Cancer Research

# Explore, Visualize, and Analyze Functional Cancer Proteomic Data Using the Cancer Proteome Atlas

Jun Li[1], Rehan Akbani[1], Wei Zhao[2], Yiling Lu[2], John N. Weinstein[1,2], Gordon B. Mills[2], and Han Liang[1,2]

## Abstract

Reverse-phase protein arrays (RPPA) represent a powerful functional proteomic approach to elucidate cancer-related molecular mechanisms and to develop novel cancer therapies. To facilitate community-based investigation of the large-scale protein expression data generated by this platform, we have developed a user-friendly, open-access bioinformatic resource, The Cancer Proteome Atlas (TCPA, http://tcpaportal.org), which contains two separate web applications. The first one focuses on RPPA data of patient tumors, which contains >8,000 samples of 32 cancer types from The Cancer Genome Atlas and other independent patient cohorts. The second application focuses on the RPPA data of cancer cell lines and contains >650 independent cell lines across 19 lineages. Many of these cell lines have publicly available, high-quality DNA, RNA, and drug screening data. TCPA provides various analytic and visualization modules to help cancer researchers explore these datasets and generate testable hypotheses in an effective and intuitive manner. *Cancer Res; 77(21); e51–54. ©2017 AACR.*

## Introduction

Proteins comprise the basic functional units in biological processes and represent major targets for cancer therapy. However, large-scale cancer proteomic data have been relatively limited, in contrast to the recent explosion of next-generation sequencing data at both DNA and RNA levels. Importantly, the relatively low correlation of DNA and RNA levels with protein levels and with posttranslationally modified proteins, in particular, requires direct assessment of protein levels. Reverse-phase protein arrays (RPPA) represent a powerful functional proteomic approach that can assess a sizable number of selected protein markers across many samples in a cost-effective, sensitive, and high-throughput manner (1–3). This quantitative antibody-based assay has been widely used to investigate molecular events that drive tumor initiation/progression and to evaluate biomarkers and mechanisms that underlie sensitivity/resistance to cancer therapy (4, 5). More recently, we have employed this platform to characterize >8,000 patient samples through The Cancer Genome Atlas (TCGA; ref. 6), and the current RPPA platform contains approximately 300 protein markers, covering all major cancer signaling pathways. However, it remains a major informatic challenge for researchers to access and analyze RPPA data effectively, which limits the translation of these valuable functional proteomic data into clinical utility. To better serve the cancer research community, we have developed an open-access bioinformatic resource, The Cancer Proteome Atlas (TCPA). This resource is publicly available at http://tcpaportal.org, which substantially reduces the barrier to analyzing such complex RPPA data that biomedical researchers face.

## Materials and Methods

We generated the RPPA data through TCGA Research Network. TCPA web interface was implemented in JavaScript, tabular results were generated by DataTables, box and scatter plots were generated by HighCharts, and interactive network views were implemented by Cytoscape.js library.

## Results

### TCPA overview

As shown in Fig. 1A, the proteins profiled by the RPPA platform are first extracted from either patient tumor tissues or cultured cell lines, followed by serial dilutions. Then, the diluted proteins are arrayed on nitrocellulose-coated slides and probed with validated antibodies. Antibody validation is an ongoing process, with continuous reassessment of antibodies used on the platform as well as new antibodies. The data obtained from the arrays are further collected and normalized using a bioinformatic pipeline that consists of several cutting-edge algorithms for data normalization, curve fitting, and processing imaging data (7, 8). The standardized and normalized data are then curated and linked with other data in the TCPA portal.

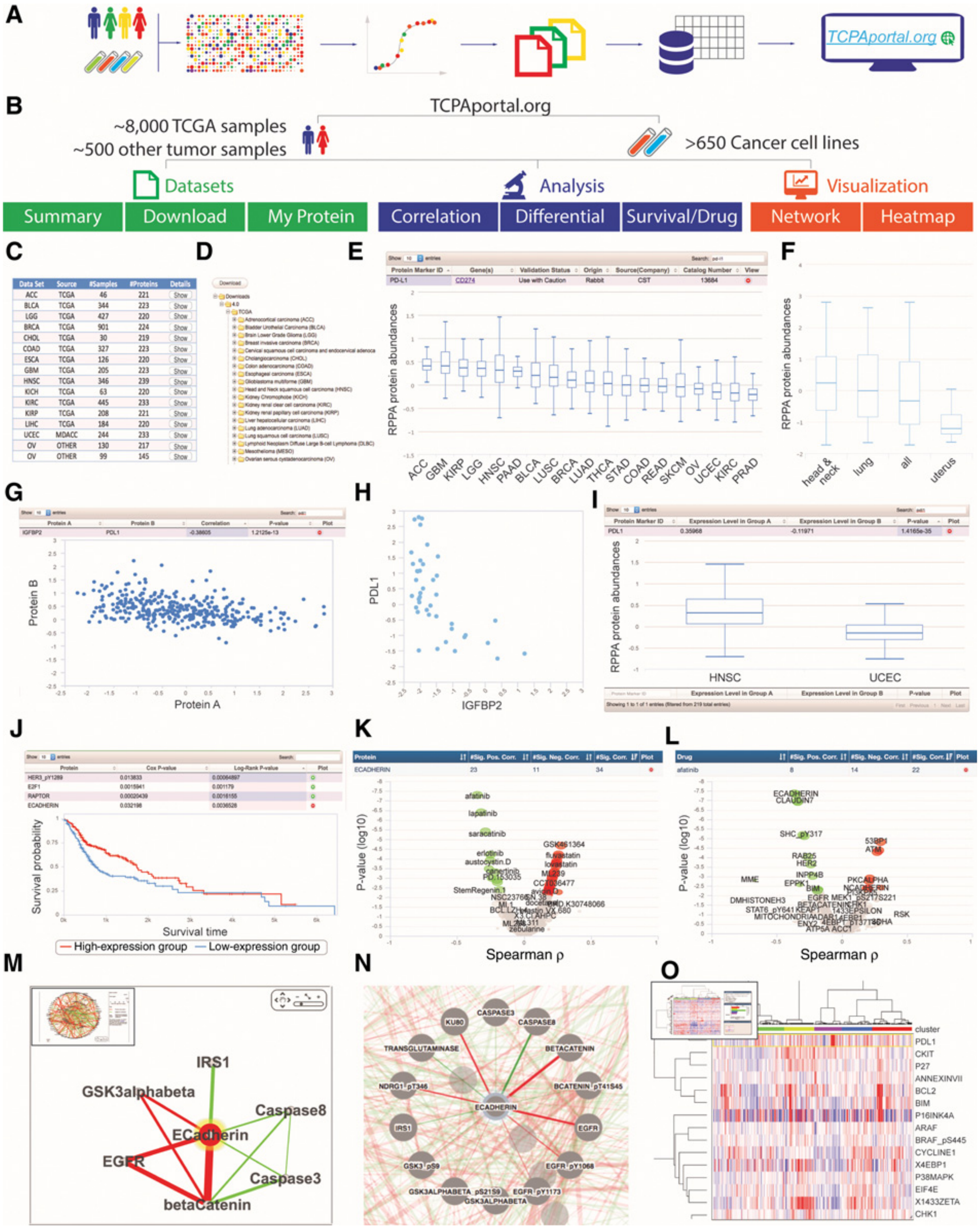
Currently, TCPA portal contains two separate web applications for exploring and analyzing the proteomic data. One focuses on patient cohort data (9), and the other is for cell line data. For the web application of patient cohort data, the latest release contains the data from approximately 8,000 samples of 32 TCGA cancer types and another approximately 500 samples from independent patient cohorts (Fig. 1B). For the web application of cell line data,

[1]Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, Texas. [2]Department of Systems Biology, The University of Texas MD Anderson Cancer Center, Houston, Texas.

**Note:** Supplementary data for this article are available at Cancer Research Online (http://cancerres.aacrjournals.org/).

**Corresponding Authors:** Han Liang, The University of Texas MD Anderson Cancer Center, 1400 Pressler Street, Houston, TX 77030. Phone: 713-745-9815; Fax: 713-563-4242; E-mail: hliang1@mdanderson.org; and Gordon B. Mills, gmills@mdanderson.org

**doi:** 10.1158/0008-5472.CAN-17-0369



| Data Set | Source | #Samples | #Proteins | Details |
|---|---|---|---|---|
| ACC | TCGA | 46 | 221 | Show |
| BLCA | TCGA | 344 | 223 | Show |
| LGG | TCGA | 427 | 220 | Show |
| BRCA | TCGA | 901 | 224 | Show |
| CHOL | TCGA | 30 | 219 | Show |
| COAD | TCGA | 327 | 223 | Show |
| ESCA | TCGA | 126 | 220 | Show |
| GBM | TCGA | 205 | 223 | Show |
| HNSC | TCGA | 346 | 239 | Show |
| KICH | TCGA | 63 | 220 | Show |
| KIRC | TCGA | 445 | 233 | Show |
| KIRP | TCGA | 208 | 221 | Show |
| LIHC | TCGA | 184 | 220 | Show |
| UCEC | MDACC | 244 | 233 | Show |
| OV | OTHER | 130 | 217 | Show |
| OV | OTHER | 99 | 145 | Show |

Download

- Downloads
  - 4.0
    - TCGA
      - Adrenocortical carcinoma (ACC)
      - Bladder Urothelial Carcinoma (BLCA)
      - Brain Lower Grade Glioma (LGG)
      - Breast invasive carcinoma (BRCA)
      - Cervical squamous cell carcinoma and endocervical adenoca
      - Cholangiocarcinoma (CHOL)
      - Colon adenocarcinoma (COAD)
      - Esophageal carcinoma (ESCA)
      - Glioblastoma multiforme (GBM)
      - Head and Neck squamous cell carcinoma (HNSC)
      - Kidney Chromophobe (KICH)
      - Kidney renal clear cell carcinoma (KIRC)
      - Kidney renal papillary cell carcinoma (KIRP)
      - Liver hepatocellular carcinoma (LIHC)
      - Lung adenocarcinoma (LUAD)
      - Lung squamous cell carcinoma (LUSC)
      - Lymphoid Neoplasm Diffuse Large B-cell Lymphoma (DLBC)
      - Mesothelioma (MESO)
      - Ovarian serous cystadenocarcinoma (OV)

Show 10 entries Search: pd-l1

| Protein Marker ID | Gene(s) | Validation Status | Origin | Source(Company) | Catalog Number | View |
|---|---|---|---|---|---|---|
| PD-L1 | CD274 | Use with Caution | Rabbit | CST | 13684 | ● |

Show 10 entries Search: pdl1

| Protein A | Protein B | Correlation | P-value | Plot |
|---|---|---|---|---|
| IGFBP2 | PDL1 | -0.38605 | 1.2125e-13 | ● |

Show 10 entries Search: pdl1

| Protein Marker ID | Expression Level in Group A | Expression Level in Group B | P-value | Plot |
|---|---|---|---|---|
| PDL1 | 0.35968 | -0.11971 | 1.4165e-35 | ● |

Showing 1 to 1 of 1 entries (filtered from 219 total entries)

Show 10 entries Search:

| Protein | Cox P-value | Log-Rank P-value | Plot |
|---|---|---|---|
| HER3_pY1289 | 0.013833 | 0.00064897 | ● |
| E2F1 | 0.0015941 | 0.001179 | ● |
| RAPTOR | 0.00020439 | 0.0016155 | ● |
| ECADHERIN | 0.032198 | 0.0036528 | ● |

| Protein | #Sig. Pos. Corr. | #Sig. Neg. Corr. | #Sig. Corr. | Plot |
|---|---|---|---|---|
| ECADHERIN | 23 | 11 | 34 | ● |

| Drug | #Sig. Pos. Corr. | #Sig. Neg. Corr. | #Sig. Corr. | Plot |
|---|---|---|---|---|
| afatinib | 8 | 14 | 22 | ● |

**Figure 1.**

Overview of The Cancer Proteome Atlas. **A,** Information flow of RPPA data. **B,** Overview of TCPA portal. **C,** "Summary" module. **D,** "Download" module. **E,** "My Protein" module for patient cohort data. **F,** "My Protein" module for cell line data. **G,** "Correlation Analysis" for patient cohort data. **H,** "Correlation Analysis" module for cell line data. **I,** "Differential Expression" module for patient cohort data. **J,** "Patient Survival" module. **K,** "Drug-Protein" module by protein. **L,** "Drug-Protein" module by drug. **M,** Zoomed-in view of "Network" visualization for patient cohort data, with a snapshot of the full view at the top left corner. **N,** Zoomed-in view of "Network" visualization for cell line data. **O,** "NG-CHM" visualization, with a snapshot of the full view at the top left corner.

the current release contains RPPA data of >650 independent cancer cell lines from the MD Anderson Cell Lines Project (10). Notably, the RPPA data of cell lines were generated by the same platform as the patient cohort data, providing independent sets for validating patterns observed in patient cohorts. TCPA routinely releases the latest RPPA data with a strict versioning approach and also allows for the efficient exploration of these datasets. Both patient cohort and cell line web applications provide three major types of modules: (i) "Datasets" modules provide the detailed information of samples and protein markers curated in TCPA, as well as a tree-view interface for data downloading; (ii) "Analysis" modules enable users to perform various common protein-centered analyses, including correlation, differential expression, patient survival, and drug sensitivity analyses; and (iii) "Visualization" modules allow users to examine the global RPPA data patterns interactively. The TCPA portal provides detailed online tutorials on how to use these modules (see Supplementary Video S1).

### "Datasets" modules

Users can examine sample information from the "Summary" module, such as data source, cancer type, number of samples, and number of antibodies (Fig. 1C). Moreover, by clicking the "show" button for each dataset, users can obtain more detailed information about how data were processed and normalized. Through the "Download" module, users can select datasets of interest in a tree-view menu (Fig. 1D). For each dataset, we provide both level 3 and level 4 data, with level 3 data representing the normalized data from independent batches (slides) and level 4 data representing merged data across multiple batches. These data have been precompressed in a zip file. Level 3 and level 4 data are equal in terms of analyzing the data generated in a single batch; level 4 data should be used when analyzing data across multiple batches. After downloading and decompressing the data, users can find the dataset metadata as well as the RPPA data matrix in the same folder. "My protein" module provides detailed information about each protein marker assessed, through which, users can obtain corresponding gene(s), validation status, antibody source, and catalog number. This module also allows for visualization of the protein expression level across different cancer types (Fig. 1E). For example, by clicking the green "plus" sign in the view column of the protein PD-L1, users can see the boxplots of the protein expression levels in 19 TCGA cancer types. To check whether the protein has a similar pattern across cell lines, one can go to the same module in the cell line application and search for PD-L1. The sorted boxplots in different cell line lineages show the same order as those in the patient tumors (median values from high to low): head and neck (HNSC), lung (LUSC and LUAD), and uterine (UCEC) cancers (Fig. 1F).

### "Analysis" modules

We implemented several commonly used approaches in TCPA, including correlation, differential expression, patient survival, and drug sensitivity analysis. For "Correlation Analysis," users can examine whether the expression levels of two proteins correlate with each other. Through this module, users can select an RPPA dataset by cancer type/lineage and obtain all pairwise correlations in a table view (Fig. 1G). The first two columns show protein pairs, followed by correlation coefficient and corresponding $P$ value. Furthermore, by clicking the plus sign, one can visualize the data in a scatter plot. For example, the plot shows that IGFBP2 and PD-L1 have a negative correlation in HNSC, and such a correlation can be also confirmed in cell line data (Fig. 1H). This independent validation increases the confidence of the patterns observed in patient cohorts. Another useful analysis is "Differential Expression." For patient tumor samples, users can identify which proteins are most differentially expressed between any two tumor types or subtypes. In Fig. 1I, PD-L1 is elevated in HNSC compared with UCEC ($t$ test $P = 1.42 \times 10^{-35}$). For cancer cell lines, users can identify which proteins are differentially expressed between cell lines with and without a mutated gene. In addition, the patient cohort application contains a unique module of "Survival Analysis," and the cell line application contains a unique module of "Protein-Drug Analysis." These two modules can be used to identify potential prognostic and predictive protein markers. As shown in Fig. 1J, the expression levels of several proteins significantly correlate with patient survival times. For example, patients with high ECADHERIN expression show better survival than those with low expression (11). To further check whether this protein has some predictive power for drug sensitivity, "Protein-Drug Analysis" can be used to assess which drugs correlate with this protein. For example, the expression of ECADHERIN shows negative correlations with sensitivity to afatinib, lapatinib, and saracatinib (Fig. 1K). Similarly, one can query by the drug to find which proteins show correlations with sensitivity to a specific drug (Fig. 1L).

### "Visualization" modules

TCPA provides two innovative visualization modules for exploring global patterns of RPPA datasets. One is "Network" visualization, in which protein markers are nodes and the interactions are colored edges linking two nodes. The red/green color represents a positive/negative correlation between two proteins. Users can search any protein using the search box to further focus on the neighbors of that protein. For example, Fig. 1M shows correlations between ECADHERIN and its neighbors. As in the other analytic modules, users can perform parallel analysis in cell line samples (Fig. 1N). Here, most of the correlations observed in Fig. 1M can be recaptured in Fig. 1N. The other module is "NG-CHM" visualization (Fig. 1O). This dynamic, interactive heatmap module enables users to visually check the global protein expression pattern for each cancer type or cell line lineage. In each heatmap, users can zoom in/out, check the expression level for each dot and related external resources, and obtain high-resolution figures for further analysis or publication.

## Discussion

We have developed an open-access, user-friendly bioinformatic resource, TCPA, which contains expression levels of key cancer proteins from approximately 10,000 samples of both patient tumors and cancer cell lines. TCPA is not only a one-stop data repository for obtaining high-quality RPPA data, but also a powerful web platform for making sense of these data through user-friendly, intuitive analyses. In the future, we will add more cancer RPPA data and integrate other genomic and clinical data into this resource. We expect that TCPA will continue to serve as a highly valuable resource that helps researchers generate testable hypotheses, validate findings of interest, and ultimately facilitate the development of novel cancer therapies.

## Disclosure of Potential Conflicts of Interest

G.B. Mills reports receiving commercial research grants from Adelson Medical Research Foundation, AstraZeneca, Breast Cancer Research Foundation, Critical Outcome Technologies, Komen Research Foundation, Nanostring, Karus, Illumina, Takeda/Millenium Pharmaceuticals, and Pfizer, has received speakers bureau honoraria from Allostery, AstraZeneca, ImmunoMet, ISIS Pharmaceuticals, Lilly, MedImmune, Novartis, Pfizer, Symphogen, and Tarveda, has ownership interest (including patents) in Catena Pharmaceuticals, ImmunoMet, Myriad Genetics, PTV Ventures, and Spindletop Ventures, and is a consultant/advisory board member for Adventist Health, Allostery, AstraZeneca, Catena Pharmaceuticals, Critical Outcome Technologies, ImmunoMet, ISIS Pharmaceuticals, Lilly, MedImmune, Novartis, Precision Medicine, Provista Diagnostics, Signalchem Lifesciences, Symphogen, Takeda/Millenium Pharmaceuticals, Tarveda, Tau Therapeutics. Han Liang is a consultant/advisory board member for INFOTECH Soft Inc. No potential conflicts of interest were disclosed by the other authors.

## Authors' Contributions

**Conception and design:** J. Li, G.B. Mills, H. Liang
**Development of methodology:** J. Li, R. Akbani, H. Liang
**Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.):** J. Li, Y. Lu, G.B. Mills, H. Liang
**Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis):** J. Li, R. Akbani, W. Zhao, G.B. Mills, H. Liang
**Writing, review, and/or revision of the manuscript:** J. Li, R. Akbani, G.B. Mills, H. Liang
**Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases):** J. Li, R. Akbani, J.N. Weinstein, H. Liang
**Study supervision:** Y. Lu, H. Liang
**Other (performed functional proteomics study by RPPA analysis):** Y. Lu

## Grant Support

This study was supported by the NIH (CA168394, CA098258, CA143883 and HG008100 to G.B. Mills; CA175486 to H. Liang; CA209851 to H. Liang and G.B. Mills; CA201949 and CA210950 to G.B. Mills, R. Akbani, and J.N. Weinstein; and the MD Anderson Cancer Center Support grant, CA016672), a grant from the Cancer Prevention and Research Institute of Texas (RP140462 to H. Liang), the Lorraine Dell Program in Bioinformatics for Personalization of Cancer Medicine (to J. N. Weinstein), and the Adelson Medical Research Foundation (to G.B. Mills).

Received February 7, 2017; revised March 9, 2017; accepted August 2, 2017; published online November 1, 2017.

## References

1. Hennessy BT, Lu Y, Gonzalez-Angulo AM, Carey MS, Myhre S, Ju Z, et al. A technical assessment of the utility of reverse phase protein arrays for the study of the functional proteome in non-microdissected human breast cancers. Clin Proteomics 2010;6:129–51.
2. Nishizuka S, Charboneau L, Young L, Major S, Reinhold WC, Waltham M, et al. Proteomic profiling of the NCI-60 cancer cell lines using new high-density reverse-phase lysate microarrays. Proc Natl Acad Sci U S A 2003;100:14229–34.
3. Tibes R, Qiu Y, Lu Y, Hennessy B, Andreeff M, Mills GB, et al. Reverse phase protein array: validation of a novel proteomic technology and utility for analysis of primary leukemia specimens and hematopoietic stem cells. Mol Cancer Ther 2006;5:2512–21.
4. Sheehan KM, Calvert VS, Kay EW, Lu Y, Fishman D, Espina V, et al. Use of reverse phase protein microarrays and reference standard development for molecular network analysis of metastatic ovarian carcinoma. Mol Cell Proteomics 2005;4:346–55.
5. Spurrier B, Ramalingam S, Nishizuka S. Reverse-phase protein lysate microarrays for cell signaling analysis. Nat Protoc 2008;3:1796–808.
6. Cancer Genome Atlas Research N, Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, et al. The Cancer Genome Atlas pan-cancer analysis project. Nat Genet 2013;45:1113–20.
7. Ju Z, Liu W, Roebuck PL, Siwak DR, Zhang N, Lu Y, et al. Development of a robust classifier for quality control of reverse-phase protein arrays. Bioinformatics 2015;31:912–8.
8. Akbani R, Ng PK, Werner HM, Shahmoradgoli M, Zhang F, Ju Z, et al. A pan-cancer proteomic perspective on the cancer genome atlas. Nat Commun 2014;5:3887.
9. Li J, Lu Y, Akbani R, Ju Z, Roebuck PL, Liu W, et al. TCPA: a resource for cancer functional proteomics data. Nat Methods 2013; 10:1046–7.
10. Li J, Zhao W, Akbani R, Liu W, Ju Z, Ling S, et al. Characterization of human cancer cell lines by reverse-phase protein arrays. Cancer Cell 2017;31:225–39.
11. Ren X, Wang J, Lin X, Wang X. E-cadherin expression and prognosis of head and neck squamous cell carcinoma: evidence from 19 published investigations. Onco Targets Ther 2016;9:2447–53.