MITRE | SOLVING PROBLEMS FOR A SAFER WORLD®

Use Case Exercise 3:

Interpreting and explaining ‘omics results

The views, opinions, and/or findings contained in this report are those of The MITRE Corporation and should not be construed as an official Government position, policy, or decision, unless designated by other documentation. This technical data deliverable was developed using contract funds under Basic Contract No. W56KGU-18-D-0004.

Use Case Exercise 3: Interpreting and explaining 'omics results

The purpose of this exercise is to develop tools to make it easier for researchers to interpret results, understand their context, and generate new hypotheses. These tools should enable researchers to interact via natural language and possibly other low/no-code means.

General Task: You are given a results dataset and some documentation about the experiment's purpose and methods. Semi-automatically identify significant results in the dataset, provide relevant contextual information about these results, and generate explanations and hypotheses from these results using scientific knowledge from papers and databases.

Exercise Scenario: This exercise is based on an analysis from the paper: Pan-cancer analysis of post-translational modifications reveals shared patterns of protein regulation (Geffen et al., Cell). Researchers have a results dataset about the relative abundances of phosphorylated forms of proteins from two classes of tumor samples. Some of the phosphorylated proteins are significantly more abundant in one tumor class than the other. The researchers want to know more about these phosphorylated proteins, how these results relate to other work in the field, and how these phosphorylated proteins might be involved in the traits that differentiate the two tumor classes.

Dataset and Documentation to Use: The dataset is located in Geffen et al.'s Supplementary Table S3. In this file, the results dataset to use is in the sheet labeled 'Table 3G'. This Excel file also has a ReadMe sheet with descriptions of the variables in Table 3G.

The text of the paper provides additional information about this analysis. The paper's Results section briefly describes the goal of the analysis in Table S3G, saying "*HRD cancers rely on alternative repair pathways to mitigate double-strand break (DSB) damage.*[47] *To investigate the PTM-directed activities of repair proteins in HRD cancers, we performed differential expression analyses (across all feature types) between HRD and homologous recombination-proficient (HRP) tumors across DNA repair genes*". The paper's Introduction and Methods sections also provide useful information for interpretation of Table S3G.

For your understanding, here is some additional background and explanation that may help: Homologous recombination deficient (HRD) tumors are tumors that have lost the ability to repair double-stranded DNA breaks via homologous recombination (the exchange of genetic information between similar stretches of DNA sequence). As repairing mutations is important to persistence, these tumors are thought to rely on other DNA repair mechanisms to compensate. This study compared the activity of DNA repair mechanisms between HRD tumors and homologous recombination-proficient

(HRP) tumors, which have retained homologous recombination repair mechanisms. To investigate the activity of DNA repair mechanisms, the study measured the abundances of proteins known to be involved in DNA repair as well as the abundances of modified forms of these DNA repair proteins. Protein modifications include phosphorylation and acetylation, in which phosphate groups and acetyl groups are added at certain amino acid sites within proteins. These modifications frequently alter a protein's activity and can consequently increase or decrease the activity of molecular pathways and processes.

Table S3G has results about the relative abundances of DNA repair proteins and their phosphorylated and acetylated forms between HRD and HRP tumors. The column 'gene_name' contains the protein names, the column 'feature' indicates whether the protein is in phosphorylated or acetylated form, and the column 'variableSites' indicates the sites (the amino acid positions in the proteins) that are phosphorylated or acetylated. Alternatively, the column 'prot_residue' contains protein names and modified sites as combined values. A statistical analysis was run to identify the protein forms that were different in abundance between the two tumor classes. The column 'adj.P.Val' contains the statistical significance of differences in abundance after multiple test correction. The relative abundance between the tumor classes is under "logFC", with positive numbers indicating greater abundance in HRD tumors than in HRP tumors.

Generating Interpretations: You can use LLMs to generate interpretation text and can use some text from the paper or other sources as input. However, to compare interpretations generated by your tools with those in the paper, you will need to exclude text in the paper about interpretation of these results. The relevant interpretation portions of the paper text to exclude are 1) a small portion of the Results section, starting with the sentence "*These comparisons revealed significant differences in the phosphorylation of* …" through the remainder of that paragraph and 2) the entire Discussion section.

Interpretation and Explanation Questions to Address:

- First, semi-automatically identify the statistically significant results in the dataset. Which phosphorylated proteins are more abundant in the HRD tumor classes? Provide the protein names and amino acid sites that are phosphorylated, e.g., PARP1 at site S782.
- Which proteins are known to regulate phosphorylation of these proteins at these sites? E.g., PKA phosphorylates PARP1 at S782. Provide citations for this information.
- For the significant phosphorylated proteins, what is known about the effects of phosphorylation at these sites on the protein's activity? E.g., phosphorylation of

PARP1 at S782 has been shown to regulate the activity of PARP1. Provide citations.

- Is there prior evidence that these phosphorylated forms might be involved in the trait that differentiates the two tumor classes, homologous recombination deficiency? If so, how might these phosphorylated forms be involved? E.g., from the Geffen paper, "*phosphorylation of EXO1 S714 … has been proposed to attenuate EXO1 activity and hinder homologous recombination (HR) as a result.*" Provide citations.
- Are there drug candidates that target these significant proteins, and if so, what are they? What is known about their effectiveness in treating cancers?
- Are there results in this dataset that are particularly interesting scientifically, and if so, how? Provide citations for supporting information.

Interpretations Given in the Paper to Compare With: In Table S3G, there are more than 200 phosphorylated protein forms (i.e., protein forms with the "feature" column value "phosphoproteome") that have adjusted p-values less than 0.055 (the value that seems to be used as the cut-off for selecting ones to talk about in the paper). The paper text includes discussion about just a few of these, below; these could be used for comparison to your results:

- PARP1 S782 (FDR = 0.05); more abundant in HRD tumors; phosphorylation of this site is PKA mediated; phosphorylation at this site is known to regulate PARP1 activity; refs: Li et al., Gupte et al., Brunyanszki et al.
- PARP1 S179 (FDR = 0.05); more abundant in HRD tumors; phosphorylation of this site is ATR mediated; phosphorylation at this site is known to regulate PARP1 activity; refs: Li et al., Gupte et al., Brunyanszki et al.
- POLQ S1587 (FDR = 0.05); more abundant in HRD tumors; the paper does not indicate what phosphorylates this; it says "*the functional effects of S1587 phosphorylation … have not been well-studied*"; the paper also says POLQ promotes MMEJ (microhomology-mediated end-joining) by inhibiting RAD51-mediated HR, and its loss has been shown to elicit synthetic lethality in HRD tumors, including in cell lines resistant to PARP inhibition"; refs: Zatreanu et al., Ceccaldi et al.
- EXO1 S714 (FDR = 0.04); more abundant in HRD tumors; phosphorylation of this site is ATR mediated; " *increased phosphorylation of* <this site> ... *has been proposed to attenuate EXO1 activity and hinder homologous recombination (HR) as a result.*"; ref: Bolderson et al.
- For PARP and POLQ proteins, the paper notes that there are drugs that target them. The Discussion section says "… *the variation in response to therapies that target DNA repair genes (e.g., PARP and POLQ inhibitors, etc.)*".