

1 **Rapid proteotyping reveals cancer biology and drug response determinants in the NCI-**
2 **60 cells**

3
4 Tiannan Guo ^{1,19} *[#], Augustin Luna ^{2,3}*, Vinodh N Rajapakse ⁴*, Ching Chiek Koh ¹*[†],
5 Zhicheng Wu ¹⁹, Michael P Menden ^{5,21}, Yongran Cheng ¹⁹, Laurence Calzone ⁶, Loredana
6 Martignetti ⁶, Alessandro Ori ⁷, Murat Iskar ⁸¹, Ludovic Gillet ¹, Qing Zhong ^{9,20}, Sudhir
7 Varma ¹⁰, Uwe Schmitt ¹¹, Peng Qiu ¹², Yaoting Sun ¹⁹, Yi Zhu ^{1,19}, Peter J Wild ⁹, Mathew J
8 Garnett ¹³, Peer Bork ^{8, 14,15,16}, Martin Beck ^{8, 17}, Julio Saez-Rodriguez ⁵, William C. Reinhold
9 ⁴, Chris Sander ^{2,3}, Yves Pommier ⁴*, Ruedi Aebersold ^{1, 18} #

10

11 * Equal contribution

12 # correspondence

13

14 **Affiliations**

15 1, Department of Biology, Institute of Molecular Systems Biology, ETH Zurich, Switzerland

16 2, cBio Center, Department of Biostatistics and Computational Biology, Dana-Farber Cancer
17 Institute, Boston, MA 02115, USA

18 3, Department of Cell Biology, Harvard Medical School, Boston, MA 02115, USA

19 4, Developmental Therapeutics Branch, Center for Cancer Research, National Cancer
20 Institute, National Institutes of Health, Bethesda, MD 20892, United States

21 5, RWTH Aachen University, Faculty of Medicine, Joint Research Centre for Computational
22 Biomedicine (JRC-COMBINE), Germany

23 6, Institut Curie, PSL Research University, INSERM, U900, Mines Paris Tech, F-75005,
24 Paris, France.

25 7, Leibniz Institute on Aging, Fritz Lipmann Institute (FLI), Beutenbergstrasse 11, 07745
26 Jena, Germany

27 8, Structural and Computational Biology Unit, European Molecular Biology Laboratory,
28 69117 Heidelberg, Germany

29 9, Institute of Surgical Pathology, University Hospital Zurich, Zurich, Switzerland

30 10, HiThru Analytics, Laurel, MD 20707, USA

31 11, Scientific IT Services, ETH Zurich, Switzerland

32 12, Department of Biomedical Engineering, Georgia Institute of Technology and Emory
33 University, 313 Ferst Dr., Atlanta, GA 30332, US

34 13, Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK

35 14, Molecular Medicine Partnership Unit, University of Heidelberg and European Molecular
36 Biology Laboratory, 69120 Heidelberg, Germany

37 15, Max Delbrück Centre for Molecular Medicine, 13125 Berlin, Germany

38 16, Department of Bioinformatics, Biocenter, University of Würzburg, 97074 Würzburg,
39 Germany

40 17, Cell Biology and Biophysics Unit, European Molecular Biology Laboratory, 69117
41 Heidelberg, Germany

42 18, Faculty of Science, University of Zurich, Zurich, Switzerland

43 19, Westlake Institute for Advanced Study, Westlake University, Hangzhou, Zhejiang, P. R.
44 China

45 20, Cancer Data Science Group, Children's Medical Research Institute, University of Sydney,
46 Sydney, New South Wales, Australia.

47 21, Bioscience, Oncology, IMED Biotech Unit, AstraZeneca, Cambridge, UK

48

49 †, current address: Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge
50 CB10 1SA, UK

51 ‡, current address: Division of Molecular Genetics, German Cancer Research Center (DKFZ),
52 Im Neuenheimer Feld 280, 69120 Heidelberg, Germany

53

54 **Summary**

55

56 We describe the rapid and reproducible acquisition of quantitative proteome maps for the
57 NCI-60 cancer cell lines and their use to reveal cancer biology and drug response
58 determinants. Proteome datasets for the 60 cell lines were acquired in duplicate within 30
59 working days using pressure cycling technology and SWATH mass spectrometry. We
60 consistently quantified 3,171 proteotypic proteins annotated in the SwissProt database across
61 all cell lines, generating a data matrix with 0.1% missing values, allowing analyses of protein
62 complexes and pathway activities across all the cancer cells. Systematic and integrative
63 analysis of the genetic variation, mRNA expression and proteomic data of the NCI-60 cancer
64 cell lines uncovered complementarity between different types of molecular data in the
65 prediction of the response to 240 drugs. We additionally identified novel proteomic drug
66 response determinants for clinically relevant chemotherapeutic and targeted therapies. We
67 anticipate that this study represents a significant advance toward the translational application
68 of proteotypes, which reveal biological insights that are easily missed in the absence of
69 proteomic data.

70 **Introduction**

71

72 To date, mainly owing to the maturity and availability of high throughput DNA- and
73 RNA- based techniques, forays into the molecular landscape of diseases, in particular cancers,
74 have primarily focused on genomics and transcriptomics¹⁻³. Protein-level measurements,
75 although showing great potential for providing the granularity and details necessary for
76 personalized therapeutic decisions, are underutilized due to technical hurdles. Advances in
77 data-dependent acquisition (DDA) mass spectrometry (MS) have permitted quantitative
78 proteomic profiling of about 100 tumor samples using multi-dimensional fractionated MS
79 analyses of each sample⁴⁻⁶, demonstrating the added value of protein measurement in
80 classifying tumor samples. Nevertheless, such DDA workflows suffer from relatively lower
81 sample-throughput, relatively higher sample consumption and technical complexity,
82 precluding their routine use in clinically relevant applications (*e.g.* drug response prediction)
83 on the speed and scale achieved by genomic and transcriptomic approaches^{2,3}.

84

85 To achieve reproducible and high throughput proteomic profiling, we have developed
86 a workflow^{7,8} integrating pressure cycling technology (PCT), an emerging sample
87 preparation method that accelerates and standardizes sample preparation for proteomic
88 profiling⁹, together with SWATH-MS, an MS-based proteomic technique that consists of
89 data independent acquisition (DIA) and a targeted data analysis strategy with unique
90 advantages over other MS-based proteomic methods^{10,11}. With this technique all MS-
91 measurable peptides of a sample are fragmented and recorded in a recursive fashion, thus
92 generating digital proteome maps that can be used to reproducibly detect and quantify
93 proteins across high numbers of samples without the need of isotope labeling. The PCT-
94 SWATH technique thus significantly increases the sample throughput and data reproducibility
95 providing excellent quantitative accuracy, and also reduces sample consumption to ca. 1
96 microgram of total peptide mass per sample^{7,8}.

97

98 In this study, we describe the acquisition of proteome maps of the NCI-60 cell lines in
99 duplicate by PCT-SWATH. The 120 proteome maps were acquired within 30 working days
100 on a single instrument and each sample consumed ca. 1 microgram of total peptide mass. We
101 consistently quantified 3,171 SwissProt proteotypic proteins across all cell lines, generating a
102 data matrix (120 proteomes vs. 3171 proteins) with 0.1% missing values. Raw signals of each
103 peptide and protein in each sample were curated with an expert system. The NCI-60 human

104 cancer cell line panel contains 60 lines from 9 different tissue types ¹². The NCI-60 have been
105 molecularly and pharmacologically characterized with unparalleled depth and coverage,
106 offering a prime *in vitro* model to further our understanding of cancer biology and cellular
107 responses to anti-cancer agents ^{12, 13}. Discoveries enabled by the NCI-60 in recent years
108 include the development of the FDA approved drugs oxaliplatin for the treatment of colon
109 cancers ¹⁴, eribulin for metastatic breast cancers ¹², bortezomib for the treatment of multiple
110 myeloma ¹⁵, and rhomidepsin for cutaneous T-cell lymphomas ¹⁶. The sensitivity of the NCI-
111 60 has been measured for over 100,000 synthetic or natural compounds derived from a wide
112 range of academic and industrial sources ¹², constructing the most comprehensive resource for
113 cancer pharmacological research. The proteomic data complement the existing NCI-60
114 molecular landscapes, allowing systematic investigation of the complementarity among
115 genomics, transcriptomics and proteomics in a number of applications.

116

117 The proteome of the NCI-60 cells has been analyzed previously by data dependent
118 analysis (DDA), a commonly used discovery mass spectrometry technique ¹⁷. Whereas the
119 study reported the cumulative identification of 10,350 IPI proteins from about 1,000
120 fractionated and kinase-enriched sample runs, only 492 proteins were quantified across the
121 NCI-60 cell lines without missing value. The present study thus extends the number of
122 proteins consistently quantified in duplicates analyses to 3,171, a ca. six-fold increase. The
123 high quality proteomic data were used for pharmacoproteomic analysis of the response of the
124 cell panel to 240 anti-cancer drugs, resulting in the identification of novel proteomic drug
125 response determinants for clinically relevant chemotherapeutic and targeted therapies.

126 **Results**

127

128 **Acquisition of the NCI-60 proteome maps**

129

130 We applied the PCT-SWATH workflow ⁷ to generate quantitative proteome maps of
131 the NCI-60 cell lines in technical replicates, resulting in the generation of 120 SWATH maps
132 with high reproducibility at the raw data level (**Supplementary Fig. 1**). The PCT-assisted
133 sample preparation took about 18 working days and the SWATH-MS data acquisition
134 consumed about 12 working days. Thus, the entire process, from sample preparation to data
135 acquisition, was accomplished within 30 working days. This constitutes an unprecedented
136 sample-throughput compared to other cancer proteomic workflows of similar scale ^{4-6, 17}. This
137 is the result of the elimination of multidimensional fractionation and the consequential sample
138 processing of each sample through using one barocycler to one mass spectrometer in which a
139 single file per sample was acquired (**Supplementary Fig. 1, Supplementary Table 1**).

140

141 SWATH proteome maps contain fragment ion chromatograms from all MS-measurable
142 peptides, albeit in a highly convoluted form. To interpret the SWATH maps, we built a human
143 cancer cell line spectral library containing 86,209 proteotypic peptides, *i.e.* peptides that
144 uniquely identify a specific protein, from 8,056 SwissProt proteins (**Supplementary Table**
145 **1**). Using this library and the OpenSWATH software ¹¹, we identified 6,556 protein groups,
146 covering 81% of the library (**Supplementary Fig. 2**). To avoid ambiguity of peptide/protein
147 quantification, we limited our analyses to canonical and proteotypic peptides and proteins by
148 excluding protein isoforms, un-reviewed protein sequences, peptide/protein sequence variants
149 and protein groups that could not be deconvoluted.

150

151 We evaluated the technical variation of each measurement through manual inspection
152 of the OpenSWATH results based on the replicated measurement for each cell line and
153 observed in substantial technical variation. This is probably due to the fact that cell type-
154 specific interfering signals leads to invalid SWATH assays, and the presence of irregular
155 liquid chromatography (LC) and MS behavior of certain peptides in the highly variable
156 proteomic context of the NCI-60 cells. These phenomena have also been observed previously
157 in selected reaction monitoring (SRM)-based targeted proteomics studies ¹⁸.

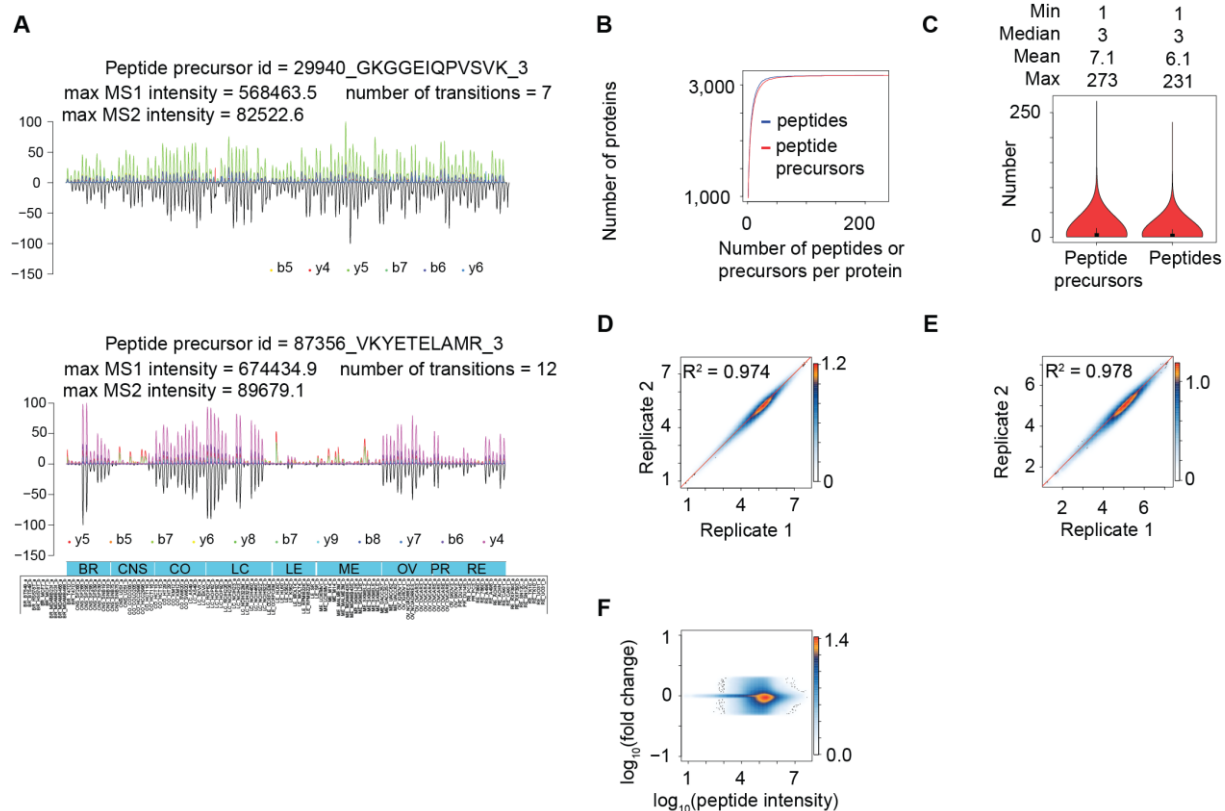
158

159 To obtain high accuracy quantitative data for the cell lines, we further developed an
160 expert system, *i.e.* DIA-expert (see Methods), to refine the peptide identification and
161 quantification provided by automated analysis tools like OpenSWATH (**Fig. 1A**). We thus
162 excluded proteins and peptides that were not reproducibly quantified in technical replicates
163 and focused our analyses on a shorter list of 22,554 proteotypic peptides from 3,171 proteins,
164 with 8% missing values at the peptide level and 0.1% missing values at the protein level
165 across all MS runs (**Supplementary Table 1**). On average, 7 peptide precursors and 6 unique
166 peptide sequences were identified for each protein (**Fig. 1B**). Several proteins were identified
167 with more than 200 peptides (**Fig. 1C**). The proteins excluded by DIA-expert may not be
168 incorrect identifications, but rather proteins that could not achieve reproducible quantification
169 by the existing algorithm across all cell lines due to either technical issues, for instance the
170 signal-to-noise ratio, or biological issues such as post-translational modifications or splicing
171 variants. Improved computational methods will likely rescue some of them in the future.

172

173 Most peptides for the 3,171 proteins were consistently quantified in all cell lines at
174 both MS1 and MS2 levels. Two representative peptides are shown in **Fig. 1A**. The coefficient
175 of determination (R^2) between technical replicates, for overall expression of peptides (**Fig.**
176 **1D**) and proteins (**Fig. 1E**), were 0.974 and 0.978, respectively, with a dynamic range over 5
177 orders of magnitude (**Fig. 1F**). We provide the raw MS signals for each quantitative value in
178 **Supplementary File 1**, allowing visual inspection of the MS signal for every peptide in each
179 sample. When we set the minimal number of peptides identified per protein to 2, 3 or 4,
180 respectively, fewer proteins (2200, 1741, 1428 proteins respectively) were quantified,
181 however, the quantitative accuracy did not substantially improve, indicating that protein
182 quantification by a single, reliably identified proteotypic peptide is similarly accurate as
183 quantification by multiple proteotypic peptides (**Supplementary Figure 3**).

184



185
 186 **Figure 1. Acquisition of NCI-60 proteotype.** (A) Representative peptide signals as curated and visualized by
 187 the DIA-expert software. (B) The cumulative number of peptides and peptide precursors identified for each
 188 protein. (C) The distribution of peptide precursors and peptides per protein. The overall Pearson correlation
 189 between technical replicates at the peptide level (D) and the protein level (E). Here, the log₁₀ transformed
 190 intensity of each peptide/protein in each cell line technical replicate is plotted in the heatmap. (F) Dynamic range
 191 of the MS signals for 22,968 proteotypic peptides.

192

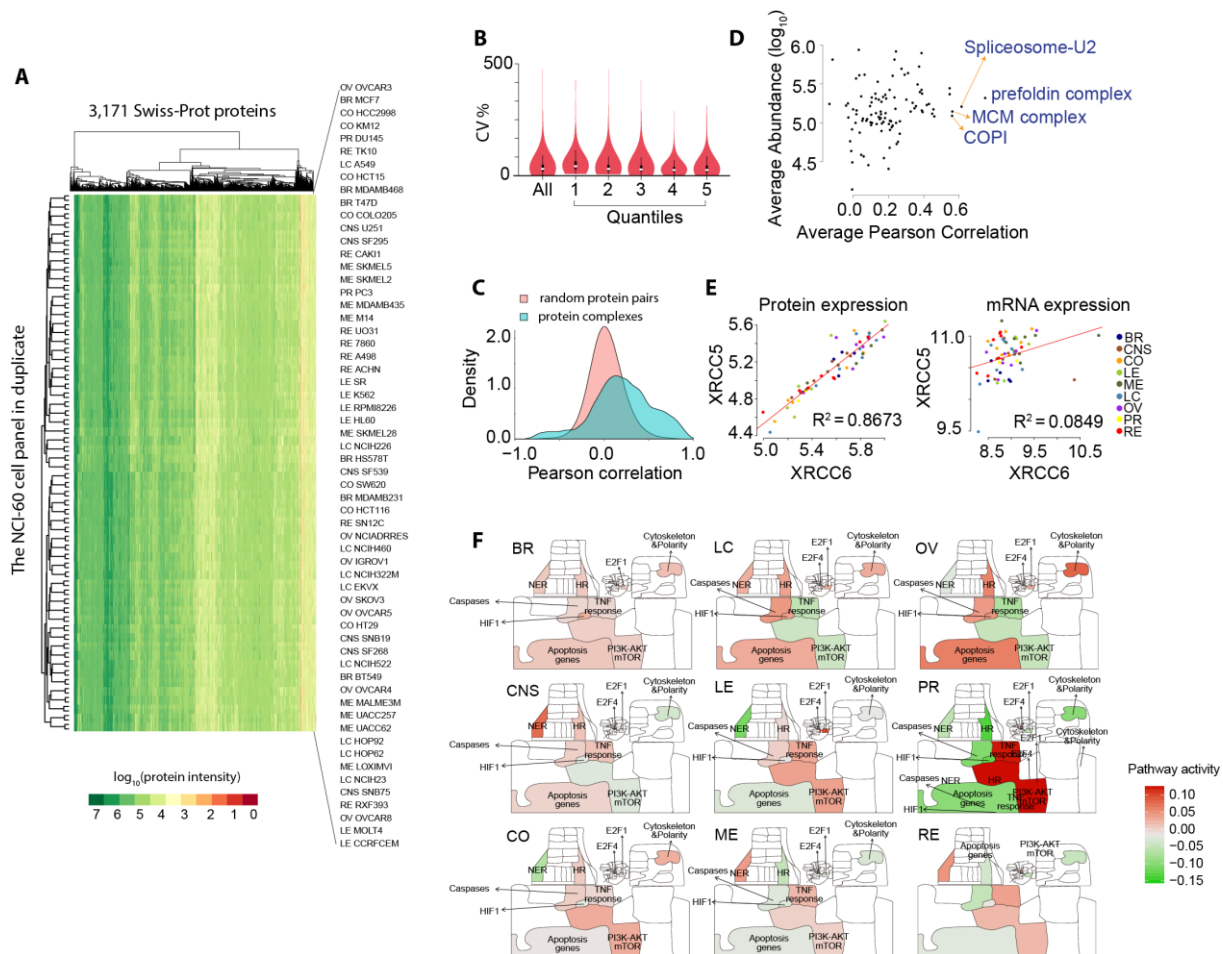
193 Characterization of the NCI-60 quantitative proteomes

194

195 The landscape of the 120 thus measured proteotypes is displayed in **Fig. 2A**. All
 196 technical replicates were clustered together using an unsupervised method based on the
 197 quantified proteotypes, confirming high quantitative accuracy. In most cases, the proteotypes
 198 are not strikingly different across different cancer cell lines, in sharp contrast with the distinct
 199 proteomes of tumor versus non-tumor kidney tissues⁷. The median coefficient of variation
 200 (CV) of the protein intensity in different cells was 48%. The CV demonstrated a low
 201 dependence on protein abundance, as evident from the distribution of its values for different
 202 expression level quantile groups of the measured proteins (**Fig. 2B**). We then compared the
 203 data acquired in this study with the DDA-MS proteomic data previously reported of the NCI-
 204 60 cells¹⁷. Whereas the DDA data reported comparable number of IPI protein groups to the
 205 SwissProt proteotypic protein number from this SWATH data set per cell line

206 (Supplementary Table 2), the SWATH data exhibited much higher degree of consistency
 207 (Supplementary Fig. 5) and better quantitative accuracy (Supplementary Fig. 6-7).

208



209

210

211 **Figure 2. Characterizing NCI-60 quantitative proteomes.** (A) Heatmap overview of NCI-60 proteotype data
 212 matrix. 3,171 Swiss-Prot proteins were quantified in 120 SWATH runs. (B) Variation of protein expression,
 213 for all proteins (All) and proteins in each abundance quantile group (from low abundance to high abundance). (C)
 214 Density plot of correlation of determination between pairs of random proteins versus pairs of proteins within a
 215 complex. (D) Stoichiometry variation of protein complexes in the NCI-60 cells. The x-axis shows the average
 216 Pearson correlation of each protein complex across the NCI-60. The y-axis shows the average abundance of
 217 proteins in a complex. Stable complexes tend to show higher values of average Pearson correlation. (E) Protein
 218 and mRNA expression of XRCC6/Ku70 and XRCC5/Ku80. (F) Visualization of pathway activity in NCI-60
 219 proteotypes. More detailed pathway annotations for this Google map are provided in **Supplementary File 2**.

220

221 Quantification of drug-responsiveness related proteins

222

223 The proteotypes covered 105 protein targets for FDA-approved anti-cancer
 224 compounds, 661 protein drug targets annotated in DrugBank¹⁹ (including 68 drug

225 metabolizing enzymes, 5 drug carriers, and 15 drug transporters), 694 proteins known to
226 participate in human diseases ^{19,20}, and 58 human protein kinases (**Supplementary Table 3**).
227 Some kinases were found to be broadly expressed in most cells with high abundance,
228 including MST4 and WNK1 (**Supplementary Fig. 4**), consistent with previous reports ^{21,22}.
229 Other kinases were highly expressed in specific cell lines, for example, EGFR in the breast
230 cancer cell line MDAMB468, ERBB2 in SKOV3 cells, and CDK6 in MOLT4 cells, in
231 agreement with previous studies using antibody-based methods ^{20,23}.

232

233 A unique benefit of our proteomic data set, compared to genomic and transcriptomic
234 data, is its capacity to reveal more accurate information about the abundance of protein
235 complexes and their stoichiometry ²⁴. Our measurements included 101 protein complexes
236 comprising 1,045 proteins (**Supplementary Table 4**) from a curated resource ²⁴. Significantly
237 higher Pearson correlation coefficients for pairs of proteins that are part of a complex further
238 supported the quantitative accuracy of our data matrix (**Fig. 2C**). We applied our
239 computational pipeline for analyzing co-expression of protein complex numbers ²⁴ to the
240 NCI-60 proteotype data and confirmed conserved stoichiometry of protein complexes such as
241 the prefoldin and MCM complexes in various cell lines (**Fig. 2D**). In a specific case, we
242 observed a high correlation between the protein expression of XRCC6/Ku70 and
243 XRCC5/Ku80, a critical heterodimer involved in DNA repair and responsible for resistance to
244 radiotherapy and chemotherapy. Ku80 is degraded when not bound to Ku70 ^{25,26}.
245 Remarkably, this correlation is not detectable using mRNA measurements (**Fig. 2E**),
246 indicating that expression of Ku80 is tightly regulated by protein degradation mechanisms
247 independent of cancer types. Indeed, a recent report has shown that RNF8, an E3 ubiquitin
248 ligase, regulates the expression of Ku80 via its removal from DNA double strand break sites
249 and its degradation through ubiquitination ²⁷.

250

251 **Google-map-based visualization of cancer signaling pathways**

252

253 The NCI-60 proteotypes cover 648 proteins in the Atlas of Cancer Signaling Networks
254 (ACSN), a manually curated pathway database presenting published facts about biochemical
255 reactions involved in cancer using a Google-Maps-style visualization (**Supplementary Fig. 8**)
256 ²⁸. When mapping the mean protein expression per cancer type, we found that multiple
257 pathways in different cell types, including apoptosis, cell survival, motility and DNA repair
258 among others, displayed a similar pattern (**Supplementary File 2**), consistent with the fact

259 that the immortal cells retain cancer hallmarks after artificial culturing²⁹. An example of a
260 clear proteotypic pattern is the delta isoenzyme of protein kinase C, *i.e.* PRKCD, involved in
261 DNA repair and a drug target that has been tested in various cancers³⁰. It was reported to be
262 absent in four renal clear cell carcinoma lines³¹. In agreement, this protein stood out in our
263 visualization, with significantly lower protein expression in renal carcinoma, relative to the
264 average expression in the NCI-60 panel. We provided detailed instructions on how to navigate
265 through the atlas and explore protein abundance in each cancer cell line (see **Supplementary**
266 **File 2**).

267

268 We next compared the activity of cellular pathways using ROMA (Representation and
269 quantification Of Module Activities)³² (**Fig. 2F**), a gene-set-based quantification algorithm.
270 This approach revealed substantial diversity of pathway activity between different proteotypes
271 as evidenced by two-tailed *t*-tests of activity scores (P -value < 0.05). When mapping activity
272 scores onto ACSN, some tissue specificities were revealed, with particular cell line
273 proteotypes displaying distinct patterns of pathway activity. For instance, the activity of
274 apoptosis (with both Caspases and Apoptosis Genes modules) was found to be significantly
275 higher in ovarian cell lines (see **Supplementary Table 5**). Although there are only two
276 prostate cancer cell lines in the panel, our analysis was able to highlight modules including
277 “AKT-mTOR” and “Apoptosis”, whose differential activity can be attributed to HSP90AA1
278 and PRDX. The latter protein has been independently reported to be overexpressed in prostate
279 tumors³³.

280

281 **Accessibility of the NCI-60 proteotypes**

282

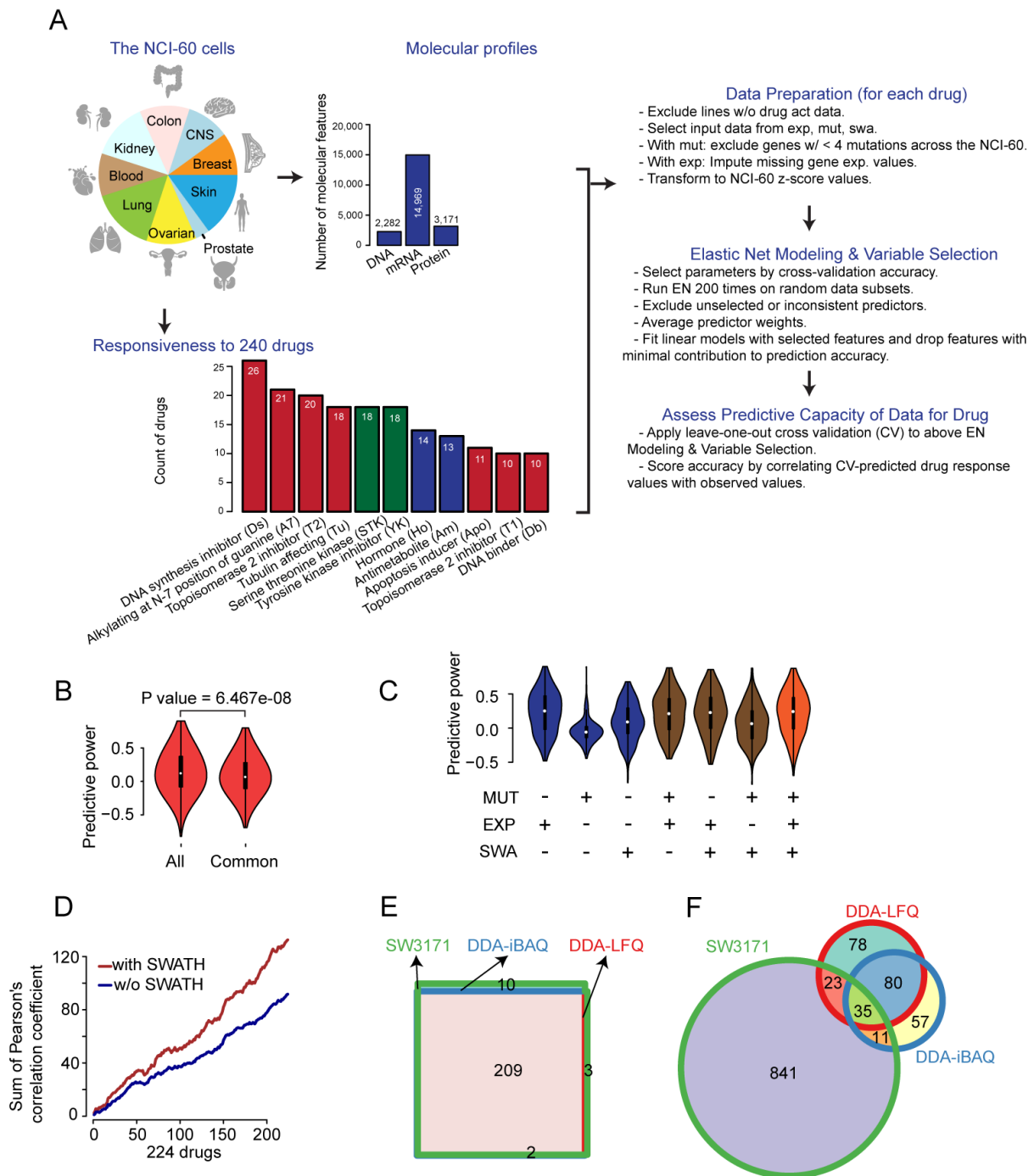
283 To enable easy data access, visualization, and comparison with other NCI-60 data sets,
284 we have incorporated the SWATH data into the CellMiner database^{13,34}. CellMiner allows
285 the direct download of the data, as well as comparative and integrative analyses with other
286 molecular data and pharmacological data, *e.g.* sensitivity of each cell line to over 20,000
287 compounds, and the manual inspection of specific genes, up to 150 per query. The detailed
288 instructions for using this resource are provided on the project website
289 (<https://discover.nci.nih.gov/cellminer/>) and in **Supplementary Fig. 9**. We have also
290 deposited raw data and processed data matrices of the NCI-60 proteotype in public databases,
291 including PRIDE³⁵ and ExpressionArray³⁶.

292

293 **Predicting drug responsiveness**

294

295 The robust, quantitative proteomic data, with almost no missing values, permitted
296 systematic investigation of whether integration of the SWATH-based proteotype with existing
297 genomic and transcriptomic features improves the prediction of drug responsiveness
298 (**Supplementary Table 6**). We generated various combinations of molecular features, and
299 evaluated their predictive power using the Pearson correlation between predicted and
300 observed drug response values for 240 FDA-approved or investigational compounds in
301 CellMiner^{13, 34, 37}. Each compound is assigned a NSC (National Service Center) identifier
302 upon submission to the National Cancer Institute for evaluation in the NCI-60 panel. The
303 largest groups of drugs with target annotations are those that interfere with DNA synthesis
304 and the DNA damage response, including topoisomerase inhibitors. The drug set also contains
305 dozens of targeted agents, including 18 serine and threonine kinase inhibitors and 18 tyrosine
306 kinase inhibitors (**Fig. 3A**).



307

308

309 **Figure 3. Prediction of drug responsiveness.** (A) Workflow for drug responsiveness prediction. Drug groups

310 with at least ten drugs are shown. (B) Distribution of predictive power (Pearson's correlation of cross-validation

311 predicted vs. observed response) for 240 compounds using all molecular features (All) versus common features

312 (Common) available for all molecular data types. (C) Distribution of predictive power for different molecular

313 data sets and their combinations. (D) Cumulative sum of Pearson correlation coefficients from drug

314 responsiveness prediction in 224 drugs. (E) Venn diagram of drugs successfully modeled using elastic net using

315 the SWATH data containing 3171 proteins (SW3171), and the DDA data based on iBAQ (DDA-iBAQ) and LFQ

316 (DDA-LFQ). (F) Venn diagram of protein predictors using the SWATH and DDA data sets.

317

318 Using the elastic net algorithm, we then developed multivariate linear models to
319 predict the NCI-60 response for each compound based on genomic, transcriptomic and
320 proteomic features. The Pearson's correlation between observed drug response values and
321 leave-one-out cross validation-predicted response values was applied to evaluate the
322 performance of each predictive model.

323

324 As different numbers of features were measured for each omics data set, two strategies
325 were adopted in the modeling analyses. First, we used all omics features (2,282 DNA
326 mutations, 14,969 mRNAs and 3,171 proteins), separately and in combination, as inputs to
327 evaluate the general performance. Second, we selected 1,566 features that were available for
328 all three molecular data types (denoted as common features). In both cases, we obtained valid
329 models for 224 (93%) of the drugs. The predictive power achieved with all features was
330 slightly higher than that obtained using the common features for all three data types (**Fig. 3A**);
331 a likely reason for this is that the latter excluded some genomic and transcriptomic features
332 not detected at the protein level. We accordingly derived our main analysis results from data
333 including all available molecular features. Our modeling led to the discovery of valid
334 biomarkers for drug responsiveness prediction. For instance, we found that the mRNA
335 expression of SLFN11, strikingly responsible for the sensitivity of 45 compounds, out of
336 which 39 were FDA-approved drugs including topoisomerase inhibitors, alkylating agents,
337 and DNA synthesis inhibitors, was the most dominant indicator, in agreement with our
338 previous report³⁸ (**Supplementary Table 7**). Fourteen ATP-binding cassette family
339 transporters, detected as mutation, transcript or protein levels, were found responsible for
340 sensitivity prediction of 51 compounds including chemotherapeutic agents and protein-
341 targeting agents such as HDAC inhibitor Depsipeptide, HSP90 inhibitor Alvespimycin,
342 mTOR inhibitor Temsirolimus and BCR-ABL inhibitor Nilotinib (**Supplementary Table 7**).

343

344 For ease of reproducibility of data analysis, we developed a Docker container
345 (described in **Methods**) that includes our code and other essential dependencies, allowing all
346 analyses to be replicated and extended for this and other omics data sets.

347

348

349 **Synergies among mutations, transcripts and proteins**

350

351 Our pipeline led to the identification of valid models for 224 compounds
352 (**Supplementary Table 7**). Given the relatively small sample size, it was not surprising that
353 accurate predictive models could not be found for every drug, particularly those with limited
354 numbers of responsive lines. We found that the SWATH-MS derived proteotypes displayed a
355 higher percentage of predictive features than mutations and transcripts. 1,090 (34%) out of
356 3,171 SWATH features are predictive, while 284 (12%) out of 2,282 features for mutations
357 and 1,976 (13%) out of 14,969 transcripts were selected in the models. In general, the
358 SWATH data outperformed the mutation data, however, the mRNA expression data set has
359 about a five to six-fold higher number of features than the protein and mutation data sets (**Fig.**
360 **3A**) and exhibited better overall performance (**Fig. 3C**).

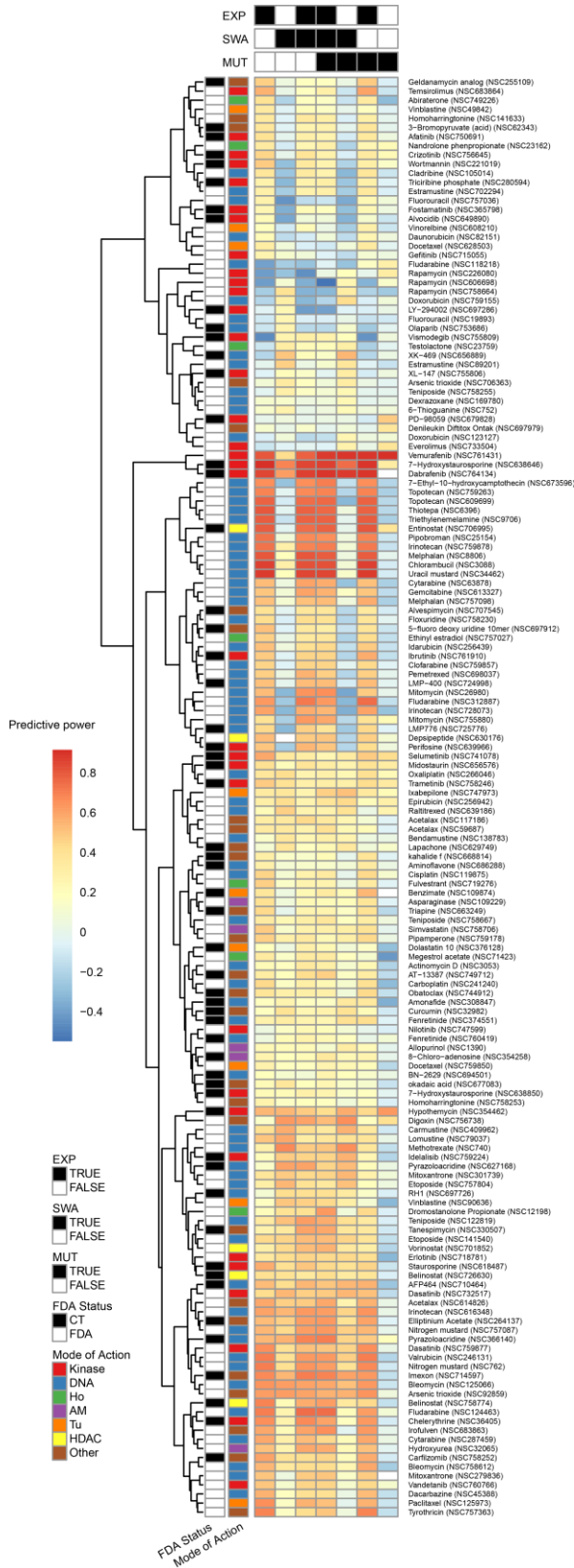
361
362 Our analyses revealed notable synergies among the different molecular measurements.
363 Each type of molecular data set demonstrated indispensable benefits in predicting the
364 response to certain drugs/compounds. The responsiveness of 35 compounds (16%) out of 224
365 was best predicted with SWATH data, whereas 107 compounds (48%) were best predicted by
366 SWATH data or by combining SWATH data with transcripts and/or DNA data. The most
367 accurate models for over half of the compounds required at least two different types of
368 molecular features. We then computed accumulative sum of Pearson correlation coefficient
369 based on drug responsiveness prediction and observed significant contribution of SWATH
370 data (**Fig. 3D**). We also compared the predictive power of the DDA data to the SWATH data.
371 While the DDA data were able to generate elastic net models for comparable number of drugs
372 (**Fig. 3E**), the number of protein predictors is much lower than SWATH data over some
373 overlap (**Fig. 3F**), indicating a higher degree of information content and robustness of the
374 signatures achieved with the SWATH data.

375 376 **Drug responsiveness prediction**

377
378 Based on the integration of various data sets, global drug response patterns were
379 predicted for the 158 well-modeled drugs (**Fig. 4**, see Methods), with predictive molecular
380 features for individual compounds provided in **Supplementary Table 7**. The data generated
381 from this computational pipeline were validated by the recovery of established
382 pharmacogenomic knowledge. For instance, the mutational status of BRAF was the top
383 predictive molecular feature for sensitivity to BRAF inhibitors, *e.g.* vemurafenib (NSC
384 761431) and dabrafenib (NSC 764134), and this association was particularly evident in

385 melanomas. Activated BRAF mutational status also sensitized cells to the MEK inhibitor
 386 hypothemycin (NSC: 354462), as has been previously described³⁹.

387
 388



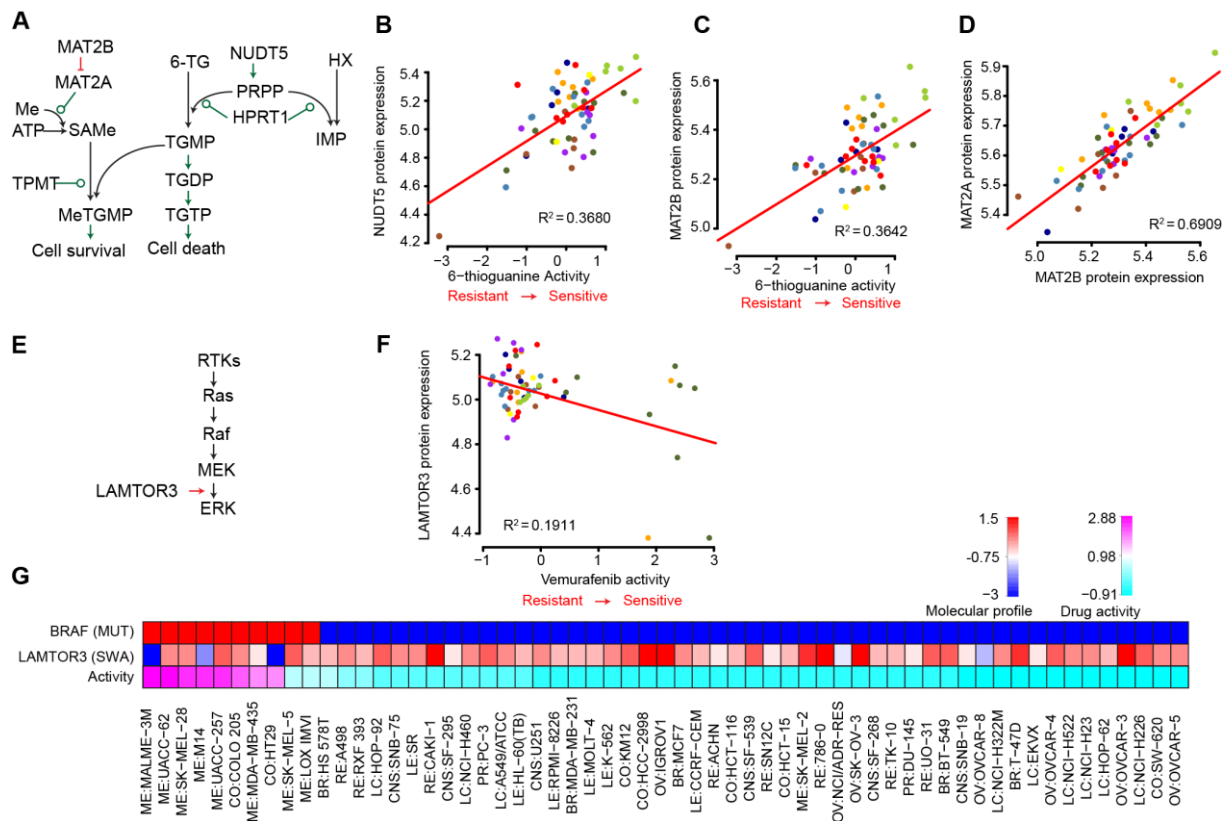
389

390 **Figure 4.** Predictive power for 224 compounds using different types of omics data. We applied elastic net and
391 cross validation to evaluate the drug response predictive accuracy for each omics data set and combinations of
392 data sets for 224 drugs which could be effectively modeled. Drug response prediction accuracies across input
393 data types are clustered without supervision. MoA of compounds and clinical status of the compounds are
394 colored. Each column indicates an input data type or combination of types; each row represents a compound.
395 The color indicates the predictive power measured by Pearson correlation of cross-validation predicted versus
396 observed drug response values. Black indicates that a valid elastic net model could not be obtained.

397

398 Sensitivity to the antimetabolite 6-thioguanine (6-TG, NSC: 752) (**Fig. 5A**) was
399 predicted by expression levels of proteins NUDT5 and MAT2B within an elastic net model
400 composed of 5 proteomic features: NUDT5, MAT2B, CD47, STX12 and GFAP. The cross-
401 validation accuracy with this compound and the SWATH-MS data was relatively low ($r =$
402 0.27), probably due to instability in the selected predictive features with limited sample size.
403 Still, we find that for the two strongest predictors in the model, NUDT5 and MAT2B, the
404 expression data were significantly correlated with the activity of 6-TG (Fig. 5B and 5C).
405 Additionally, we were able to relate the inter-connected activities of these two proteins to the
406 mechanism of action for 6-TG. In the purine salvage pathway, HPRT1 catalyzes synthesis of
407 inosine monophosphate from hypoxanthine and phosphoribosyl pyrophosphate (PRPP), with
408 production of the latter stimulated by NUDT5. 6-TG can substitute for hypoxanthine,
409 ultimately yielding altered nucleotides that are toxic upon incorporation into DNA⁴⁰. PRPP is
410 still required, so low NUDT5 expression could possibly induce 6-TG resistance. This is
411 consistent with our NCI-60 data and recent experimental work showing that depletion of
412 NUDT5 confers resistance to 6-TG⁴¹. As noted in **Fig. 5A**, a metabolite of 6-TG,
413 thioguanosine monophosphate (TGMP) can be inactivated by methylation. Production of the
414 methyl group donor, S-adenosylmethionine (SAME), is catalyzed by the methionine
415 adenosyltransferase II α (MAT2A) enzyme. The MAT2B protein, exhibiting high correlation
416 with MAT2A (**Fig. 5D**), is a regulatory component of MAT which may enhance feedback
417 inhibition by SAME⁴². Increased MAT inhibition and diminished TGMP methylation may
418 shunt more TGMP toward DNA incorporation, enhancing the 6-TG response. In spite of its
419 relatively low cross-validation accuracy, the presented model may provide a starting point for
420 further exploration, in light of the supporting prior research.

421



422

423 **Figure 5. Drug responsiveness predicted by SWATH data. (A)** molecular mechanisms of 6TG. **(B)**
 424 correlation between NUDT5 protein expression and 6-TG activity. **(C)** correlation between MAT2B protein
 425 expression and 6-TG activity. **(D)** correlation between MAT2B and MAT2A protein expression. **(E)**
 426 LAMTOR3 facilitates MEK/ERK pathway activation by binding MEK and ERK. **(F)** correlation between
 427 LAMTOR3 protein expression and Vemurafenib activity. **(G)** Association of BRAF mutation and LAMTOR3
 428 protein expression with Vemurafenib activity.

429

430 Analysis of the protein kinase inhibitor vemurafenib (NSC 761431) yielded a
 431 multivariate model based on BRAF V600E activating mutation status⁴³ and the protein
 432 expression level of LAMTOR3. LAMTOR3 (MP1) is part of an endosomal scaffolding
 433 complex that interacts with components of the RAF/MEK/ERK mitogenic signaling pathway
 434 (**Fig. 5E**). In particular, LAMTOR3 binds MEK1 and ERK1, facilitating activation of the
 435 latter protein⁴⁴. Elevated LAMTOR3 protein expression was correlated with vemurafenib
 436 resistance ($r = 0.44$, **Fig. 5F**), consistent with the hypothesis that LAMTOR3 has the capacity
 437 to enhance RAF/MEK/ERK pathway signaling downstream from RAF. In particular,
 438 increased protein expression of LAMTOR3 was observed in two BRAF mutant cell lines,
 439 ME:SK-MEL-5 and ME:LOXIMVI, which are relatively resistant to Vemurafenib (**Fig. 5G**).
 440 Due to the limited number of cell lines in the NCI-60 compendium that contained BRAF
 441 mutations exhibiting relative drug resistance (*i.e.* 2 cell lines), additional statistical analyses
 442 with sufficient power were not possible. Robust statistical validation of this model may be

443 possible when larger cell line databases (e.g. the Sanger and Broad resources) expand to
444 include proteomic coverage of LAMTOR3. Still, this finding remains relevant in light of the
445 recent research into the activity of LAMTOR3, including the observation that reduced
446 LAMTOR3 protein levels decreased the activation of MEK1/2 and ERK1/2 ^{44, 45}.
447 Additionally, LAMTOR3 has been shown to affect proliferation of pancreatic and breast
448 cancers ^{46, 47}, and has been patented as a diagnostic biomarker for breast cancer ⁴⁷.

449

450 Our elastic net analysis also produced multiple recurrent predictors with plausible drug
451 response associations. ABCC4 was a negatively weighted predictor for several alkylating
452 agents, including chlorambucil (NSC: 3088), uracil mustard (NSC: 34462), nitrogen mustard
453 (NSC: 762), consistent with its established role as a drug efflux pump ⁴⁸. Another recurrent,
454 negatively-weighted predictor was CTNND1, which was identified for several compounds,
455 including bendamustine (NSC: 138783), etoposide (NSC: 141540), valrubicin (NSC:
456 246131), and carmustine (NSC: 409962). CTNND1 encodes delta-catenin, whose
457 overexpression promotes cell survival through activation of Wnt pathway signaling ⁴⁹. The
458 resulting inhibition of apoptosis ⁵⁰ could plausibly confer resistance to the mentioned DNA-
459 damage inducing drugs.

460

461 **Discussion**

462

463 Due to the complementarity of protein and transcript data ^{4-6, 51}, it can be expected that
464 the rapid and consistent quantification of thousands of proteins across a large sample cohort
465 will reveal new biological information that is not apparent from the commonly used transcript
466 profiles. However, due to technical limitations, such proteomic cohort datasets have been
467 challenging to acquire. Here, using the NCI-60 cell line compendium, we demonstrate the
468 ability of the PCT-SWATH proteomic technique to consistently quantify in excess of 3000
469 proteins across the 60 cell lines measured in duplicate. The data were acquired in 30 working
470 days on a single mass spectrometer and for each sample measurement ca. 1 microgram of
471 total peptide mass was consumed. This has been enabled by the pressure cycling technology
472 which minimizes samples consumption and the data-independent MS data acquisition using
473 SWATH-MS ⁷. The data generated and their use to reveal cancer biology and drug response
474 determinants represent a significant advance in the field.

475

476 The proteome of the NCI-60 cells has been previously measured by extensive sample
477 fractionation and DDA-MS analysis of over 1,000 fractionated samples¹⁷. In this study, data
478 acquisition for each cell line required an average of about 29.16 hours MS instrument time.
479 That shotgun proteomics study reported the cumulative identification of 10,350 IPI proteins
480 over the NCI-60 cell lines. However, only 492 proteins were quantified in all cell lines
481 without missing value. The PCT-SWATH methodology adopted in this study offers an over
482 10-fold increase in sample-throughput, which has allowed us to acquire the proteotype for
483 each cell in the NCI-60 panel in duplicate, with standardized sample preparation, within 30
484 working days. In addition, our data have 0.1% amount of missing values at protein level
485 owing to the data acquisition strategy and improvements in bioinformatics analysis. This
486 study demonstrates that the human proteotype can be obtained with a throughput comparable
487 to genomic and transcriptomic analyses, though still at relatively lower coverage.

488

489 Two aspects of our workflow ensure robust and quantitatively accurate protein
490 expression measurements. First, we obtained technical duplicates for the entire set of NCI-60
491 proteotypes, which was feasible due to the unparalleled high sample-throughput of the PCT-
492 SWATH methodology which is now gaining popularity in proteomic profiling of clinical
493 specimens. In addition, we developed an expert system software (manuscript in preparation)
494 to further curate peptide and protein identification and quantification. Applying stringent
495 criteria, 3,171 proteins were included for further analyses. The raw MS signal for each of
496 these quantified proteins, in each cell line, was inspected by the expert system, simulating
497 manual inspection, and is available for visual inspection in the supplementary data. We further
498 compared the expression of a few proteins with known expression in certain cell lines,
499 obtaining good agreement. Nevertheless, we cannot conclude that the peptides and proteins
500 that failed to pass curation by the expert system are not biological signals, due to the
501 unpredictable degree of biological heterogeneity, and the fact that we did not analyze non-
502 canonical peptide variants and post-translational modification. The latter can be potentially
503 dissected and quantitated by future *in silico* analyses of our SWATH maps. Since the NCI-60
504 cell lines are widely used in cell biology, we anticipate broad utility of this highly curated
505 proteomic data. Additionally, our rapid proteotype acquisition pipeline using PCT-SWATH
506 requires little biological material, making it suitable for clinical settings and in precision
507 medicine efforts^{7, 8, 52}.

508

509 Compared to other omics data, the proteotypes obtained here offered unique insights
510 into the coordinated expression of protein complexes. Interactions amongst their component
511 subunits contribute to our understanding of protein function, as well as human diseases^{24, 53-}
512⁵⁵. Several protein complexes have been identified as biomarkers of disease, including cancer
513 progression⁵⁶. Our high quality proteomic data allowed systematic investigation of the
514 composition of 101 protein complexes in 60 cell lines. We expect that this represents a proof-
515 of-principle for a generic, high-throughput approach, applicable to clinical specimens⁷, for
516 exploring the association between protein complexes and biological/disease phenotypes.

517

518 The NCI-60 continues to enable important contributions that have come and continue
519 to come from this resource, and often emerging technologies are first tested on this cell line
520 panel due to its diversity and depth of surrounding knowledge^{3, 12, 57-59}. Each cancer cell line
521 in the NCI-60 has been tested against tens of thousands of compounds, including the 240
522 FDA-approved and investigational drugs featured in our analyses. With the addition of the
523 SWATH proteomic data, the NCI-60 remains positioned as one of most comprehensive
524 models for cancer research and drug discovery^{12, 15}. It uniquely enabled our thorough,
525 integrative analysis of different molecular profiles (genomic, transcriptomic, and proteomic)
526 in predicting drug responsiveness. Our findings strengthen the body of work highlighting the
527 importance of integrative omic approaches in understanding drug mechanisms and establish
528 the benefit of large-scale proteomic measurements. Therefore, we expect this work to become
529 a seminal work in the area of pharmacoproteomics, the benefit of which will grow with
530 anticipated expansion of sample size, proteomic coverage, including extension to
531 phosphoproteomic expression, as well as extension to mouse models⁶⁰ and human specimens
532⁷.

533

534 The existing SWATH data specifically enabled the use of advanced analysis
535 techniques to produce multivariate models of drug response. Great effort was put into making
536 our work accessible to a large audience through data submission to the NCI-60 CellMiner
537 database and availability through an accompanying R package, rcellminer. We expect this
538 pipeline based on the widely used elastic net method will continue to evolve and enable future
539 studies on additional data sets and phenotypes. And while the strengths of the elastic net
540 method over other related methods have been previously described^{61, 62}, the resulting models
541 still require careful scrutiny by individual researchers. The interpretation of the models
542 developed here, and by others using our pipeline, should be guided by understanding of the

543 biological activities of the associated predictors in the context of the mechanisms of action for
544 the input drugs. From the models generated by the current analyses, we identified several
545 potential determinants of drug responses, including NUDT5 and MAT2B protein levels for
546 the antimetabolite 6-TG, as well as complementary markers, such as LAMTOR3 protein
547 levels in conjunction with BRAF mutational status for Vemurafenib and other BRAF
548 inhibitors. These determinants may provide clinically relevant insights toward understanding
549 mechanisms of resistance to these and other agents. Together, these results invite further
550 investigation of this unique proteomic data resource. For example, the analysis of protein
551 complexes in the current study identified discrepancies between data at the transcriptomic and
552 proteomic levels. This observation has been similarly made in tumor samples, with additional
553 variation across tissue types⁶³. These differences can be used in future studies to develop
554 drug response models with non-redundant predictor sets including both data types. However,
555 due to the tissue diversity of the NCI-60 cells and the limited number of cell lines, data from a
556 higher number of cancer cell lines of specific tissue type and extension to clinical specimens
557 are required to advance our findings to clinical applications.

558

559 **Acknowledgements**

560 We thank Margot Sunshine who developed CellMiner and the NCI-DTP team (Dr.
561 Jerry Collins and Dr. James H. Doroshow) for the drug data and support for data acquisition,
562 Emanuel Gonçalves for comments to the manuscript. The work was supported by the
563 SystemsX.ch project PhosphoNetX PPM (to R.A.), the Swiss National Science Foundation
564 (grant no. 3100A0-688 107679 to R.A.), the European Research Council (grants no. ERC-
565 2008-AdG 233226 and ERC-20140AdG 670821 to R.A.), the Ruth L. Kirschstein National
566 Research Service Award (grant no. F32 CA192901 to A.L.), the National Resource for
567 Network Biology (NRNB) from the National Institute of General Medical Sciences (NIGMS)
568 (grant no. P41 GM103504 to C.S.), and the Center for Cancer Research, Intramural Program
569 of the National Cancer Institute (grant no. Z01 BC006150 to Y.P), and the Wellcome Trust
570 Award (102696) to M.J.G. We thank An Guo for help in preparing the graphics.

571

572 **Author contributions**

573 T.G. designed and coordinated the project with supervision from R.A. C.C.K.
574 processed the samples. L.G., C.C.K. and T.G. acquired the SWATH data. T.G. performed the
575 SWATH data interpretation and benchmarking with help from C.C.K., and the expert system
576 analysis with help from U.S. A.L., V.N.R. and Z.W. performed the drug response prediction

577 analysis, and developed the reproducible research infrastructure, with critical inputs from
578 M.P.M., J.S.R., M.J.G., S.V., W.C.R., C.S., and Y.P.. L.C. and L.M. performed the pathway
579 analysis. A.L., V.N.R., W.C.R. and S.V. integrated the SWATH data into rcellminer and
580 CellMiner. A.O., M.I. and R.C. performed the protein complex analysis, with help from A.L.,
581 Z.W., Y.C., V.N.R., C.S., Y.S., Y.Z., Y.P., P.Q. and Q.Z. contributed to the data analysis.
582 T.G., A.L. and V.N.R. wrote the manuscript with inputs from all co-authors. P.J.W., P.B.,
583 M.R., J.S.R., W.C.R., C.S., Y.P. and R.A. supervised the project.

584

585 **Competing financial interests**

586 R.A. holds shares of Biognosys AG, which operates in the field covered by the article.
587 The research group of R.A. is supported by SCIEX, which provides access to prototype
588 instrumentation, and Pressure Biosciences, which provides access to advanced sample
589 preparation instrumentation.

590

591

592 **Materials and Methods**

593

594 **PCT-assisted sample preparation for MS analyses**

595

596 The NCI-60 cells were obtained as frozen, non-viable cell pellets from the
597 Developmental Therapeutics Program (DTP), National Cancer Institute (NCI-NIH) and
598 processed using Barocycler® NEP2320 (PressureBioSciences Inc, South Easton, MA). The
599 IDs of the NCI-60 cells in our study matching to the IDs in Cellminer and a previous
600 proteomic study by the Kuster group are provided in **Supplementary Table 1**. Briefly, cell
601 pellets were lysed in a buffer containing 8M urea, 0.1M ammonium bicarbonate, and
602 Complete™ protease inhibitor using barocycler program (20 seconds 45 kpsi, 10 seconds 0
603 kpsi, 120 cycles) at 35°C⁷. Whole cell lysates were sonicated for 25 seconds with 1 min
604 interval on ice for 3 times. Cellular debris was removed by centrifugation and sample protein
605 concentration was determined by BCA assay prior to protein reduction with 10 mM TCEP for
606 20 min at 35°C, and alkylation with 40 mM iodoacetamide in the dark for 30 min at room
607 temperature. Lys-C digestion (1/50, w/w) was performed in 6 M urea using PCT program (25
608 seconds 25 kpsi, 10 seconds 0 kpsi 75 cycles) at 35°C; whereas trypsin digestion (1/30, w/w)
609 was performed in further diluted urea (1.6M) using PCT program (25 seconds 25 kpsi, 10
610 seconds 0 kpsi, 160 cycles) at 35°C. Digestion was stopped by acidification with
611 trifluoroacetic acid to a final pH of around 2 before C18 column desalting using SEP-PAK
612 C18 cartridges (Waters Corp., Milford, MA, USA).

613

614 **Off-gel electrophoresis**

615

616 To create a comprehensive spectral library for SWATH-MS analysis, we pooled 20-
617 40% of desalted peptide solutions from each NCI-60 sample and performed off-gel
618 fractionation. Briefly, pooled peptides were resolubilised in OGE buffer containing 5% (v/v)
619 glycerol, 0.7% (v/v) acetonitrile (ACN) and 1% (v/v) carrier ampholytes mixture (IPG buffer
620 pH 3.0-10.0, GE Healthcare). Fractionation was performed on a 3100 OFFGEL (OGE)
621 Fractionator (Agilent Technologies) using a 24 cm pH3-10 IPG strip (Immobilised pH
622 Gradient strip from GE Healthcare) according to manufacturer's instructions using a program
623 of 1 h rehydration at a maximum of 500 V, 50 µA and 200 mW followed by separation at a
624 maximum of 8000 V, 100 µA and 300 mW until 50 kVh were reached. Each of 24 fraction
625 was recovered and cleaned up by C18 reversed-phase MicroSpin columns (The Nest Group

626 Inc.). Based on the sample complexity (based on Nanodrop, A280 measurement), for each
627 strip, the following fractions were pooled into 12 samples for MS injections: pool 1 (fraction
628 1-2), pool 2 (fraction 3), pool 3 (fraction 4), pool 4 (fraction 5), pool 5 (fraction 6-7), pool 6
629 (fraction 8-9), pool 7 (fraction 10-11), pool 8 (fraction 12-15), pool 9 (fraction 16-19), pool 10
630 (fraction 20-21), pool 11 (fraction 22), pool 12 (fraction 23-24). Those were injected in
631 quadruplicate, resulting in 48 DDA injections of fractionated samples.

632

633 **DDA MS for spectral library generation**

634

635 For spectral library generation, a SCIEX TripleTOF 5600 System mass spectrometer
636 was operated essentially as described before⁶⁴: all samples were analyzed on an Eksigent
637 nanoLC (AS-2/1Dplus or AS-2/2Dplus) system coupled with a SWATH-MS-enabled AB
638 SCIEX TripleTOF 5600 System. The HPLC solvent system consisted of buffer A (2% ACN
639 and 0.1% formic acid, v/v) and buffer B (95% ACN with 0.1% formic acid, v/v). Samples
640 were separated in a 75 μm diameter PicoTip emitter (New Objective) packed with 20 cm of
641 Magic 3 μm , 200A C18 AQ material (Bischoff Chromatography). The loaded material was
642 eluted from the column at a flow rate of 300 nL min^{-1} with the following gradient: linear 2 -
643 35% B over 120 min, linear 35 - 90% B for 1 min, isocratic 90% B for 4 min, linear 90 - 2%
644 B for 1 min and isocratic 2% solvent B for 9 min. The mass spectrometer was operated in
645 DDA mode using a top20 method, with 500 ms and 150 ms acquisition time for the MS1 and
646 MS2 scans respectively, and 20 s dynamic exclusion for the fragmented precursors. Rolling
647 collision energy using the following equation ($0.0625 \times m/z - 3.5$) with a collision energy
648 spread of 15 eV was used for fragmentation regardless of the charge state of the precursors, to
649 mimic as close as possible the fragmentation conditions of the precursors in SWATH-MS
650 mode. Altogether, we had 66 DDA-MS injections, including the 48 OGE samples and another
651 18 pooled peptide samples from the unfractionated cell lysate of the NCI-60 cells.

652

653 **Spectral and assay library generation**

654

655 All raw instrument data were centroided using Proteowizard msconvert (version 2.0).
656 The assay library was generated using an established protocol⁶⁴. In short, the shotgun data
657 sets were searched individually using X!Tandem⁶⁵ (2011.12.01.1) with k-score plugin⁶⁶,
658 Myrimatch⁶⁷ (2.1.138), OMSSA⁶⁸ (2.1.8) and Comet⁶⁹ (2013.02r2) against the reviewed
659 UniProtKB/Swiss-Prot (2014_02) protein sequence database containing 20,270 proteins

660 appended with 11 iRT peptides and decoy sequences. Carbamidomethyl was used as a fixed
661 modification and oxidation as the variable modification. Maximally two missed cleavages
662 were allowed. Peptide mass tolerance was set to 50 ppm, fragment mass error to 0.1 Da. The
663 search identifications were combined and statistically scored using PeptideProphet ⁷⁰ and
664 iProphet ⁷¹ available within the Trans-Proteomics Pipeline (TPP) toolset (version 4.7.0) ⁷².
665 MAYU ⁷³ (v. 1.07) was used to determine the iProphet cutoff (0.999354) corresponding to a
666 protein FDR of 1.03%. SpectraST was used in library generation mode with CID-QTOF
667 settings and iRT normalisation at import against the iRT Kit ⁷⁴ peptide sequences (-
668 c_IRTirt.txt -c_IRR) and a consensus library was consecutively generated. An in-house
669 python script, spec-trast2tsv.py31 (msproteomicstools 0.2.2) was then used to generate the
670 assay library with the following settings: -l 350,2000 -s b,y -x 1,2 -o 6 -n 6 -p 0.05 -d -e -w
671 swath32.txt -k openswath (fragment ions between 350 and 2000 m/z, b and y ions authorized,
672 fragment charges 1+ and 2+, 6 most intense transitions, precision of fragment ion retrieved
673 0.05 Da, exact fragment ion mass calculated, exclude fragments in the swath window). The
674 OpenSWATH tool, ConvertTSVToTraML converted the TSV file to TraML format; Open-
675 SwathDecoyGenerator generated the decoy assays in shuffle mode and appended them to the
676 TraML assay library. In this study, we built a SWATH assay library containing 86,209
677 proteotypic peptide precursors in 8,056 proteotypic SwissProt proteins. This library is
678 supplied in PRIDE project PXD003539.

679

680 **SWATH-MS**

681

682 The SWATH-MS data acquisition in a Sciex TripleTOF 5600 mass spectrometer was
683 performed as described before ¹⁰, using 32 windows of 25 Da effective isolation width (with
684 an additional 1 Da overlap on the left side of the window) and with a dwell time of 100 ms to
685 cover the mass range of 400 - 1200 *m/z* in 3.3 s. The collision energy for each window was set
686 using the collision energy of a 2+ ion centered in the middle of the window (equation: 0.0625
687 $\times m/z - 3.5$) with a spread of 15 eV. The sequential precursor isolation window setup was as
688 follows: [400-425], [424-450], [449-475], ..., [1174-1200].

689

690 **Protein identification using OpenSWATH**

691

692 We analyzed the SWATH data using OpenSWATH software ¹¹ using parameters as
693 described previously ²⁴. We identified 48,374 peptides from 6,556 protein groups from the

694 NCI-60 panel with < 1% false discovery rate at both peptide and protein level evaluated by
695 OpenSWATH ¹¹ and Mayu ⁷⁵ (supplied in PRIDE project PXD003539).

696

697 **DIA-expert analyses**

698

699 The DIA-expert software read OpenSWATH output result file which contains
700 statistical scores (*i.e.* mProphet score or mScore) indicating the confidence of identification
701 for each peptide precursor in each sample, and from there selected the sample in which a
702 peptide precursor was identified with highest confidence. It then obtained extracted ion
703 chromatograms (XICs) for the target peptide precursor and all associated annotated *b* and *y*
704 fragments in the reference sample, and refined fragments based on the peak shape of each
705 fragment and its peak boundary. The refined fragments and precursor XIC traces from each of
706 the rest samples were subsequently compared with the reference peak group using empirical
707 expert rules, based on which the best matched peak group in each sample was picked and
708 visualized. Duplicated measurements were used to evaluate the accuracy of peptide and
709 protein quantification. The protein quantity was normalized based on total ion
710 chromatography of the MS1 spectra from each raw SWATH file. All codes are provided in
711 Github <https://github.com/tiannanguo/dia-expert>.

712

713 **Protein complexes analysis**

714

715 For this analysis, technical replicates were averaged to generate the NCI-60
716 proteotypes. To assess the coverage of protein complexes by NCI-60 proteotypes, we
717 retrieved a large resource of mammalian protein complexes assembled from CORUM ⁷⁶,
718 COMPLEAT ⁷⁷ and literature-curated complexes ^{24, 78}. This resource contains 2,041 proteins
719 as members of 279 distinct complexes and it is available at <http://variablecomplexes.embl.de/>.
720 101 complexes were represented in the NCI-60 proteotypes with at least 5 members
721 quantified. These complexes, in total, contain 1,045 distinct proteins quantified in the NCI-60
722 proteotypes. Pearson's correlation coefficient was calculated for all the pairwise comparisons
723 of 3,171 proteins across the NCI-60 cell lines. All pairwise comparisons were classified into
724 two categories: either two proteins were members of the same complex or not. Average
725 abundance, standard deviation and average Pearson correlation of each complex were
726 calculated based on the abundance of complex members in the NCI-60 proteotypes.

727

728 For this analysis, technical replicates were averaged to generate the NCI-60
729 proteotypes. To assess the coverage of protein complexes by NCI-60 proteotypes, we
730 retrieved a large resource of mammalian protein complexes assembled from CORUM ⁷⁶,
731 COMPLEAT ⁷⁷ and literature-curated complexes ^{24, 78}. This resource contains 2041 proteins
732 as members of 279 distinct complexes and it is available at <http://variablecomplexes.embl.de/>.
733 158 complexes were represented in the NCI-60 proteotypes with at least 5 members
734 quantified. These complexes, in total, contain 1,045 distinct proteins quantified in the NCI-60
735 proteotypes. Pearson's correlation coefficient was calculated for all the pairwise comparisons
736 of 3,171 proteins across the NCI-60 cell lines. All pairwise comparisons were classified into
737 two categories: either two proteins were members of the same complex or not. Average
738 abundance and standard deviation of each complex were calculated based on the mean
739 abundance of complex members in the NCI-60 proteotypes.

740

741 **Pathway activity analysis**

742

743 The activity of pathways, as they are described in ACSN, has been computed using
744 ROMA ³². Among all the modules defined in ACSN, only 11 show a significant dispersion
745 over the data set: AKT_MTOR, HR (Homologous Recombination), NER (nucleotide
746 Excision Repair), TNF response, Death Receptors regulators, Apoptosis, caspases, E2F3 and
747 E2F4 targets, HIF1 and cytoskeleton polarity. For these modules, the mean activity score for
748 each type of cancer cell lines was computed and mapped onto the atlas (from bright green for
749 low values to bright red for high values). To assess module differential activity between
750 proteotypes, we computed a *t*-test on the activity scores in cell lines of a cancer type versus
751 the activity of all other cancer cell lines. The definition of genes composing each module can
752 be found in <http://acsn.curie.fr>

753

754

755 **Drug sensitivity prediction using elastic net**

756

757 The elastic net regularized regression algorithm was applied to predict drug response
758 for 240 FDA-approved or investigational NSC-designated compounds. Some widely studied
759 drugs are represented by more than one NSC identifier, with each identifier associated with a
760 distinct compound sample and series of NCI-60 drug activity assays. For each compound, 7
761 combinations of input data were evaluated. These included NCI-60 mRNA expression, gene-

762 level mutation, and SWATH-MS protein expression, both alone and in all possible
763 combinations. mRNA expression data was available for 14,969 genes, and derived from
764 CellMiner , with missing values imputed using the impute.knn function (with default
765 parameters) of the Bioconductor impute package. Gene-level mutation profiles were
766 available for 2,282 genes, and were obtained from CellMiner exome sequencing data, with
767 values indicating the percent conversion to a variant form for the case of expected function-
768 impacting alterations (frameshift, nonsense, splice-sense, missense mutations by
769 SIFT/PolyPhen2 analysis). SWATH-MS based protein expression data was available for
770 3,171 proteins.

771

772 Elastic net analysis was done using the glmnet R package ⁷⁹. The elastic net analysis
773 was conducted using a multi-step pipeline involving cross-validations performed in a nested
774 manner. The “outer” cross-validation is a leave-one-out cross validation that is conducted
775 over all computational steps present in the “inner” pipeline, and it is used to validate model
776 performance. The “inner” cross-validation are conducted to select elastic net hyperparameters
777 (alpha and lambda) and for predictor set trimming, using data from a set of ~59 cell lines.

778

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

789

790 To obtain a robust estimate of performance on unseen data, leave-one-out cross-
791 validation was applied to the overall procedure as part of the “outer” pipeline. Specifically,
792 drug response for each cell line was predicted using an elastic net model derived using the
793 remaining held out data (and the steps outlined above). The vector of predicted response
794 values was then correlated with the actual response values, with the Pearson’s correlation

795 coefficient providing an estimate of the predictive value of the applied input data
796 combination. More details of the elastic net algorithm are provided in File S3.

797
798 Elastic net analysis was done using the rcellminerElasticNet R package
799 (https://bitbucket.org/cbio_mskcc/rcellminerelasticnet), which facilitates the application of the
800 glmnet R package (which provides the elastic net algorithm code) to data from the rcellminer
801 and rcellminerData packages⁸⁰. rcellminerElasticNet also provides utility functions for
802 summarizing and visualizing elastic net results.

803
804 Results for the elastic net analysis are available from this URL:
805 [https://discover.nci.nih.gov/cellminerreviewdata/swath_analysis/swathOutput_062316_all.tar.](https://discover.nci.nih.gov/cellminerreviewdata/swath_analysis/swathOutput_062316_all.tar.gz)
806 [gz](https://discover.nci.nih.gov/cellminerreviewdata/swath_analysis/swathOutput_062316_all.tar.gz). This compressed file contains results for the analysis run with all features and selected
807 common features. Each drug compound has three files for each combination of molecular
808 features used in a particular run of the elastic net algorithm: 1) a knitr report R Markdown
809 (.Rmd) file containing the code that was run, 2) an RData (.Rdata) file containing the results
810 of each elastic net run (see elasticNet() documentation in the rcellminerElasticNet package),
811 3) the rendered knitr report as a webpage (.html).

812
813 Beyond the knitr report containing code, the elastic net pipeline is made reproducible
814 using a Docker image. Docker (www.docker.com) is an emerging platform for conducting
815 reproducible research in the biomedical research community. All necessary software and
816 dependencies to run the described analysis have been embedded in the available Docker
817 container to provide readers an environment that runs on all major operating systems
818 (including Windows, OSX, and Linux), making Docker containers self-contained, portable,
819 and capable of performing at levels similar to the host system.

820
821 The Docker container is available at the Docker Hub repository: [cannin/swath](https://hub.docker.com/r/cannin/swath)
822 (<https://hub.docker.com/r/cannin/swath/>). Key dependencies installed, include: RStudio
823 Server (<https://www.rstudio.com/>), rcellminer/rcellminerData⁸⁰, and rcellminerElasticNet.
824 With these installed dependencies, readers have the opportunity to 1) re-run analysis for
825 specific drug compounds and modify the code in order to extend the analysis using RStudio
826 Server, a web-based version of the RStudio R editor, and 2) use an R Shiny app web-based
827 data explorer to further understand described results. Instructions on the usage of the Docker

828 container are located at the rcellminerElasticNet project page
829 (https://bitbucket.org/cbio_mskcc/rcellminerelasticnet).

830

831 **Data deposition**

832

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

835 Reviewer account details:

836 Username: reviewer15254@ebi.ac.uk

837 Password: dWdyptzf

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

841 Reviewer account details:

842 Username: Reviewer_E-PROT-2

843 Password: gdgywGco

844 The protein data matrix is also accessible in CellMiner website¹³ and R package
845 rcellminer³⁷.

846

847 **References**

848

- 849 1. Cancer Genome Atlas Research, N. et al. The Cancer Genome Atlas Pan-Cancer analysis
850 project. *Nat Genet* **45**, 1113-1120 (2013).
- 851 2. Barretina, J. et al. The Cancer Cell Line Encyclopedia enables predictive modelling of
852 anticancer drug sensitivity. *Nature* **483**, 603-607 (2012).
- 853 3. Garnett, M.J. et al. Systematic identification of genomic markers of drug sensitivity in cancer
854 cells. *Nature* **483**, 570-575 (2012).
- 855 4. Zhang, B. et al. Proteogenomic characterization of human colon and rectal cancer. *Nature*
856 **513**, 382-387 (2014).
- 857 5. Mertins, P. et al. Proteogenomics connects somatic mutations to signalling in breast cancer.
858 *Nature* **534**, 55-62 (2016).
- 859 6. Zhang, H. et al. Integrated Proteogenomic Characterization of Human High-Grade Serous
860 Ovarian Cancer. *Cell* (2016).
- 861 7. Guo, T. et al. Rapid mass spectrometric conversion of tissue biopsy samples into permanent
862 quantitative digital proteome maps. *Nat Med* (2015).
- 863 8. Shao, S. et al. Minimal sample requirement for highly multiplexed protein quantification in
864 cell lines and tissues by PCT-SWATH mass spectrometry. *Proteomics* (2015).
- 865 9. Powell, B.S., Lazarev, A.V., Carlson, G., Ivanov, A.R. & Rozak, D.A. Pressure cycling technology
866 in systems biology. *Methods Mol Biol* **881**, 27-62 (2012).
- 867 10. Gillet, L.C. et al. Targeted data extraction of the MS/MS spectra generated by data-
868 independent acquisition: a new concept for consistent and accurate proteome analysis.
869 *Molecular & cellular proteomics : MCP* **11**, O111 016717 (2012).
- 870 11. Rost, H.L. et al. OpenSWATH enables automated, targeted analysis of data-independent
871 acquisition MS data. *Nat Biotechnol* **32**, 219-223 (2014).
- 872 12. Shoemaker, R.H. The NCI60 human tumour cell line anticancer drug screen. *Nat Rev Cancer*
873 **6**, 813-823 (2006).
- 874 13. Reinhold, W.C. et al. CellMiner: A Web-Based Suite of Genomic and Pharmacologic Tools to
875 Explore Transcript and Drug Patterns in the NCI-60 Cell Line Set. *Cancer Res* **72**, 3499-3511
876 (2012).
- 877 14. Fojo, T. et al. Identification of non-cross-resistant platinum compounds with novel
878 cytotoxicity profiles using the NCI anticancer drug screen and clustered image map
879 visualizations. *Crit Rev Oncol Hematol* **53**, 25-34 (2005).
- 880 15. Holbeck, S.L., Collins, J.M. & Doroshow, J.H. Analysis of Food and Drug Administration-
881 approved anticancer agents in the NCI60 panel of human tumor cell lines. *Mol Cancer Ther* **9**,
882 1451-1460 (2010).
- 883 16. Bates, S.E. et al. Romidepsin in peripheral and cutaneous T-cell lymphoma: mechanistic
884 implications from clinical and correlative data. *Br J Haematol* **170**, 96-109 (2015).
- 885 17. Gholami, A.M. et al. Global Proteome Analysis of the NCI-60 Cell Line Panel. *Cell Rep* **4**, 609-
886 620 (2013).
- 887 18. Picotti, P. & Aebersold, R. Selected reaction monitoring-based proteomics: workflows,
888 potential, pitfalls and future directions. *Nat Methods* **9**, 555-566 (2012).
- 889 19. Law, V. et al. DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res* **42**,
890 D1091-1097 (2014).
- 891 20. Uhlen, M. et al. Proteomics. Tissue-based map of the human proteome. *Science* **347**,
892 1260419 (2015).
- 893 21. Lin, J.L. et al. MST4, a new Ste20-related kinase that mediates cell growth and transformation
894 via modulating ERK pathway. *Oncogene* **20**, 6559-6569 (2001).
- 895 22. Huang, C.L., Cha, S.K., Wang, H.R., Xie, J. & Cobb, M.H. WNKs: protein kinases with a unique
896 kinase domain. *Exp Mol Med* **39**, 565-573 (2007).

- 897 23. Xu, H. et al. Epidermal growth factor receptor (EGFR)-related protein inhibits multiple
898 members of the EGFR family in colon and breast cancer cells. *Mol Cancer Ther* **4**, 435-442
899 (2005).
- 900 24. Ori, A. et al. Spatiotemporal variation of mammalian protein complex stoichiometries.
901 *Genome Biol* **17**, 47 (2016).
- 902 25. Kanungo, J. Exogenously expressed human Ku70 stabilizes Ku80 in *Xenopus* oocytes and
903 induces heterologous DNA-PK catalytic activity. *Mol Cell Biochem* **338**, 291-298 (2010).
- 904 26. Chang, H.W. et al. Effect of beta-catenin silencing in overcoming radioresistance of head and
905 neck cancer cells by antagonizing the effects of AMPK on Ku70/Ku80. *Head Neck* **38 Suppl 1**,
906 E1909-1917 (2016).
- 907 27. Feng, L. & Chen, J. The E3 ligase RNF8 regulates KU80 removal and NHEJ repair. *Nat Struct*
908 *Mol Biol* **19**, 201-206 (2012).
- 909 28. Kuperstein, I. et al. Atlas of Cancer Signalling Network: a systems biology resource for
910 integrative analysis of cancer data with Google Maps. *Oncogenesis* **4**, e160 (2015).
- 911 29. Hanahan, D. & Weinberg, R.A. Hallmarks of cancer: the next generation. *Cell* **144**, 646-674
912 (2011).
- 913 30. Mackay, H.J. & Twelves, C.J. Targeting the protein kinase C family: are we there yet? *Nat Rev*
914 *Cancer* **7**, 554-562 (2007).
- 915 31. Engers, R. et al. Protein kinase C in human renal cell carcinomas: role in invasion and
916 differential isoenzyme expression. *Br J Cancer* **82**, 1063-1069 (2000).
- 917 32. Martignetti, L., Calzone, L., Bonnet, E., Barillot, E. & Zinovyev, A. ROMA: Representation and
918 Quantification of Module Activity from Target Expression Data. *Front Genet* **7**, 18 (2016).
- 919 33. Ummanni, R. et al. Peroxiredoxins 3 and 4 are overexpressed in prostate cancer tissue and
920 affect the proliferation of prostate cancer cells in vitro. *J Proteome Res* **11**, 2452-2466 (2012).
- 921 34. Shankavaram, U.T. et al. CellMiner: a relational database and query tool for the NCI-60
922 cancer cell lines. *BMC genomics* **10**, 277 (2009).
- 923 35. Jones, P. et al. PRIDE: a public repository of protein and peptide identifications for the
924 proteomics community. *Nucleic Acids Res* **34**, D659-663 (2006).
- 925 36. Brazma, A. et al. ArrayExpress--a public repository for microarray gene expression data at the
926 EBI. *Nucleic Acids Res* **31**, 68-71 (2003).
- 927 37. Luna, A. et al. rcellminer: exploring molecular profiles and drug response of the NCI-60 cell
928 lines in R. *Bioinformatics* (2015).
- 929 38. Zoppoli, G. et al. Putative DNA/RNA helicase Schlafen-11 (SLFN11) sensitizes cancer cells to
930 DNA-damaging agents. *Proc Natl Acad Sci U S A* **109**, 15030-15035 (2012).
- 931 39. Solit, D.B. et al. BRAF mutation predicts sensitivity to MEK inhibition. *Nature* **439**, 358-362
932 (2006).
- 933 40. de Boer, N.K., van Bodegraven, A.A., Jharap, B., de Graaf, P. & Mulder, C.J. Drug Insight:
934 pharmacology and toxicity of thiopurine therapy in patients with IBD. *Nat Clin Pract*
935 *Gastroenterol Hepatol* **4**, 686-694 (2007).
- 936 41. Doench, J.G. et al. Optimized sgRNA design to maximize activity and minimize off-target
937 effects of CRISPR-Cas9. *Nat Biotechnol* **34**, 184-191 (2016).
- 938 42. Halim, A.B., LeGros, L., Geller, A. & Kotb, M. Expression and functional interaction of the
939 catalytic and regulatory subunits of human methionine adenosyltransferase in mammalian
940 cells. *The Journal of biological chemistry* **274**, 29720-29725 (1999).
- 941 43. Flaherty, K.T. et al. Inhibition of mutated, activated BRAF in metastatic melanoma. *The New*
942 *England journal of medicine* **363**, 809-819 (2010).
- 943 44. Schaeffer, H.J. et al. MP1: a MEK binding partner that enhances enzymatic activation of the
944 MAP kinase cascade. *Science* **281**, 1668-1671 (1998).
- 945 45. Teis, D., Wunderlich, W. & Huber, L.A. Localization of the MP1-MAPK scaffold complex to
946 endosomes is mediated by p14 and required for signal transduction. *Dev Cell* **3**, 803-814
947 (2002).
- 948 46. Jun, S. et al. PAF-mediated MAPK signaling hyperactivation via LAMTOR3 induces pancreatic
949 tumorigenesis. *Cell Rep* **5**, 314-322 (2013).

- 950 47. De Araujo, M.E. et al. Polymorphisms in the gene regions of the adaptor complex
951 LAMTOR2/LAMTOR3 and their association with breast cancer risk. *PLoS One* **8**, e53768
952 (2013).
- 953 48. Borst, P. & Elferink, R.O. Mammalian ABC transporters in health and disease. *Annu Rev*
954 *Biochem* **71**, 537-592 (2002).
- 955 49. Tang, B. et al. Overexpression of CTNND1 in hepatocellular carcinoma promotes carcinous
956 characters through activation of Wnt/beta-catenin signaling. *J Exp Clin Cancer Res* **35**, 82
957 (2016).
- 958 50. Chen, S. et al. Wnt-1 signaling inhibits apoptosis by activating beta-catenin/T cell factor-
959 mediated transcription. *J Cell Biol* **152**, 87-96 (2001).
- 960 51. Liu, Y., Beyer, A. & Aebersold, R. On the Dependency of Cellular Protein Levels on mRNA
961 Abundance. *Cell* **165**, 535-550 (2016).
- 962 52. Shao, S. et al. Reproducible Tissue Homogenization and Protein Extraction for Quantitative
963 Proteomics Using MicroPestle-Assisted Pressure-Cycling Technology. *J Proteome Res* **15**,
964 1821-1829 (2016).
- 965 53. Dudley, A.M., Janse, D.M., Tanay, A., Shamir, R. & Church, G.M. A global view of pleiotropy
966 and phenotypically derived gene function in yeast. *Mol Syst Biol* **1**, 2005 0001 (2005).
- 967 54. Wang, Q. et al. Community of protein complexes impacts disease association. *Eur J Hum*
968 *Genet* **20**, 1162-1167 (2012).
- 969 55. Fraser, H.B. & Plotkin, J.B. Using protein complexes to predict phenotypic effects of gene
970 mutation. *Genome Biol* **8**, R252 (2007).
- 971 56. Le, D.H. A novel method for identifying disease associated protein complexes based on
972 functional similarity protein complex networks. *Algorithms Mol Biol* **10**, 14 (2015).
- 973 57. Barretina, J. et al. The Cancer Cell Line Encyclopedia enables predictive modelling of
974 anticancer drug sensitivity. *Nature* **483**, 603-607 (2012).
- 975 58. Weinstein, J.N. Drug discovery: Cell lines battle cancer. *Nature* **483**, 544-545 (2012).
- 976 59. Abaan, O.D. et al. The exomes of the NCI-60 panel: a genomic resource for cancer biology
977 and systems pharmacology. *Cancer Res* **73**, 4372-4382 (2013).
- 978 60. Gao, H. et al. High-throughput screening using patient-derived tumor xenografts to predict
979 clinical trial drug response. *Nat Med* **21**, 1318-1325 (2015).
- 980 61. Papillon-Cavanagh, S. et al. Comparison and validation of genomic predictors for anticancer
981 drug sensitivity. *J Am Med Inform Assoc* **20**, 597-602 (2013).
- 982 62. Jang, I.S., Neto, E.C., Guinney, J., Friend, S.H. & Margolin, A.A. Systematic assessment of
983 analytical methods for drug sensitivity prediction from cancer cell line data. *Pacific*
984 *Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 63-74 (2014).
- 985 63. Kostj, I., Jain, N., Aran, D., Butte, A.J. & Sirota, M. Cross-tissue Analysis of Gene and Protein
986 Expression in Normal and Cancer Tissues. *Sci Rep* **6**, 24799 (2016).
- 987 64. Schubert, O.T. et al. Building high-quality assay libraries for targeted analysis of SWATH MS
988 data. *Nat Protoc* **10**, 426-441 (2015).
- 989 65. Craig, R. & Beavis, R.C. A method for reducing the time required to match protein sequences
990 with tandem mass spectra. *Rapid Commun Mass Spectrom* **17**, 2310-2316 (2003).
- 991 66. MacLean, B., Eng, J.K., Beavis, R.C. & McIntosh, M. General framework for developing and
992 evaluating database scoring algorithms using the TANDEM search engine. *Bioinformatics* **22**,
993 2830-2832 (2006).
- 994 67. Tabb, D.L., Fernando, C.G. & Chambers, M.C. MyriMatch: highly accurate tandem mass
995 spectral peptide identification by multivariate hypergeometric analysis. *J Proteome Res* **6**,
996 654-661 (2007).
- 997 68. Geer, L.Y. et al. Open mass spectrometry search algorithm. *J Proteome Res* **3**, 958-964 (2004).
- 998 69. Eng, J.K., Jahan, T.A. & Hoopmann, M.R. Comet: an open-source MS/MS sequence database
999 search tool. *Proteomics* **13**, 22-24 (2013).
- 1000 70. Keller, A., Nesvizhskii, A.I., Kolker, E. & Aebersold, R. Empirical statistical model to estimate
1001 the accuracy of peptide identifications made by MS/MS and database search. *Anal Chem* **74**,
1002 5383-5392 (2002).

- 1003 71. Shteynberg, D. et al. iProphet: multi-level integrative analysis of shotgun proteomic data
1004 improves peptide and protein identification rates and error estimates. *Molecular & cellular*
1005 *proteomics : MCP* **10**, M111 007690 (2011).
- 1006 72. Keller, A., Eng, J., Zhang, N., Li, X.J. & Aebersold, R. A uniform proteomics MS/MS analysis
1007 platform utilizing open XML file formats. *Mol Syst Biol* **1**, 2005 0017 (2005).
- 1008 73. Reiter, L. et al. Protein identification false discovery rates for very large proteomics data sets
1009 generated by tandem mass spectrometry. *Molecular & cellular proteomics : MCP* **8**, 2405-
1010 2417 (2009).
- 1011 74. Escher, C. et al. Using iRT, a normalized retention time for more targeted measurement of
1012 peptides. *Proteomics* **12**, 1111-1121 (2012).
- 1013 75. Reiter, L. et al. Protein Identification False Discovery Rates for Very Large Proteomics Data
1014 Sets Generated by Tandem Mass Spectrometry. *Molecular & Cellular Proteomics* **8**, 2405-
1015 2417 (2009).
- 1016 76. Ruepp, A. et al. CORUM: the comprehensive resource of mammalian protein complexes--
1017 2009. *Nucleic Acids Res* **38**, D497-501 (2010).
- 1018 77. Vinayagam, A. et al. Protein complex-based analysis framework for high-throughput data
1019 sets. *Sci Signal* **6**, rs5 (2013).
- 1020 78. Ori, A. et al. Cell type-specific nuclear pores: a case in point for context-dependent
1021 stoichiometry of molecular machines. *Mol Syst Biol* **9**, 648 (2013).
- 1022 79. Friedman, J., Hastie, T. & Tibshirani, R. Regularization Paths for Generalized Linear Models via
1023 Coordinate Descent. *J Stat Softw* **33**, 1-22 (2010).
- 1024 80. Luna, A. et al. rcellminer: exploring molecular profiles and drug response of the NCI-60 cell
1025 lines in R. *Bioinformatics* **32**, 1272-1274 (2016).
- 1026