interest, model and estimate drug and radiosensitivity based on (phospho)proteome abundance or cross-interrogate hypotheses 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. 1c).

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 (refs. 810). At the time of writing, the UniProt Mus musculus complete proteome set lacked experimental evidence at the protein level for 7,999 genes (roughly 36%). Our analysis confirmed expression of 13,538 genes with previous 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 and Supplementary Fig. 4d), 1,663 splice variants of 776 genes were identified from the proteomic data. By applying rigorous FDR control and computational validation (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.
Protein expression and phosphorylation levels show broadly than half of all mouse proteins are phosphorylated at least once. The phosphoproteomic data show that more protein expression and phosphorylation characteristics of tissues clearly segregated from the rest but anatomically adjacent tissues clearly segregated from the rest but anatomically adjacent tissues. The functional specialization of a tissue was also clearly 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.

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 indicate that sORFs are canonical components of the mouse proteome but these need to be validated independently (Fig. 2c–e and Supplementary Table 3).

We further explored the data for unexpected proteoforms such as mutations or PTMs. Using an open-search strategy11,12 (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 (Methods), 2,418 new 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 and Supplementary Table 3). Future investigations can be envisaged that make use of this data, for example, 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 show 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 25 times 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 quantities11 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 that 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...
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 posttranslational level. Several insights have been obtained by integrating (paired) transcriptomic and proteomic data collected for tissues of the same species\textsuperscript{8,14–16}. Here, we asked to what extent molecular determinants of protein abundance regulation are conserved between species. We collected RNA-sequencing (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 and 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 Figs. 6b and 7 and 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 messenger RNA levels for controlling protein abundance (Supplementary Fig. 6c).

We next investigated whether molecular features previously identified for human and Arabidopsis\textsuperscript{8–10} 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). 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’ untranslated region (UTR), coding region and 3’ UTR differed substantially between species indicating that regulatory mRNA sequence elements are used with variable preferences (Supplementary Fig. 6f). 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\textsuperscript{17} (representing 65% of all human–mouse orthologs, Fig. 4b and Supplementary Table 6). Overall, we observed a high degree of correspondence (Pearson R = 0.75; Fig. 4c and Supplementary Fig. 8a,b). However, principal component analysis (PCA) 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) (Methods) and whose abundances vary more between species than tissues (Fig. 5e and Supplementary Fig. 8c).

Based on transcriptome data, the mouse ENCODE Project\textsuperscript{19} concluded that most orthologous genes show conserved coexpression between human and mouse. Neighborhood analysis of conserved coexpression (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 ten-times 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. A few of these genes (that is, Abca1, Abca6, Cyp1a1, Cyp2el and Cyp7a1) have been reported to oscillate in terms of abundance\textsuperscript{16}. It is, therefore, possible that the source of the differences between mouse and human may be in part the result of an unsynchronized circadian rhythm at the time of the liver 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,

Fig. 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. PPI, protein–protein interactions. 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 coexpression conservation. Each dot represents a gene annotation category (molecular functions, biological processes or cellular components). Across each tissue pair, when restricted to only the members of a given category, the proteome expression is highly correlated between mouse and human for most 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 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 Fig. 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 SVOs and TVOs. The density estimation is calculated independently for each of the three sections of the plot, denoted by the dashed lines. f, 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 coexpressed with the same set of orthologs in both species. The boxes indicate the IQR, the black horizontal lines indicate median values and whiskers extend to ±1.5xIQR; no outliers are shown. g, Percentage of orthologs having a certain fold change when comparing each tissue pair. Between the two species, orthologs can differ as much as 100-fold. The colored lines indicate the different tissues. h, Scatter plot depicting proteome-based expression levels of mouse and human genes with 1:1 orthologs, highlighting differentially expressed genes in heart (b) 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 log(10) and log(100).

NATURE METHODS | VOL 19 | JULY 2022 | 803–811 | www.nature.com/naturemethods
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.

**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. Despite limitations, murine cell lines are still important tools for research and drug discovery, particularly for diseases such as 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 five doses of ionizing radiation (Fig. 5a and Supplementary Table 7).

Using machine learning (that is, combining a recursive partitioning tree for feature selection and a random approach forest for predicative modeling; Methods), we integrated the quantitative
Fig. 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 (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 (Pearson correlation, \( n = 66 \), \( R = -0.66 \), \( P < 10^{-6} \)). The filled area indicates the 95% confidence interval, in blue is the regression line. d, Same as Fig. 5b, but for p-sites. e, STRING-based interaction networks as before. DNA damage and chromatin modifying enzyme networks are highly enriched in p-sites positively correlated with radiation sensitivity or resistance (Fig. 5b). The strongest correlation coefficient (R) is observed for Mrps27 (R = 0.75, \( P = 1.2 \times 10^{-4} \)). The filled area indicates the 95% confidence interval, in blue is the regression line. 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 Shrooms2 is a sensitivity marker for five drugs targeting tubulin. ∆AUC indicates the difference between the maximum and minimum value of the standardized 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.
BAF chromatin remodeling complex was strongly associated with radiation resistance \( (P = 1.7 \times 10^{-7}, R = 0.59, n = 66) \) confirming previous observations\(^{36}\). Training a classifier based on these 20 proteins showed reasonable accuracy in predicting radiation response \( (R = 0.74, 0.75 \text{ and } 0.66 \text{ for training, validation and hold-out data, respectively; Fig. 5b inset})\). Taking the same approach for the phosphorylation data also identified candidate markers that showed reasonable prediction accuracy \( (R = 0.80, 0.81 \text{ and } 0.81 \text{ for training, validation and hold-out data, respectively; Fig. 5d})\). 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 \( (P = 3.62 \times 10^{-4}, R = 0.62, n = 66) \) is associated with radiation resistance. Following DNA damage, it has been shown that the nucleolar localization of NBS1 (ref. 27), a main player in DNA repair, is triggered in a Tcof1 phosphorylation-dependent manner, with both CK2 and ATM kinases responsible of the phosphorylation of Tcof1 (ref. 28). Consistently, silencing of TCOF1 expression radiosensitized rat progenitor cells\(^{30}\), while phosphorylation levels of Tcof1 disproportionately increased in radiosensitive 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 \( (R > 0.5 \text{ or } R < -0.5 \text{ and } P < 0.0001) \) to STRING analysis\(^{16}\). 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. 5c). Among these are p-sites of Brca1 (pSer1149, \( P = 1.4 \times 10^{-3}, R = 0.78, n = 22 \); pSer152, \( P = 1.2 \times 10^{-4}, R = 0.54, n = 44 \); pSer1154, \( P = 2.3 \times 10^{-4}, R = 0.52, n = 45 \)), the last two are known substrates of ATM and CDK1 (refs. 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 and Supplementary Table 7) gave rise to the number of interesting pharmacodynamics biomarkers. For instance, elastic net regression (Fig. 5f and Supplementary Fig. 11c) and Combimetbin (\( P = 8.4 \times 10^{-4}, R = -0.62, n = 35 \); Supplementary Fig. 11d). While this has been noted before\(^{16}\), our analysis places OTSS14 (\( P = 6.1 \times 10^{-4}, R = -0.55, n = 35 \); Supplementary Fig. 11e), an inhibitor of the MAPKK-like kinase TOPK, and its target into the same functional context\(^{16}\). Another example is Shroom2, whose abundance is correlated with KX2-391 sensitivity \( (P = 1.3 \times 10^{-3}, R = -0.70, n = 31) \) (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\(^{17}\) as well as regulating the accumulation of \( \gamma \)-tubulin and cell morphology\(^{18}\). 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 previous suggestions that Rigosertib also acts as a (direct or indirect) tubulin polymerization inhibitor\(^{19,20}\). 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 \( (P = 1.2 \times 10^{-3}, R = 0.66, n = 36) \), the CHEK1 inhibitor Robusertib \( (P = 4.2 \times 10^{-3}, R = 0.63, n = 36) \) or the ataxia telangiectasia and Rad3-related (ATR) inhibitor AZD6738 \( (P = 9.6 \times 10^{-4}, R = 0.67, n = 36) \) (Fig. 5b). Mical2 is a nuclear monooxygenase that regulates transcription via demethylation 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 explaining 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.

**Online content**

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41592-022-01526-y.

Received: 17 January 2022; Accepted: 17 May 2022; Published online: 16 June 2022

**References**

1. The mouse genome. Nature 420, 510 (2002).
2. Consortium, M. G. S. et al. Initial sequencing and comparative analysis of the mouse genome. Nature 420, 520–562 (2002).
3. Su, A. I. et al. A gene atlas of the mouse and human protein-encoding transcriptionomes. Proc. Natl. Acad. Sci. USA https://doi.org/10.1073/pnas.0400782101 (2004).
4. Geiger, T. et al. Initial quantitative proteomic map of 28 mouse tissues using the SILAC mouse. Mol. Cell. Proteomics 12, 1709–1722 (2013).
5. Huttlin, E. L. et al. A tissue-specific atlas of mouse protein phosphorylation and expression. Cell 143, 1174–1189 (2010).
6. Gessulat, S. et al. Prosite: proteome-wide prediction of peptide tandem mass spectra by deep learning. Nat. Methods 16, 509–518 (2019).
7. Wilhelm, M. et al. Mass-spectrometry-based draft of the human proteome. Nature https://doi.org/10.1038/nature13319 (2014).
8. Mergner, J. et al. Mass-spectrometry-based draft of the Arabidopsis proteome. Nature https://doi.org/10.1038/s41586-020-2094-2 (2020).
9. Bult, C. J. et al. Mouse Genome Database (MGD) 2019. Nucleic Acids Res. https://doi.org/10.1093/nar/gky1056 (2019).
10. The UniProt Consortium. UniProt: a worldwide hub of protein knowledge. The UniProt Consortium. Nucleic Acids Res. https://doi.org/10.1093/nar/gky1049 (2019).
11. Chick, J. M. et al. A mass-tolerant database search identifies a large proportion of unassigned spectra in shotgun proteomics as modified peptides. Nat. Biotechnol. 33, 743–749 (2015).
12. Kong, A. T., Leprevost, F. V., Antonovom, D. M., Mellacheru, D. & Nesvizhskii, A. I. MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. Nat. Methods 14, 513–520 (2017).
13. Uhlén, M. et al. Tissue-based map of the human proteome. Science https://doi.org/10.1126/science.1260419 (2015).
14. Edfors, F. et al. Gene-specific correlation of RNA and protein levels in human cells and tissues. Mol. Cell. Biol. https://doi.org/10.1128/mcb.20167144 (2016).
15. Frankis, A., Airoldi, E. & Slavov, N. Post-transcriptional regulation across human tissues. PLoS Comput. Biol. https://doi.org/10.1371/journal.pcbi.1005535 (2017).
16. Eraslal, B. et al. Quantification and discovery of sequence determinants of protein-per-mRNA amount in 29 human tissues. Nat. Biotechnol. 35; e8513 (2019).
17. Wang, D. et al. A deep proteome and transcriptome abundance atlas of 29 healthy human tissues. Mol. Syst. Biol. 15, e8503 (2019).
18. Yue, F. et al. A comprehensive encyclopedia of DNA elements in the mouse genome. Nature 515, 355–364 (2014).
19. Robles, M. S., 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, 10, (2014).

20. Hackam, D. G. & Redelmeier, D. A. Translation of research evidence from animals to humans. JAMA https://doi.org/10.1001/jama.296.14.1731 (2006).

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

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

23. Nusinow, D. P. et al. Quantitative proteomics of the Cancer Cell Line Encyclopedia. Cell https://doi.org/10.1016/j.cell.2019.12.023 (2020).

24. Yang, P. et al. The cytosolic nucleic acid sensor LRRFIP1 mediates the production of type I interferon via a β-catenin-dependent pathway. Nat. Immunol. https://doi.org/10.1038/nii.2010.10.

25. Weichselbaum, R. R. 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. USA https://doi.org/10.1073/pnas.0809242105 (2008).

26. Warmerdam, D. O. et al. PHF6 promotes non-homologous end joining and G2 checkpoint recovery. EMBO Rep. https://doi.org/10.15252/embr.201948460 (2020).

27. Larsen, D. H. et al. The NBS1-Treacle complex controls ribosomal RNA transcription in response to DNA damage. Nat. Cell Biol. https://doi.org/10.1038/ncb3007 (2014).

28. Ciccia, A. et al. Treacher Collins syndrome TCOF1 protein cooperates with NBS1 in the DNA damage response. Proc. Natl. Acad. Sci. USA https://doi.org/10.1073/pnas.1422488112 (2014).

29. Weber, T. J. et al. Stable acinar progenitor cell model identifies treacle-dependent radioresistance. Radiat. Res. https://doi.org/10.1667/RR15342.1 (2019).

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. https://doi.org/10.1093/nar/gky1131 (2019).

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

32. Kim, S. S. 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. https://doi.org/10.1158/0008-5472.CAN-09-2418 (2009).

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

34. Strub, T. et al. SIRT6 haploinsufficiency induces BRAFV600E melanoma cell resistance to MAPK inhibitors via IGF signalling. Nat. Commun. https://doi.org/10.1038/ncomms15966 (2018).

35. Kim, E. J. & Juhn, Y. S. 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-regulation. J. Biol. Chem. https://doi.org/10.1074/jbc.M114.633198 (2015).

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

37. Yuan, J. et al. SHROOM2 inhibits tumor metastasis through RhoA–ROCK pathway-dependent and -independent mechanisms in nasopharyngeal carcinoma. Cell Death Dis. https://doi.org/10.1038/cdqd0191-1325-7 (2019).

38. Baker, S. J. et al. A contaminant impurity, not rigosertib, is a tubulin binding agent. Mol. Cell https://doi.org/10.1016/j.molcel.2020.05.024 (2020).

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2022
Methods

Tissue preparation. C57BL/6N mice were maintained in individually ventilated cages with water and standard mouse chow according to the directive 2010/63/EU, German laws and German Mouse Clinic housing conditions (www.mouseclinic.de). Mice (16 weeks of age, male and female) were euthanized with CO₂. Afterwards, animals 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[c]. Cells were cultured in Dulbecco’s Modified Eagle Medium (DMEM) high glucose medium (Sigma) supplemented with 10% fetal calf serum (FCS), assay (Thermo Fisher Scientific). Cells were irradiated with 0, 2, 4, 8 or 16 Gy 24 h rate of 0.90 Gy per min using the RS225A irradiation device (Gulmay/Xstrahl). Radiation was delivered at 220 kV and 15 mA with a dose of 80 °C until further use.

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

Radiosensitivity assay. Radiation was delivered at 220 kV and 15 mA with a dose rate of 0.90 Gy per 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 irradiated with 0, 2, 4, 8 or 16 Gy 24 h after seeding. The AlamarBlue reagent was added 72 h after irradiation. After an incubation time of 4 h at 37 °C in a humidified 5% CO₂ atmosphere, proliferation of cells was measured by the AlamarBlue signal. The outlet of the column was divided and two fractions were collected: one containing nonphosphorylated peptides (flow-through), the other containing phosphorylated peptide. Both fractions were dried in a vacuum centrifuge and stored at −80 °C.

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

For the PDAC cells, phosphopeptides were enriched by using 5 μl of 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 depleted 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⁻¹, where 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-through containing nonphosphorylated 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 resuspended in 100 μl of buffer A (25 mM ammonium formate, pH 10), and subjected to high pH RP fractionation with the AssayMAP Bravo platform (Agilent). The cartridges were connected 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⁻¹. The cartridges were equilibrated with 100 μl of buffer A at a flow rate of 10 μl min⁻¹, peptides were then loaded at 5 μl min⁻¹ and the flow-through 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% to 25%, 10% to 30%, 15% to 40% and 20% to 50%), 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. x 2 cm, packed-in-house with 5 μm of Reprofil C18 resin; Dr. Maisch) and washed using 0.1% formic acid at a flow rate of 5 μl min⁻¹ for 10 min. Subsequently, peptides were transferred to an analytical column (75 μm i.d. x 4 cm, packed-in-house with 3 μm Reprofil C18 resin, Dr. Maisch) at a flow rate of 300 μl min⁻¹. Peptides were chromatographically separated using a linear gradient of solvent B (0.1% formic acid, 5% DMSO in ACN) and solvent A (0.1% formic acid, 5% DMSO in water). Linear gradients were as follows: from 0 to 32% of B in 82 min for full proteome analysis, and from 0 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 UHPLC (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⁻¹, using a 60 min linear gradient (3 to 28% of solvent B) (0.1% formic acid, 5% DMSO in ACN) and solvent A (0.1% formic acid, 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% formic acid and solvent A (0.1% formic acid, 5% DMSO in water), from 4 to 15% 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 to automatically switch between MS and MS/MS. For label-free samples full scan MS spectra were recorded in the Orbitrap from 360 to 1,600 m/z, at a resolution of 60,000 using an automatic gain control (AGC) target value of 3 x 10^6 charges and maximum injection time (maxIT) of 10 ms (50 ms for HF-X and 45 ms for Exaporis 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 HF-X and 1.2 m/z for Exaporis 480) for high-collision dissociation fragmentation and fragment ions were recorded in the Orbitrap at a resolution of 15,000 (30,000 for Exaporis 480), AGC of 1 x 10^6 and maxIT of 25 ms (22 for HF-X and 54 ms for Exaporis 480). For phosphoproteome analysis of tissues samples, the 15 most intense precursors were selected instead, while AGC and maxIT were set to 2 x 10^5 and 50 ms, respectively. For the Exaporis 480, the data-dependent mode was set to cycle time, the cycle time was set to 0.8 s and internal compensation voltage stepping experiments with compensation voltage values of −35, −45 and −60 V were applied. Normalized collision energy was set to 25% for the HF and 28% for the HF-X and Exaporis 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 and 27 s for the measurements on the HF and HF-X, respectively. For the Exaporis 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 (v.1.5.3.8), using default settings with the following minor changes: oxidized methionine (M) and acetylation (protein N terminal) 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 with a minimum peptide length of seven amino acids. Searches were performed against the M. musculus UniProt FASTA database (UP000000589, March 2018, 61,307 entries), containing canonical and isoform sequences, as well as against a common contaminants database.

The FDR was set to 1, meaning 100%, to allow for a subsequent peptide-spectrum match (PSM) rescoring via Prosit and Percolator, implemented in ProteomicsDB. After rescoring, PSMs and peptides were filtered at an FDR of 1% (calculated on sample level) and protein- and gene-level FDR estimates were calculated using the picked target-decoy approach 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) 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, log2 transformed and right-shifted by 10log2 units into positive numerical space.

Further bioinformatics analysis was carried out with Perseus (v.1.6.5.0), Microsoft Excel (v.2013) and R (v.3.6.3). Protein and gene annotations were extracted from the UniProt, Ensembl and MGI databases.

**Identification and quantitation of phosphorylation sites.** Raw mass spectrometry data were processed with MaxQuant (v.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 threshold for the localization of phosphorylated residues.

Phosphorylation site intensities were normalized by the total sum approach, log2 transformed and right-shifted by 10log2 units into positive numerical space.

**Analysis of protein expression.** Protein expression analysis including normalization, hierarchical cluster analysis and PCA were performed using R (v.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 analysis of was performed either in Perseus (v.1.6.15.0) or with the R package topGO, using annotations from UniProt and MGI, respectively. The Fisher exact test was performed with a FDR < 0.05.

Only long lists of significantly (q-value < 0.05) enriched gene ontology terms were further uploaded to REVIGO for summarization, using default parameters, restricting the database to the M. 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 catalog for human and mouse analysis, 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 label 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 individual proteins, to annotate the UniProt existence level on the basis of the identifications reported here 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 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 (based on the peptide level), peptides mapped to any gene of interest using the EMBL or the National Center for Biotechnology Information (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 sORF.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 (LC–MS/MS analysis section). Experimental and synthetic peptide spectra were extracted from the raw files and used for similarity calculation without any previous 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 threshold for candidate validation. For variant peptide identification, the full proteome dataset was additionally processed applying the so-called ‘open-search’ strategy, overcoming the limitations of obtaining a comprehensive protein database for each sample from exome or RNA-seq. To this end, raw mass spectrometry data were converted to mzML format using msconvert in the ProteoWizard Toolkit. Peak lists were searched with MSFragger (v.2.4) using the same UniProt database as mentioned above, but with a 0.1 Da evolution and 0.02 Da. Precursor ion mass tolerance was set to 500 Da and remaining settings were used at their default value. The search results were processed with PeptideProphet and ProteinProphet as part of the MSFragger suite, and identification were filtered to retain only hits with a probability > 0.9. At least 90% of all the PSMs were identified outside the −0.01 to 0.01 Da mass bin, representing the modified peptides. For roughly 4 million PSMs (roughly 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, 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, being translational, chemical or postranslational (for example, Ala → Ser, Oxidation and Phe → Tyr all result in the same shift of +15.9949 Da), we excluded the analysis amino acid substitution that could be explained by any of the Unimod annotated modifications (representing potential artifact or known modifications). We then defined as illogical all those substitutions involving a nonalkylated cysteine residue, as well as those requiring more than one nucleotide substitution.

To further validate the existence of a peptide variant, for any given PSM, each potential modification was computed individually and obtained sequences were subjected to spectra prediction by Prosi, using the same charge state detected for the endogenous peptide. To this end, spectra of endogenous peptides (that is, those identified in the full proteome and phosphoproteome datasets) were compared against those obtained by Prosi. SA score was calculated using an in-house R script, and a SA score of at least 0.7 was used as threshold 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 artifact during sample preparation, ProSti-validated substitutions to aspartic acid and glutamic acid whose nondeamidated peptide sequence was found to be identical and not part of an annotated protein were considered ambiguous and therefore discarded.

For all spectra comparisons, only those peaks that were 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. PDAC raw mass spectrometry data were processed using MaxQuant (v.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 (STV) as a 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) 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 log_{10} transformed.

For each dose-response dataset (radiation or drugs), data were normalized into a range between 1 (so response or full viability) and 0 (full response or no viability). Afterward, 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 being between zero and the upper bound or maximum predicted viability (y_{max}), divided by the area under y=y_{max} from the lowest to the highest concentration or absorbed radiation tested.

Radiation sensitivity prediction. To test whether we can use proteins or p-sites (predictions) to predict 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 the recursive partitioning tree procedure. The first step is feature selection using a 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 sets were further randomly split into training (32 cell lines) and validation (17 cell lines) set. Recursive partitioning trees 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 the validation and training set:

\[
s = \sum_{i=1}^{100} (p_{ih} x_{\text{train}} + p_{ih} x_{\text{test}})\]

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 importance \(g\) of a node is

\[
g = \sum_{j=1}^{J} \{P_j (1 - P_j)\}
\]

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' and g") 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 \(g'\) in the branch, the Gini importance index for the node \(I\) is calculated as

\[
I = g - \frac{m}{n} g' - \frac{n - m}{n} g''
\]

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 \(i\)th sampling. The performance of models from training set is included to lower the weight of deficient models who give an 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 \(^{13}\) ) 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 (that is, 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 a higher chance for true positive results to be present compared to selecting only one or two predictors. The functional interaction among selected predictors was derived from the STRING database.

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 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 pairwise incomplete observations. Only signal peptide and terminus peptide pairs with pairwise 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 where the hyperparameter \(\alpha\) (0.01, 0.05, 0.1) is used to control the balance between the L2-penalty (\(\alpha = 0\)) and the L1 penalty (\(\alpha = 1\)). In addition, a second hyperparameter \(\lambda\) controls the degree of regularization\(^{26}\). In our analysis, \(\alpha\) (three options are 0.01, 0.05, 0.1) and \(\lambda\) were optimized using tenfold cross-validation with mean-squared-error 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. Finally, the performance of the random forest models was evaluated using the R package glmnet\(^{27}\). Cell lines that do not have drug response data were excluded in the analysis, resulting in 36 cell lines.

RNA-seq. 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. complementary DNA libraries were prepared using the TruSeq Stranded mRNA Sample Preparation kit (Illumina) according to the manufacturer’s 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 (v.0.11.8; www.bioinformatics.babraham.ac.uk/projects/fastqc), and RNA-seq data were trimmed to remove adapter contaminations and poor-quality base calls using Trim galore (v.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\(^{29}\) (v.0.4.4) 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\(^{30}\), and a cut-off of \(\geq 3\) Z-score relative to the Gaussian fit (\(zTPM\))\(^{31}\) was used as lower limit for detection across samples. Unless otherwise stated, genes TPMs were normalized using the total sum approach, as for the proteome dataset.

PTR. The Pearson correlation coefficient was used for correlating proteome and transcription 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 regression using the lm0deg2 R package\(^{32}\), as independent and dependent variables are not expressed in the same units and the error variances of the two variables differ along their value.

Features for protein-level prediction models. Predictors selected for this analysis were: mRNA levels, codon usage, nonsynonymous-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 3–7 nucleotides).

De novo motif identification. Raw mass spectrometry data were processed using MaxQuant (v.1.6.0.13), with a 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 modified peptides. iBAQ was used as protein abundance estimator. The iBAQ protein intensities were log_{10} 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 Protein chain 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 (v.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 Biocoutr package DESeq2 (ref. 64), log_{10} transformed and median centered as described for the proteome data. Transcripts with ten reads per kb were treated as transcribed.

Tissue-specific protein-to-mRNA ratios (PTRs) were computed using the normalized protein and transcript levels.

For Human and Arabidopsis thaliana, PTRs were obtained from to Eraslan et al.14 and to Mercq et al.16, respectively. GEMMA software (v.0.94.17) was used to identify de novo motifs in 5' UTR, CDS and 3' UTR regions similar to Eraslan et al.14, 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 other k-mers (main 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 (q-value < 0.1) 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 subsequence (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 GenInfo Identifier 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 (x3) of a given codon by the full length of the CDS.

Synonymous and nonsynonymous substitution rates. The d_s/d_N substitution rates were calculated from CDS pairs of closest species between M. musculus and Mus spretus, Homo sapiens and Pan troglodytes, and A. thaliana and A. lyrata. Reciprocal best BLAST hits with a cutoff of ≤1 × 10^{-4} were used to identify orthologs between each pair. Individual CDS pairs were aligned using PRANK and Gblocks was applied to eliminate poorly aligned positions in an alignment with a cutoff of eight contiguous nonconserved positions and none gap position allowed. The codeml package from the program PAML was applied for pairwise sequence comparison was used to estimate synonymous substitution and nonsynonymous substitution rates (d_s and d_N, 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 AraExpress.

Protein–protein interactions and phosphorylation status. Protein–protein interactions were downloaded from STRING (https://string-db.org/). The number of protein interaction partners was extracted for each gene. No scoring 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 (that is, 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 the 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 two models were summarized in an incidence matrix along with the effect direction (positive or negative effect on protein levels). Our previous work 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 fit, R^2 variance decomposition was performed using the 'genric' metric, which is implemented in the relaimpo R package. Relative feature contributions were calculated for each tissue as well as averaged across all tissues.

Orthologous genes quantification. The recently published proteome atlas by Wang et al.15 contained a protein database, was set as the Buxco as our 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 challenging.

To overcome these limits, 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 six 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 orthologs was obtained from the Ensembl BioMart (https://www.ensembl.org/biomart/mview).

Before variance decomposition and NACC (below) the proteome dataset was filtered for those orthologous genes quantified in at least six mouse and human tissues pairs. Missing values were then replaced by imputation with values representing a normal distribution around the detection limit18, for each tissue 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 model, where variance was accounted for by tissue and species, respectively. SVOs are genes whose abundance is twofold 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 (for example, highly 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, and are therefore potentially better suited to model human biology. Remaining genes were classified as mixed-variable orthologs.

NACC. To evaluate the degree of conservation of orthologous genes abundance between mouse and human, at the proteome level we carried out NACC analysis as described previously15 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 orthologous mouse–human gene pair was calculated as follows:

\[
\text{NACC} = \frac{(Mh - M) + (Hm - H)}{2}
\]
A random set of NACC values was also calculated using the same procedure, but randomizing the orthology between the two species (that is, selecting 20 random nonorthologous genes to calculate Mb and Hm). We carried out the analysis 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 (v0.44.0) to the human GENCODE 29 transcriptome, gene-level summarized and filtered for a rTPM value \( > 3 \), before being merged to the mouse data.

The lower the NACC value, the stronger is the tendency of a gene to be coexpressed 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\(^7\). 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 reprocessed together with our data as described above. Only class I sites were considered for the comparison.

### Data availability

The data supporting the findings of this study are available within the paper, the supplementary information and the public repositories. The \( M. \) \( \text{musculus} \) UniProt FASTA database (UP000000589) was downloaded from the UniProt website (https://www.uniprot.org/). The soRF database was downloaded from www.sorls.org. The \( M. \) \( \text{musculus} \) Ensembl, MGI or NCBI databases, along with their annotation, were downloaded from the following websites: 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 phosphoproteomics data has been deposited in the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD030983.

### References

41. Mueller, S. et al. Evolutionary routes and KRAS dosage define pancreatic cancer phenotypes. *Nature* https://doi.org/10.1038/nature25459 (2018).

42. Ruprecht, B., Zecha, J., Zolg, D. P. & Kuster, B. in *Proteomics: Methods and Protocols* Vol. 1530 (eds Comai, L. et al.) 83–98 (Springer, 2017).

43. Yu, P. et al. Trimodal mixed mode chromatography that enables efficient off-line two-dimensional peptide fractionation for proteome analysis. *Anal. Chem.* 89, 8884–8891 (2017).

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.* 26, 1367–1372 (2008).

45. Savitski, M. M., Wilhelm, M., Hahne, H., Kuster, B. & Bantscheff, M. A scalable approach for protein false discovery rate estimation in large proteomic data sets. *Mol. Cell. Proteomics* https://doi.org/10.1074/mcp.M114.046995 (2015).

46. Schwambach, B. et al. Global quantification of mammalian gene expression control. *Nature* 473, 337–342 (2011).

47. Tyanova, S. et al. The PerSeus computational platform for comprehensive analysis of (pro)teomics data. *Nat. Methods* 13, 731–740 (2016).

48. R Core Team. *R: A Language and Environment for Statistical Computing* (2014); https://www.R-project.org/.

49. Liaw, A. & Wiener, M. Classification and regression by randomForest. *J. Stat. Softw.* https://doi.org/10.18637/jss.v013.i05 (2002).

50. Soneson, C., Love, M. I. & Robinson, M. D. Differential analyses for label-free global proteomics. *J. Proteome Res.* https://doi.org/10.1021/pr501138h (2015).

51. Geiger, B. et al. Motif analysis scripts, A. Hubauer, M. Krötz-Fahning, M. Zukowska, J. Götzfried and N. Gehrke. Laboratory and the genomics and proteomics core facility at the German Cancer Center (DKFZ) in Heidelberg for RNA-seq. M. Raspe for help with the drug screen. F. Neff, V. Gailus-Durner and H. Fuchs for their contribution to early discussions of the project. B. Eralan and J. Gnegner for providing access to the source code of the de novo mRNA motif analysis scripts. A. Hubauer, M. Krötz-Fahning, M. Zukowska, J. Götzfried and J. Manolow for expert laboratory assistance. This work was supported by the Deutsche Forschungsgemeinschaft (grant nos. SFB1321: P06 to D.S., P13 to G.S., P18 to B.K. and S01 to D.S., R.R. and G.S.), the German Federal Ministry of Education and Research (Infronfer grant no. 01KX1012 to M.H.d.A., ProteomeTools grant no. 03110008A to B.K., DIAS grant no. 0311068B to B.K.), the German Center for Diabetes Research (to M.H.d.A.) and the European Research Council (ERC CoG grant 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 (grant no. 81600406). 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 Prost is operated and maintained by the SAP University Competence Center of the Technical University Munich.

### Acknowledgements

We are grateful to all members of the Kuster laboratory for fruitful discussions and creating reagents and tools used in this work. We thank the sample processing laboratory and the genomics and proteomics core facility at the German Cancer Center (DKFZ) in Heidelberg for RNA-seq. M. Raspe for help with the drug screen. F. Neff, V. Gailus-Durner and H. Fuchs for their contribution to early discussions of the project. B. Eralan and J. Gnegner for providing access to the source code of the de novo mRNA motif analysis scripts. A. Hubauer, M. Krötz-Fahning, M. Zukowska, J. Götzfried and J. Manolow for expert laboratory assistance. This work was supported by the Deutsche Forschungsgemeinschaft (grant nos. SFB1321: P06 to D.S., P13 to G.S., P18 to B.K. and S01 to D.S., R.R. and G.S.), the German Federal Ministry of Education and Research (Infronfer grant no. 01KX1012 to M.H.d.A., ProteomeTools grant no. 03110008A to B.K., DIAS grant no. 0311068B to B.K.), the German Center for Diabetes Research (to M.H.d.A.) and the European Research Council (ERC CoG grant 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 (grant no. 81600406). 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 Prost is operated and maintained by the SAP University Competence Center of the Technical University Munich.

### Author contributions

PG. performed (phospho)proteomic and transcriptomic experiments on tissue samples. Y.B., JR., JK., C.-Y.L., Y.-C.C. and F.P.B. performed (phospho)proteomic experiments on mPDAC cell lines. P.G., PS., Y.B. and B.K. interpreted and visualized data. P.G., PS., Y.B. and C.M. generated web resource databases. P.G., P.S., Y.B. and B.K. interpreted and visualized data. P.G., P.S., Y.B. 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 Prost is operated and maintained by the SAP University Competence Center of the Technical University Munich.
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.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41592-022-01526-y.

Correspondence and requests for materials should be addressed to Bernhard Kuster.

Peer review information Nature Methods thanks Leonard Foster and Lukas Käll for their contribution to the peer review of this work. Peer reviewer reports are available. Primary Handling Editor: Arunima Singh, in collaboration with the Nature Methods team.

Reprints and permissions information is available at www.nature.com/reprints.
Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

- Confirm that you have included the following for all statistical analyses
  - n/a
  - The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
  - A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
  - The statistical test(s) used AND whether they are one- or two-sided
  - Only common tests should be described solely by name; describe more complex techniques in the Methods section.
  - A description of all covariates tested
  - A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
  - A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
  - For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted Give P values as exact values wherever suitable.
  - For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
  - For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
  - Estimates of effect sizes (e.g. Cohen’s d, Pearson’s r), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about: availability of computer code

Data collection

MS measurements were performed on a QExactive Orbitrap HF, QExactive Orbitrap HF-X, and an Orbitrap Exploris 480 mass spectrometer. RNA sequencing was performed on a HiSeq 4000 platform. For the radiosensitivity assay, radiation were delivered using the RS225A irradiation device. Cell lines were screened for radiation response using the AlamarBlue proliferation assay. Proliferation of cells was measured using the ELx808 microplate reader, and data were recorded by the GenS Software.

The automated drug screen was performed using a Multidrop Combi Reagent Dispenser and a Cytomat™ 24C automated incubator. Cells were treated with the compound library using a CyBio Felix pipetting platform. Cell viability was measured using CellTiter-Glo Luminescent Cell Viability Assay in an Infinite 200 PRO microplate reader. Detailed information is available in the manuscript.

Data analysis

Data were processed with MaxQuant (version 1.5.3.8 or 1.6.0.13), MSFagger (version 2.4), FastQC (version 0.11.8), Trim galore (version 0.5.0), Kallisto (version 0.44.0), and STAR (version 2.6.1c). Data analysis was performed in ProteomicsDB, R (version 3.6.3), Perseus (version 1.6.15.0), Excel (version 2013), and GEMMA (version 0.94.1). Detailed information is available in the manuscript.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. Github). See the Nature Portfolio guidelines for submitting code & software for further information.
Data

Policy information about availability of data
All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:
- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

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 (UP000005989) was downloaded from the UniProt website (https://www.uniprot.org/). The sORF database was downloaded from www.sorfs.org. The Mus musculus Ensemble, MGI, or NCB 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 EMTAB-10276.
The mass spectrometry proteomics data have been deposited in the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD030983.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

- Life sciences
- Behavioural & social sciences
- Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/re-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size
No statistical method was used to predetermine sample size.
For each tissue and cell line, one sample was measured. Tissues samples were chosen in a way that the proteome atlas contains n ≥ 2 tissues for each anatomical system (except for the cardiovascular, hemolymphoid, and integumental systems, where n = 1). Primary murine PDAC cell lines, consisting of 66 low-passaged lines with various genetic backgrounds, were obtained from genetically engineered mice as described in https://doi.org/10.1038/nature25459
Automated drug screen experiments were performed once on a subset of the PDAC cell lines (PDACs n = 36 randomly chosen, drugs n = 407) with three replicates for each concentration tested.

Data exclusions
All data were analyzed.

Replication
At least 2 distinct tissues were measured for each anatomical system (except for the cardiovascular, hemolymphoid, and integumental systems, where n = 1). For the cardiovascular, hemolymphoid, and integumental systems only a single tissue was dissected from the mouse. Any additional tissues that could have been dissected from those systems, upon disruption and dissection did not yield the required proteins or mRNA amount for the subsequent proteome and transcriptome analysis. For reproducibility evaluation of the employed workflow, one PDAC cell line (9591) was measured 6 times and 7 tissues (Hindbrain, Occipital lobe, Pancreas, Skeletal muscle, Skin, Spinal cord, Temporal lobe) were measured twice (Supplementary Figure 2).
Radiosensitivity assay was performed once but with four radiation doses and a control.
Drug screen was performed once but with three replicates for each of the seven tested concentrations (3-fold dilution series, highest concentration 10 μM).
All attempt at replication were successful for all experiment were replication was performed.
All other experiments were performed once.

Randomization
Samples for mass spectrometry (all tissues and cell lines) and RNAseq (tissues for which a RIN value was higher than 7) measurements were prepared and measured in random order.

Blinding
Blinding was not performed in this study. It was not relevant for this study, since we present a resource data.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.
### Materials & experimental systems

| n/a | Involved in the study |
|-----|-----------------------|
| ☒   | Antibodies            |
| ☒   | Eukaryotic cell lines |
| ☒   | Palaeontology and archaeology |
| ☒   | Animals and other organisms |
| ☒   | Human research participants |
| ☒   | Clinical data         |
| ☒   | Dual use research of concern |

### Methods

| n/a | Involved in the study |
|-----|-----------------------|
| ☒   | ChIP-seq              |
| ☒   | Flow cytometry        |
| ☒   | MRI-based neuroimaging |

### Eukaryotic cell lines

**Policy information about cell lines**

- **Cell line source(s)**: Primary murine PDAC cell lines, consisting of 66 low-passaged lines with various genetic backgrounds (C57Bl/6, 129S6/SvEv), were obtained from genetically engineered mice as described in [https://doi.org/10.1038/nature25459](https://doi.org/10.1038/nature25459). The murine PDAC cell lines were prepared by the authors.

- **Authentication**: Cell line authentication by re-genotyping has previously been performed for the 2018 Nature paper ([https://doi.org/10.1038/nature25459](https://doi.org/10.1038/nature25459)) as stated in the paper’s methods section.

- **Mycoplasma contamination**: Cell lines were routinely screened for mycoplasma contamination and tested negative.

- **Commonly misidentified lines**
  - (See [ICLC register](#))
  - No commonly misidentified cell line was used.

### Animals and other organisms

**Policy information about studies involving animals**: ARRIVE guidelines recommended for reporting animal research

- **Laboratory animals**: Mus musculus, C57Bl/6N, m/f, 16 weeks old, for housing conditions please refer to [www.mouseclinic.de](http://www.mouseclinic.de)

- **Wild animals**: The study did not involve wild animals.

- **Field-collected samples**: The study did not involve samples collected from the field.

- **Ethics oversight**: 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](http://www.mouseclinic.de)).

Note that full information on the approval of the study protocol must also be provided in the manuscript.