**Description:**

Preprocessing step of REVEALER2. Mutation information of samples can be combined to single matrix of feature by sample based on user’s requirement. Also, matrix can be passed in and various filter can be applied.

**Summary:**

REVEALER2 is a powerful tool to investigate mutually exclusive binary features that are correlated with a continuous target. This preprocessing step can take raw mutation file and prepare ready-to-use input file for REVEALER2. The features can be customly generated and filtered based on its weight, frequency, or classification.

**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. REVEALER2 take matrix of binary feature by samples as input. This tool can prepare input matrix for REVEALER2 from raw MAF format mutation file. User can provide their own filtering method to generate different types of input matrix. The input matrix can be generated based on mutation classification, weight compared to phenotype, or frequency of mutation.

**Algorithm:**

**Allele Mode**: In this mode, each feature is [Gene Name]\_[Allele Name]. No combination is performed on mutations. Filtering will be applied based on gene list provided by user or frequency threshold provided by user or in default. If user is interested not only in mutually exclusive genes, but also in allele level, this mode is recommended. The output will be larger and more sparce compared to other mode, so specifying gene would be recommended.

**Mutall Mode**: In this mode, each feature is [Gene Name]\_Mut\_All. Basically, all gene-allele pair of each gene is combined into single feature. Filtering can be applied based on gene list provided by user or frequency threshold provided by user or in default.

**Class Mode**: In this mode, each feature is [Gene Name]\_[Classification Name]. For each gene-class combination feature, samples with that mutation will be set as 1. [Gene Name]\_Mut\_All feature in above mode is also generated in this mode. Filtering will be applied based on gene list provided by user or frequency threshold provided by user or in default. This is the default mode and recommended for general investigation.

**Weight Mode**: In this mode, user can provide target, which will be same with target in later REVEALER2 main part, to filter mutations. Each feature will be named as [Gene Name]\_weight\_[Threshold Provided]. First, target is normalized and raised based on user provided value. Then, IC between target and each gene-allele combination is calculated (detailed calculation method can be checked in REVEALER documentation). For each gene, all gene-allele combination that pass weight threshold will be combined to single feature. Filtering can be further applied based on gene list provided by user or frequency threshold provided by user or in default.

**Weight With Filter Mode**: This mode is mostly same with mode above with one extra filtering step. After combining gene-allele combination after filtering with weight threshold, number of positive sample is counted and if it is less than number of positive gene for this gene times given ratio, then this gene is filtered out. This filter is added to filter out genes with most allele meaningless in this situation.

**Frequency Mode**: In this mode, feature is named as [Gene Name]\_frequency\_[Top N num provided by user]. User need to provide target, which is used to break the tie. First, target is normalized and raised based on user provided value. Then, IC between target and each gene-allele combination is calculated (detailed calculation method can be checked in REVEALER documentation). For each gene, all gene-allele combination is sorted based on its frequency and IC, with frequency prioritized, and then top n gene-allele combination is combined as single feature. Filtering can be further applied based on gene list provided by user or frequency threshold provided by user or in default. This mode is not recommended in general. Mainly because with test using real data, we realized that this mode cannot reflect biological insights into the data, which lead to less meaningful result in later step.

For all mode, gmt files can be generated indication detailed allele in each feature, which can be used as input for later REVEALER.

**Parameters:**

|  |  |
| --- | --- |
| **Name** | **Description** |
| input\_file | This is a file in maf or gct format that contains mutation information. |
| protein\_change\_identifier | Name of Protein Change identifier. This is the most detailed classification for mutations. |
| mode | This is the mode preprocessing should run with. Available options are: class, freq, weight, weight\_filter, mutall, or comb. |
| prefix | Prefix of result files. |
| phenotype\_file | This is a file in gct format that contains target to run weight, weight\_filter, or freq mode. |
| Phenotype\_name | Name of target feature in phenotype file. |
| weight\_threshold | Weight threshold to filter by IC. |
| class\_file | Class file that includes list of class. This is required if gct file is provided as input. |
| direction | This indicates correlation direction between features and target expected to be observed. |
| frequency\_threshold | This indicates top frequency threshold number of gene-allele combination should be picked for each feature. |
| name\_match | This is a boolean parameter that indicate if sample of target and name in maf file is matching. For TCGA example, because maf file has longer annotation for samples, then this parameter should be set to False. |
| gene\_list | This indicates multiplication value in bandwidth calculation. |
| sample\_list | This is a text file indicating subset of samples to be utilized. This can be used if only one type of cancer is investigated. |
| ratio | Ratio of number of samples in selected gene-allele combined features over number of samples in that gene that is acceptable. Used in weight\_filter mode. |
| make\_figure | This is a boolean parameter that indicates if heatmap is need to be generated for each gene indicating distribution of each allele in that gene. Recommended only when running on few genes, otherwise too many figures are generated. |
| total\_ratio | This ratio is used to filter out too frequent genes that number of positive samples compared to total number of samples is more than this value. |
| class\_separater | Separator between gene name and later part if gct file is provided. |
| if\_gmt | This is a boolean option that indicates if gmt file is generated as output. |
| 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. |
| gzip | This is a boolean option that indicates if result files are gzipped. |
| output\_folder | This indicates name of output folder. |

**Input Files:**

1. Annotated mutation: MAF file

This file contains mutation information. Can be generated from VCF files.

1. Optional target phenotype: GCT file

This file contains continuous target to calculate IC in weight, weight\_filter, and freq mode.

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. Feature matrix: GCT file

The main result that is a matrix in gct format, with row as feature names and column as sample names. This can be directly used as REVEALER input.

1. Optional gene level heatmap: PDF or PNG image

Heatmap showing allele distribution compared to target for each gene.

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

Allele information is provided for each feature in result gct file matrix.