

MakeSubsetGctAndCls Documentation

Description: Make .cls and .gct file subset for a particular class comparison.

Author: Barbara Weir

Date: 8/2/2010

Summary

This module creates a .cls file and a new .gct file to match from an input .gct data file and a sample annotation file. Removes samples with NA and those samples in the annotation file not found in the input .gct file.

References

Cheung et al.

Input Parameters

Name	Description
annotationfile	Sample annotation file – simple format of cell lines in rows and alterations in columns, two different formats are accepted- 'external' or 'sampleinfo'.
alteration	Alteration name must match a column name in the annotation file used.
type	Type of annotation file. Allowed values= {external, sampleinfo}. (Default=external)
gctfile	Input .gct file after PMAD (or ZMAD) normalization
prefix	Output filename prefix for .cls and .gct files to be created.

Input files

1. Annotation file **REQUIRED**

Simple format of cell lines in rows and alterations in columns, two different format options are accepted. One is equivalent to the sample information file format and can be used for mutation class comparisons. The second is more general and can be used for any annotation type.

- Option #1, Sample information file:

Sample names are listed in the first column, annotations in the first row. Data for mutation columns should be annotated as: WT, gene name (ex. KRAS) or NA (for unknown). Mutation

annotations in the first row (header) must start with the gene name (ex. KRAS_status). 'NA' samples (and any other samples not found in the annotation file) will be ignored.

- Option #2, External annotation file:

Categories for classification should be given as column headers. Data should be given as '0' for target class, '1' for non-target class and if desired '2' for a third, ignored class. Samples must not be left blank, use 'NA' instead. 'NA' samples (and any other samples not found in the annotation file) will be ignored.

2. Input data file **REQUIRED**

Input data file of cell lines (columns) and shRNAs (rows) in .gct file format (after PMAD/ZMAD normalization).

Output Files

1. Class file (<prefix>.cls)

.cls file for the comparison (mutant or target class is always class 0)

2. Output data file (<prefix>.gct)

.gct file for the comparison (samples listed as NA or not found in the sample information or external annotation file are excluded, so the .gct file may be smaller than the original)

3. STDOUT file (stdout.txt)

Standard output will give some information about cell lines in the input .gct file that matched or didn't match the input annotation file. Check spelling of names in this file if you are missing samples that should be there and update the annotation file accordingly.

Platform Dependencies

Module Type: RNAi

CPU type: any

OS: any

Language: Perl