TCPA: a resource for cancer functional proteomics data

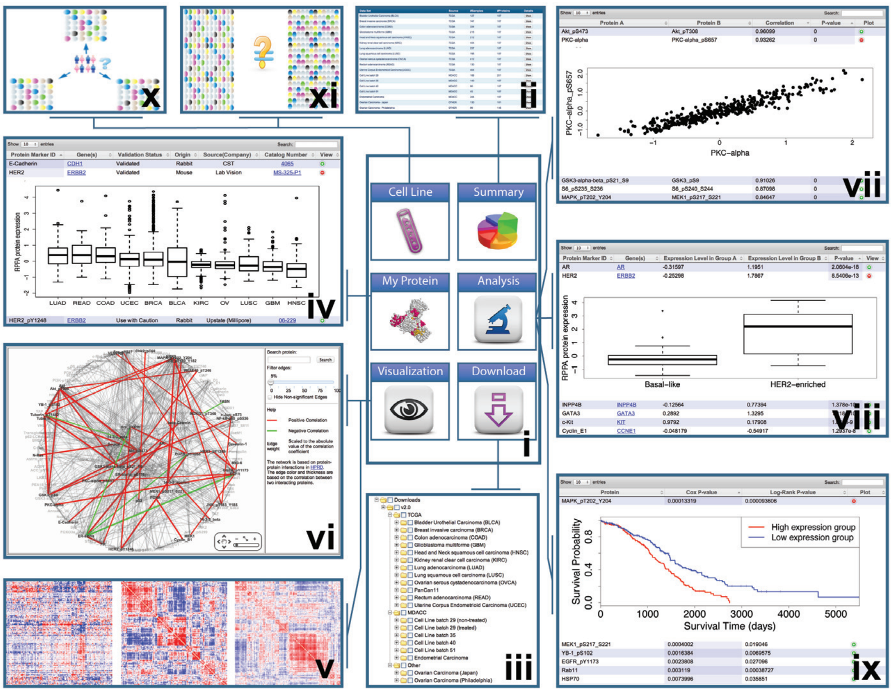
To the Editor: Functional proteomics represents a powerful approach to understand the pathophysiology and therapy of cancer. However, comprehensive cancer proteomic data have been relatively limited. As a part of The Cancer Genome Atlas (TCGA) Project and other efforts, we have generated protein expression data over a large number of tumor and cell line samples using reverse-phase protein arrays (RPPAs). RPPA is a quantitative, antibody-based technology that can assess multiple protein markers in many samples in a cost-effective, sensitive and high-throughput manner[1,2]. This technology has been extensively validated for both cell line and patient samples[3–5], and its applications range from building reproducible prognostic models[6] to generating experimentally verified mechanistic insights[7].

Our RPPA profiling platform includes extensively validated antibodies to nearly 200 proteins and phosphoproteins (**Supplementary Methods** and **Supplementary Table 1**). We are in the process of extending it to 500 independent proteins, covering all major signaling pathways, including PI3K, MAPK, mTOR, TGF-β, WNT, cell cycle, apoptosis, DNA damage, Hippo and Notch pathways. The current data release covers 4,379 tumor samples and consists of three parts (**Supplementary Table 2**). These are (i) TCGA tumor tissue sample sets: 3,467 samples from 11 cancer types, to be extended to 25 cancer types; (ii) independent tumor tissue sample sets: one endometrial tumor set (244 samples)[7] and two ovarian tumor sets (99 and 130 samples, respectively)[6], with other independent sets to be added soon; and (iii) tumor cell lines: 439 samples in four cell line sets, including both baseline and drug-treated cell lines. To our knowledge, this represents the largest publicly available collection of cancer functional proteomics data with parallel DNA and RNA data.

To facilitate broad access to these RPPA data sets, we developed a user-friendly data portal, The Cancer Proteome Atlas (TCPA; http://bioinformatics.mdanderson.org/main/TCPA:Overview). TCPA provides six modules: Summary, My Protein, Download, Visualization, Analysis and Cell Line (**Fig. 1**, **i**). The Summary module provides an overview of the RPPA data with detailed descriptions of each set (**Fig. 1**, **ii**). The Download module allows users to obtain any RPPA data set for analysis through a tree-view interface (**Fig. 1**, **iii**). The My Protein module provides detailed information about each RPPA protein: protein name, corresponding gene symbol, antibody status and source for the antibody. Users can examine the expression pattern of a protein of interest across different tumor types (for example, HER2 expression shown in **Fig. 1**, **iv**).

The Visualization module provides two ways to examine global protein expression patterns in a specific RPPA data set. One is through a "next-generation clustered heat map" (**Fig. 1**, **v**), which allows users to zoom, navigate and scrutinize clustering patterns of samples or proteins and link those patterns to relevant biological information sources. The other is through a network view (**Fig. 1**, **vi**), which overlays the correlation between any two interacting partners in the protein interaction network (curated in the Human Protein Reference Database[8]).

The Analysis module provides three analysis methods. (i) For correlation analysis, given a user-specified data set, correlations between any pair of proteins are presented in a table (**Fig. 1**, **vii**). Users can search the results by protein name, rank correlations or visualize the scatter plot of a correlation of interest (for example, there is a strong correlation between PKC-α and its phosphorylated



Data Set	Source	#Samples	#Proteins	Details
Bladder Urothelial Carcinoma (BLCA)	TCGA	127	187	Show
Breast invasive carcinoma (BRCA)	TCGA	747	187	Show
Colon adenocarcinoma (COAD)	TCGA	334	187	Show
Glioblastoma multiforme (GBM)	TCGA	215	187	Show
Head and Neck squamous cell carcinoma (HNSC)	TCGA	212	187	Show
Kidney renal clear cell carcinoma (KIRC)	TCGA	454	187	Show
Lung adenocarcinoma (LUAD)	TCGA	237	187	Show
Lung squamous cell carcinoma (LUSC)	TCGA	195	187	Show
Ovarian serous cystadenocarcinoma (OVCA)	TCGA	412	187	Show
Rectum adenocarcinoma (READ)	TCGA	130	187	Show
Uterine Corpus Endometrioid Carcinoma (UCEC)	TCGA	404	187	Show
Cell Line batch 29	MDACC	199	201	Show
Cell Line batch 35	MDACC	140	187	Show
Cell Line batch 40	MDACC	66	187	Show
Cell Line batch 51	MDACC	45	187	Show
Endometrial Carcinoma	MDACC	244	187	Show
Ovarian Carcinoma - Japan	OTHER	130	181	Show
Ovarian Carcinoma - Philadelphia	OTHER	99	148	Show

Show 10 entries Search:

Protein Marker ID	Gene(s)	Validation Status	Origin	Source(Company)	Catalog Number	View
E-Cadherin	CDH1	Validated	Rabbit	CST	4065	○
HER2	ERBB2	Validated	Mouse	Lab Vision	MS-325-P1	●
HER2_pY1248	ERBB2	Use with Caution	Rabbit	Upstate (Millipore)	06-229	○

Search protein:
Search
Filter edges:
5%
0 25 50 75 100
□ Hide Non-significant Edges
Help
Positive Correlation
Negative Correlation
Edge weight: Scaled to the absolute value of the correlation coefficient
The network is based on protein-protein interactions in HPRD. The edge color and thickness are based on the correlation between two interacting proteins.

- Downloads
  - v2.0
    - TCGA
      - Bladder Urothelial Carcinoma (BLCA)
      - Breast invasive carcinoma (BRCA)
      - Colon adenocarcinoma (COAD)
      - Glioblastoma multiforme (GBM)
      - Head and Neck squamous cell carcinoma (HNSC)
      - Kidney renal clear cell carcinoma (KIRC)
      - Lung adenocarcinoma (LUAD)
      - Lung squamous cell carcinoma (LUSC)
      - Ovarian serous cystadenocarcinoma (OVCA)
      - PanCan11
      - Rectum adenocarcinoma (READ)
      - Uterine Corpus Endometrioid Carcinoma (UCEC)
    - MDACC
      - Cell Line batch 29 (non-treated)
      - Cell Line batch 29 (treated)
      - Cell Line batch 35
      - Cell Line batch 40
      - Cell Line batch 51
      - Endometrial Carcinoma
    - Other
      - Ovarian Carcinoma (Japan)
      - Ovarian Carcinoma (Philadelphia)

Show 10 entries Search:

Protein A	Protein B	Correlation	P-value	Plot
Akt_pS473	Akt_pT308	0.96099	0	○
PKC-alpha	PKC-alpha_pS657	0.93262	0	●
GSK3-alpha-beta_pS21_S9	GSK3_pS9	0.91026	0	○
S6_pS235_S236	S6_pS240_S244	0.87098	0	○
MAPK_pT202_Y204	MEK1_pS217_S221	0.84647	0	○

Show 10 entries Search:

Protein Marker ID	Gene(s)	Expression Level in Group A	Expression Level in Group B	P-value	View
AR	AR	-0.31597	1.1951	2.0604e-18	○
HER2	ERBB2	-0.25298	1.7867	8.5406e-13	●
INPP4B	INPP4B	-0.12564	0.77394	1.378e-10	○
GATA3	GATA3	0.2892	1.3295	[illegible]	○
c-Kit	KIT	0.9792	0.17908	[illegible]	○
Cyclin_E1	CCNE1	-0.048179	-0.54917	1.2937e-8	○

Show 10 entries Search:

Protein	Cox P-value	Log-Rank P-value	Plot
MAPK_pT202_Y204	0.00013319	0.000093606	●
MEK1_pS217_S221	0.0004002	0.019046	○
YB-1_pS102	0.0016384	0.0069575	○
EGFR_pY1173	0.0023808	0.027096	○
Rab11	0.003119	0.00038727	○
HSP70	0.0073996	0.035851	○

Figure 1 | Overview of the TCPA data portal. TCPA contains six modules (**i**): the Summary module (**ii**); the Download module (**iii**); the My Protein module, which has a table view (**iv**); the Visualization module, which has a "next-generation clustered heat map" view (**v**) and network view (**vi**); the Analysis module, which offers correlation analysis (**vii**), differential analysis (**viii**) and survival analysis (**ix**); and the Cell Line module, which offers cell line–patient BLAST analysis (**x**) and drug treatment effect analysis (**xi**).

form PKC-a_pS657 in endometrial cancer, as shown in **Fig. 1**, **vii**). (ii) For differential analysis, differentially expressed protein markers between two tumor types or subtypes can be identified. Given user-defined comparison groups, the results are displayed in a table view, and for a protein of interest, users can visualize the box plots for the comparison (for example, the much higher expression of HER2 in the HER2-enriched subtype of breast cancer than in the basal-like subtype shown in **Fig. 1**, **viii**). (iii) For survival analysis, protein markers or pathway events significantly correlated with patient survival can be identified. The table view shows the univariate Cox proportional hazards model, log rank–test *P* values and a Kaplan-Meier plot for each protein in the data set (for example, phosphorylated MAPK, MEK, EGFR and YB are the top predictors of patient survival in ovarian cancer, which suggests a strong prognostic value of the tyrosine kinase receptor–RAS–MAPK pathway in this disease, as shown in **Fig. 1**, **ix**).

The Cell Line module provides two analyses for RPPA data from tumor cell lines. (i) For cell line–patient BLAST, cell lines with RPPA profiles that are most similar to those of a patient sample of interest can be selected (**Fig. 1**, **x**). The returned cell lines are externally linked with Cancer Cell Line Encyclopedia (CCLE)[9], from which selected mutations, transcriptomic profiles and sensitivity to specific drug treatments can be obtained. (ii) For drug treatment analysis, drug effects on RPPA profiles are provided (**Fig. 1**, **xi**).

Compared with other proteomic databases such as The Human Protein Atlas[10], an advantage of TCPA is the availability of quantitative protein expression data over large cohorts of well-characterized TCGA patient tumors, with linked DNA and RNA analyses. TCPA allows the validation of findings from TCGA RPPA data through independent sample cohorts and will help users select model tumor cell lines for further functional investigation. TCPA complements nucleic acid–centric cancer genomic data resources such as the CCLE, the Memorial Sloan-Kettering Cancer Center's cBioPortal for Cancer Genomics, OncoMine and the UCSC Cancer Genomics Browser. TCPA is also complementary to other protein-driven resources such as the Human Protein Reference Database, search tool for the retrieval of interacting genes/proteins (STRING) and Human Interactome Project. We will include additional data sets from TCGA and other independent cancer studies as they become available, and we will also accept (and help curate as necessary) cancer proteomic data from other groups.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper (doi:10.1038/nmeth.2650).

ACKNOWLEDGMENTS

We gratefully acknowledge contributions from the TCGA Research Network and its TCGA Pan-Cancer Analysis Working Group (contributing consortium members are listed in **Supplementary Note**). The TCGA Pan-Cancer Analysis Working Group is coordinated by J.M. Stuart, C. Sander and I. Shmulevich. This study was supported by the US National Institutes of Health (U24CA143883 to J.N.W. and G.B.M and P30CA016672); UTMDACC–G.S. Hogan Gastrointestinal Research Fund and NCI/UTMDACC Uterine SPORE Career Development Award (to H.L.); Susan G. Komen Foundation (KG081694 to G.B.M.); and Lorraine Dell Program in Bioinformatics for Personalization of Cancer Medicine (to J.N.W.).

AUTHOR CONTRIBUTIONS

G.B.M. and H.L. conceived of and supervised the project. Y.L., R.A., Z.J., W.L., J.-Y.Y., R.G.W.V. and J.L. generated the data, and J.L., P.L.R., B.M.B., D.W.K., C.W., J.N.W., G.B.M. and H.L. developed the data portal. Y.L., G.B.M. and H.L. wrote the manuscript with input from all the other authors.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Jun Li[1,4], Yiling Lu[2,4], Rehan Akbani[1], Zhenlin Ju[1], Paul L Roebuck[1], Wenbin Liu[1], Ji-Yeon Yang[1], Bradley M Broom[1], Roeland G W Verhaak[1], David W Kane[1,3], Chris Wakefield[1], John N Weinstein[1,2], Gordon B Mills[2] & Han Liang[1]

[1]Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, Texas, USA. [2]Department of Systems Biology, The University of Texas MD Anderson Cancer Center, Houston, Texas, USA. [3]SRA International, Inc., Fairfax, Virginia, USA. [4]These authors contributed equally to this work.
e-mail: gmills@mdanderson.org or hliang1@mdanderson.org

PUBLISHED ONLINE 15 SEPTEMBER 2013; DOI:10.1038/NMETH.2650

1. Sheehan, K.M. *et al. Mol. Cell. Proteomics* **4**, 346–355 (2005).
2. Spurrier, B., Ramalingam, S. & Nishizuka, S. *Nat. Protoc.* **3**, 1796–1808 (2008).
3. Tibes, R. *et al. Mol. Cancer Ther.* **5**, 2512–2521 (2006).
4. Hennessy, B.T. *et al. Clin. Proteomics* **6**, 129–151 (2010).
5. Nishizuka, S. *et al. Proc. Natl. Acad. Sci. USA* **100**, 14229–14234 (2003).
6. Yang, J.-Y. *et al. J. Clin. Invest.* doi:10.1172/JCI68509 (15 August 2013).
7. Liang, H. *et al. Genome Res.* **22**, 2120–2129 (2012).
8. Prasad, T.S.K. *et al. Nucleic Acids Res.* **37**, D767–D772 (2009).
9. Barretina, J. *et al. Nature* **483**, 603–607 (2012).
10. Uhlen, M. *et al. Nat. Biotechnol.* **28**, 1248–1250 (2010).