**Analysis of Independent Differences (AID): a TPP Analysis**

**Instructions**

AID examines the differences between the fractions of non-denatured protein in order to predict the most likely shifted proteins from thermal proteome profiling experimentsThese instructions provide an overview about AID inputs and outputs. The R script is accessible at https://github.com/alex-bio/TPP and the Shiny App, which requires no prior knowledge of R or coding, is accessible at https://gygi.med.harvard.edu/software. For further details on how AID works, please see Panov, A. & Gygi, S.P. Analysis of Independent Differences (AID) detects complex thermal proteome profiles independent of shape and identifies candidate panobinostat targets (2019).

## Input

There are five inputs for AID:

- Control file
- Treatment file
- Peptide threshold
- P-value threshold
- Temperatures

**Control and treatment files**

The control and treatment files should contain protein abundance values (spectral counts, summed reporter ion intensities, etc.) from searched mass spectra. Generally, each file should have a column identifying proteins and multiple columns containing the protein abundance values across different temperatures. The control and treatment files must be .csv (Comma Separated Values). Excel spreadsheets can be converted in Excel by File/Save As/ and changing the file type to .csv. Within each file, there must be a column with the header “Protein.Id” and a column with the header “Number.of.peptides” (note capitalization). The “Protein.Id” column is how AID matches proteins between control and treatment conditions. It is recommended to use an identifier that can discriminate between protein isoforms. If a simple gene symbol is used (e.g. AMPK), multiple isoforms may be present. If a “Protein.Id” appears more than once within either control or treatment files (e.g. AMPK appears twice in the control file), those observations will not be analyzed because isoforms/proteins are indistinguishable. Moreover, any observation that contains an “NA” value in any column will not be analyzed. Reverse hits may be present in the uploaded control and treatment files, as long as they are indicated in the “Protein.Id” column by at least two number signs, “##”. Any proteins containing at least two number signs in their identifier in the “Protein.Id” column will be filtered out.

For the protein abundance values across different temperatures, the lowest temperature column must have the header “temp_01”. All ensuing columns should take the same format, “temp_02”, “temp_03”,…, for as many columns as are present in order of increasing temperature. Make sure that other columns that do not contain relevant protein abundance values do not have the phrase “temp” in the header/column name. This analysis will work with any number of temperature channels from both isobaric tag labelled or label-free data.

All other columns within the control and treatment files may remain and will have no bearing on the analysis (see note about columns with “temp” in the header immediately above). Other columns will return in the output; see below for details.

Two example files can be found at <XXX> to run in the Shiny App or use as a template. These files contain a subset of TPP data from an experiment using a TMT-10plex with DMSO as control and 50 µM staurosporine as treatment. The corresponding temperatures are the default temperatures in the Shiny App.

**Peptide threshold**
AID will only analyze proteins with a “Number.of.peptides” value greater than or equal to the user-defined peptide threshold value. Default is 3 peptides, minimum is 1 peptide.

**P-value threshold**
The differences between the treatment and control conditions are examined for each protein. An experiment using a TMT 10-plex, for example, yields 10 difference values for each protein, one for each temperature channel. Each temperature channel across all proteins forms a Normal distribution, such that individual *p*-values may be calculated. AID first ranks proteins most likely shifted by the log of the Multivariate Normal *p*-value; however, a second level of ranking is employed by examining the number of individual Normal *p*-values within a given protein that are below the defined *p*-value threshold. For example, in a TMT 10-plex, if a given protein has 9 *p*-values less than 0.05 and 1 greater than 0.05, then an output of AID, “pval_count”, shows a value of 9.

**Temperatures**
Input the actual temperature values used in the experiment. These values are only used for graphing purposes.

## Running the analysis

There are default values for peptide threshold, *p*-value threshold, and temperatures, which may be changed as desired. Once both control and treatment .csv files are uploaded, a button appears that says “Run AID”. Click to run the analysis. A progress bar should appear, and the analysis may take a few minutes. Once complete, a button to download the output spreadsheet and a graphing utility will appear.

## Output

### Spreadsheet
The spreadsheet, an Excel .xlsx file, will have two sheets: “AID_output” and “normalized_data”.

“AID_output” will have the following columns:

- “Protein.Id”
- “temp_01”, “temp_02”,…: difference values, or $y_{i,j}$s, for each protein *i*
- “log_Multiv_Norm_pval”: log of the Multivariate Normal *p*-value
- “pval_temp_01”, “pval_temp_02”,…: Individual Normal *p*-values for each difference value, or $y_{i,j}$s, for each temperature *j*
- “pval_count”: Number of individual Normal *p*-values per protein that pass the user-defined *p*-value threshold
- “var_warning”: Warning issued if variance of the differences per protein is greater than 2 standard deviations away from the mean variance; this is qualitative and intended to highlight curves that oscillate between extreme values
- “sum_of_signs”: Sum of the signs of the difference values, which indicates approximate directionality
- “prediction”: Predicted stabilization or destabilization based on “sum_of_signs”

“AID_output” is sorted first by ascending “log_Multiv_Norm_pval”, second by descending “pval_count”, and third by decreasing magnitude of “sum_of_signs”. If the log of the Multivariate Normal *p*-value is so small it is estimated as -∞, then a value of -1000 is displayed. If a few proteins have a log Multivariate Normal *p*-value equal to -1000, then those proteins are further ranked by the columns “pval_count” and “sum_of_signs”, respectively.

“normalized_data” will have the following columns:

- “Protein.Id”
- “Number.of.peptides_control”

- “Number.of.peptides_treatment”
- “temp_01_control”, “temp_02_control”,…: Normalized values; normalized to the max channel
- “temp_01_treatment”, “temp_02_treatment”,…: Normalized values; normalized to the max channel
- All other columns included in the input control and treatment .csv files

**Graphs**

The “Protein.Id” column is listed in alphabetical order in the drop-down list. The defined temperature values are used to graph each protein melting curve.