**Description:**

Get a list of mutually exclusive binary features (e.g., mutation of gene loci) that are strongly correlated with a continuous target (e.g., geneset enrichment score, gene expression).

**Summary:**

REVEALER2.0 is a powerful tool to investigate mutually exclusive binary features that are correlated with a continuous target. It loops through the provided features and for each iteration, choose a feature that has highest conditional information coefficient value based on provided target and seed in each iteration.

**Introduction:**

REVEALER2 (repeated evaluation of variables conditional entropy and redundancy 2) is a method for identifying groups of genomic alterations that together associate with a functional activation, gene dependency or drug response profile. The combination of these alterations explains a larger fraction of samples displaying functional target activation or sensitivity than any individual alteration considered in isolation. REVEALER2 can be applied to a wide variety of problems and allows prior relevant background knowledge to be incorporated into the model. Compared to original REVEALER, REVEALER2.0 can work on much larger sample size with much higher speed.

**Algorithm:**

PUT PSEUDOCODE HERE

**Parameters:**

|  |  |
| --- | --- |
| **Name** | **Description** |
| target\_file | This is a file in gct format that contains continuous target data. |
| target\_name | Name of target in target file. Take first row in default. |
| feature\_file | This is a file in gct format that contain all the binary features. Should be feature in row and sample name in column. |
| prefix | Prefix of result files. |
| seed\_file | This is a file in gct format that contains seed to start with. |
| seed\_name | Name of seeds to be used. Multiple seed names can be provided and will be collapsed in program. |
| gene\_locus | Text file that contains the information of locus of each gene. This file can be downloaded from github page. |
| tissue\_file | This is a file in gct format that contains tissue type information of samples. |
| gmt\_file | This is a file in gmt format that contains allele information of each feature. This file can be generated in REVEALER\_preprocess |
| minimum\_iteration | This indicates how many iterations REVEALER should run. If set to -1, iteration will stop automatically when total IC start decreasing. |
| k | This indicates number of neighborhood for kernel. Higher number will lead to higher accuracy but lower speed. |
| bandwidth\_multiplication | This indicates multiplication value in bandwidth calculation. |
| bandwidth\_adjustment | This indicates adjustment value in bandwidth calculation. |
| direction | This indicates correlation direction between features and target expected to be observed. |
| num\_top | This indicates number of top scored features to be plotted in intermediate figures. |
| low\_threshold | This indicates lower threshold of feature filtering. If integer is provided, then features that have less than that number of sample mutated are removed. If float if provided, then features that have less than that ratio of sample mutation among all samples are removed. |
| high\_threshold | This indicates higher threshold of feature filtering. If integer is provided, then features that have more than that number of sample mutated are removed. If float if provided, then features that have more than that ratio of sample mutation among all samples are removed. |
| if\_collapse | This is a boolean option that indicates if exactly same features are collapsed together in intermediate figure. |
| subset | This is a text file indicating subset of samples to be utilized. This can be used if only one type of cancer is investigated. |
| figure\_format | Format of generated figure. |
| if\_pvalue | This is a boolean option that indicates if p-values are calculated. |
| if\_bootstrap | This is a boolean option that indicates if variants are calculated. |
| if\_cluster | This is a boolean option that indicates if features are clustered with NMF for intermediate figures. |
| if\_intermediate | This is a boolean option that indicates if intermediate figures are generated. |
| separater | This indicates character that is used to separate gene name and later part in input files |
| gene\_set | This is a file that include genes that are used. This is used when only subset of genes are used for analysis. |
| alpha | This is the value that is used raise target values by specific time. |
| gzip | This is a boolean option that indicates if result files are gzipped. |
| output\_folder | This indicates name of output folder. |

**Input Files:**

1. Target dataset: GCT file

This file contains continuous target phenotype to correlate with features.

1. Feature dataset: GCT file

This file contains all binary features that is used to compare.

1. Optional seed database: GCT file

This file contains seed that is used for the first iteration. If both seed file and seed name is not provided, then it will start without seed. If seed name is provided but no seed file is provided, then it will find seeds from feature file.

1. Optional gene locus: txt file

This file contains locus information of gene. This can be downloaded from github.

1. Optional allele information for each feature: gmt file

This file contains what mutations are included in each feature. This file can be generated by REVEALER\_preprocess.

1. Optional sample list: txt file

This file contains list of sample to use for analysis. If user want to run REVEALER on only subset of samples(for example, sample in only specific cancer type), this can be used.

1. Optional gene list: txt file

This file contains list of genes to use for analysis. If user want to run REVEALER on only subset of genes(for example, gene with known function), this can be used.

**Output Files:**

1. Main Result Heatmap: PDF or PNG image

The main result heatmap. The first row is target distribution and following rows are top feature picked in each iteration. Feature name is labeled on left side and optional variations and p-values will be labeled on right side.

1. Optional intermediate heatmap: PDF or PNG image

The intermediate heatmap that is generated one for each iteration. First row is target distribution and following rows are top n features that is picked in this iteration. Feature name is labeled on left side and optional variations and p-values will be labeled on right side.

1. Optional allele detail for each final feature: gmt file

Allele information is provided for each feature that is selected for the result. This file is generated only when user provide gmt file of allele information of all features as input file.

1. Optional tissue distribution image: PDF or PNG

If tissue distribution of samples is provided, then tissue distribution result compared to target is plotted. First row is target distribution and following rows are tissue type of each sample.