pepper bio
DISCOVERY IS IN THE DATA

Take-home analysis question

Thank you for taking the time to delve into our analysis question. This exercise provides you with an example of some of the types of data and questions you would encounter as a member of Pepper's R&D team. We recommend spending **no more than 4 hours** investigating the data and responding to our question. Please use programming languages (e.g., Python), visualization tools, etc… of your choice. Upon completion, please send a single zip file containing:

- A single file that includes an overview of your approach, an answer to the questions below, and supporting information (e.g., figures):
  - Recommended text length is no more than 1 page (single spaced, size 10 font).
  - Feel welcome to append relevant figures / results as you see fit, up to 4 additional pages (these pages can include figure captions).
- Your code and/or analysis files.
- Please do not spend time on formatting either of the above.
  - We will look for your thought process, and not on your formatting choices

After you submit the above, we will schedule a follow-up call to discuss. Looking forward!

Question

Consider tumor and normal adjacent tissues (NATs) in Gillette, *et al.*, Cell, 2020. Does proteomic data provide different insights than transcriptomic data? Why or why not? Please use the proteomic and transcriptomic data provided in the tables (as labeled in column B). Note, when answering this question, consider details such as the following:

- What questions can we answer using the data?
- What do we do with the results of this analysis? E.g., experiment to follow-up on.

Notes

Note 1: There is a lot of data here! Please do not feel the need to take an exhaustive approach. You are invited to set bounds as you see fit to answer the question from a perspective you find interesting.
Note 2: Please only use the data provided.
Note 3: While the paper presents an analysis of this data, we are looking for your own analysis and assessment. Feel welcome to draw inspiration from the paper as you wish, but this is given primarily to provide clarity of the data source and context, however their approach, etc…. does not need to inform how you analyze the data to answer our question.
Note 4: If after the ~4 hours there are more ideas of analyses you would like to pursue in answer to the question, feel welcome to note the additional analyses of interest in your submitted response to the question.
Note 5: If your first approach is inconclusive, there is no need to spend more than the recommended 4 hours. Feel welcome to discuss what you learned and recommendations for what you would do if you continued the analysis.
Note 6: Have fun playing with the data – please know there is no right or wrong answer. We are excited to see and later discuss how you approach this question.

Paper and dataset

The dataset of interest is from Gillette, *et al.,* Cell, *182,* 200-225, July 9, 2020. For your reference, we've provided the following in the attachments in a compressed file. Manuscript: Main text and methods combined into one document

- Metadata (Table S1): Metadata associated with LUAD patients and tumors related to STAR methods
- Transcriptomic data (Table S2D): Gene-level, upper-quartile normalized counts converted to log2-transformed RPKM values. Column Annotations (Rows 4-81) described in Table S1. Row annotations:
  - id: Row identifier (gene symbol)
  - data type: defines if meta data or transcriptomic data
  - gene_id: Ensemble gene ID
  - geneSymbol: HUGO gene symbol
  - gene_type: Gene type (protein-coding, etc…)
  - length: Gene length
- Proteomic data (Table S3A): Two-component normalized Log2 transformed protein expression. The TMT ratios are normalized and filtered (see Methods "Two-component normalization of TMT ratio distributions" and "Dataset filtering "). Column headers:
  - Id: Same as accession_number
  - data type: defines if meta data or proteomics data
  - id.description: Same as entry name
  - geneSymbol: Gene symbol obtained for accession_number from UCSC Table browser
  - numColumnsProteinObserved: The number of TMT10-plex experiments the protein subgroup was observed in (1-25)
  - numSpectraProteinObserved: The total number of identified MS/MS spectra for the protein subgroup across all 25 TMT10-plex experiments. Each TMT MS/MS spectrum measures a peptide from 10 samples (9 individual samples, and the common reference sample (mix of 103 tumors and 100 normal adjacent tissue).
  - protein_mw: Calculated protein MW (Da) of the protein sequence belonging to accession number
  - percentCoverage: Percent of the protein sequence that is covered by peptides identified by MS/MS
  - numPepsUnique: Count of distinct peptide sequences detected for the protein subgroup.
  - scoreUnique: Sum of the Spectrum Mill identification scores of the distinct peptides for the protein subgroup
  - species: HOMO SAPIENS, smORFeomeHuman
  - orfCategory: dORF, IncRNA,uORF for smORFeome entries
  - accession_number: RefSeq protein ID, MiTranscriptome Gene ID
  - accession_numbers: All accession numbers for the protein subgroup that were parsimoniously grouped together on the basis of shared peptides for the protein
  - subgroupNum: Combined number proteinGroup.subgroup
  - entry_name: Protein description for accession_number
- Proteomic data (Table S3D): Unfiltered version of S3A