TCPA: A Resource for Cancer Functional Proteomics Data

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

Supplementary Table 2 Summary of RPPA datasets currently available in TCPA (as of August 2013)

Cancer Type	Source	Sample No.	Protein No.
Bladder urothelial carcinoma (BLCA)	TCGA	127	187
Breast invasive carcinoma (BRCA)	TCGA	747	187
Colon adenocarcinoma (COAD)	TCGA	334	187
Glioblastoma multiforme (GBM)	TCGA	215	187
Head and neck squamous cell carcinoma (HNSC)	TCGA	212	187
Kidney renal clear cell carcinoma (KIRC)	TCGA	454	187
Lung adenocarcinoma (LUAD)	TCGA	237	187
Lung squamous cell carcinoma (LUSC)	TCGA	195	187
Ovarian serious cystadenocarcinoma (OV)	TCGA	412	187
Rectum adenocarcinoma (READ)	TCGA	130	187
Uterine Corpus Endometrioid carcinoma (UCEC)	TCGA	404	187
Endometrial carcinoma	MDACC	244	187
Ovarian carcinoma	Japan	130	181
Ovarian carcinoma	Philadelphia	99	145
Cell Line Set 29	MDACC	188	201
Cell Line Set 35	MDACC	140	187
Cell Line Set 40	MDACC	66	187
Cell Line Set 51	MDACC	45	187

Nature Methods: doi:10.1038/nmeth.2650

Supplementary Methods

Sample source

TCGA tumor samples were obtained from TCGA Biospecimen Core Resource. Independent endometrial tumor tissue samples were obtained from the tissue bank at the University of Texas MD Anderson Cancer Center, and the study was approved by the Institutional Review Board; and independent ovarian tumor cancer samples were obtained as described in Yang et al.[6] Tumor cell lines were obtained from multiple sources including the ATCC, the Korean Cell Line Collection and the originators of some cell lines.

Reverse phase protein arrays and data normalization

Cellular proteins were denatured by 1% SDS (with beta-mercaptoethanol) and diluted in five 2-fold serial dilutions in dilution buffer (lysis buffer containing 1% SDS). Serial diluted lysates were arrayed on nitrocellulose-coated slides (Grace Biolab) by Aushon 2470 Arrayer (Aushon BioSystems). A total of 5,808 array spots were arranged on each slide, including the spots corresponding to positive and negative controls prepared from mixed cell lysates or dilution buffer, respectively.

We assessed RPPA antibodies for specificity, quantification and sensitivity (dynamic range) using protein extracts from cultured cells or tumor tissues. Antibodies with a single or dominant band on Western blotting were assessed by direct comparison to RPPA using cell lines with differential protein expression or modulated with ligands/inhibitors or siRNA for phosphoproteins or structural proteins, respectively. Antibodies with a Pearson correlation coefficient, between RPPA and Western blotting, of greater than 0.7 were considered as "validated". Each slide was probed with a validated primary antibody plus a biotin-conjugated secondary antibody. Multiple replicates of "Control Lysates" on each slide served as a standard for "spatial correction" and "quality test". The QC score from “quality test” indicates good (above 0.8) or poor (below 0.8) antibody staining. Poorly stained slides were excluded from further data analysis.

The signal obtained was amplified using a Dako Cytomation–catalyzed system (Dako) and visualized by DAB colorimetric reaction. The slides were scanned, analyzed, and quantified using customized software (ArrayPro) to generate spot intensity.

Each dilution curve was fitted with a logistic model (“Supercurve Fitting” developed by the Department of Bioinformatics and Computational Biology at MD Anderson Cancer Center, “http://bioinformatics.mdanderson.org/OOMPA”). This fits a single curve using all the samples (i.e., dilution series) on a slide with the signal intensity as the response variable and the dilution steps as independent variables. The fitted curve is plotted with the signal intensities — both observed and fitted — on the *y*-axis and the log2-concentration of proteins on the *x*-axis for diagnostic purposes. The protein concentrations of each set of slides were then normalized by median polish, which was corrected across samples by the linear expression values using the median expression levels of all antibody experiments to calculate a loading correction factor for each sample.

Since RPPA is intrinsically a batch approach, each tumor type was run on a single batch to limit batch effects except the breast cancer set. We merged different batches of RPPA data using a

novel algorithm, called Replicates Based Normalization (RBN), which reduces batch effects. Using replicate samples that are common between batches, RBN adjusts the means and standard deviations of all the antibodies in the batches so that the means and standard deviations of the replicates become the same in all the batches. The underlying assumption is that any variation observed between replicates across batches is mainly due to systematic variations and should be canceled out. The number of common replicates varied between 69 to over 100. This method was used to combine the breast cancer set (over three separate batches) and to generate a combined TCGA Pan-Cancer dataset for cross-tumor comparison.

Web resource implementation

RPPA data and pre-calculated data are stored in CouchDB. Correlation, differential analyses, survival analyses, and cell-line/patient BLAST analyses were performed in R. Web interface was implemented by JavaScript; tables were visualized by DataTables; the embedded plots were based on HighCharts; and network visualization was implemented by Cytoscape Web.