GenePattern

PreprocessDataset Documentation

Module name: PreprocessDataset
Description: Perform several preprocessing options on a RES, GCT, or ODF input file
Author: Joshua Gould, Pablo Tamayo (Broad Institute),
Contact: gp-help@broadinstitute.org
Release: 4.0

Summary: This module performs a variety of pre-processing operations including thresholding/ceiling, variation filter and normalization:

Thresholding:
Value = threshold if Value < threshold
Value = ceiling if Value > ceiling

Variation filter (row by row exclude genes for which):
max / min < minchange
max – min < mindelta
here the max and min are computed over a row excluding the top (and bottom) "num.excl" experiments. This is to prevent one or more "spikes" to make the gene pass the filter.

The filter flag controls the application of both thresholding and the variation filter.

Independently of the application of thresholding and the variation filter the module also has a flag (preprocessing flag) to turn on row normalization of the dataset (after thres. and filtering).

Prob. Threshold allows sampling of the rows without replacement to obtain that fraction of the total number of rows.

The order of the steps in the module is as follows:
1. Thresholding
2. Remove row if n columns not >= than given threshold
3. Variation filter
4. Log Base 2

References: none

Parameters

Name	**Description**
input.filename (required)	input filename - .res, .gct, .odf
threshold.and.filter	turn on thresholding and filtering • no(0) • yes(1) (default)

floor	Value for floor threshold (default = 20)
ceiling	Value for ceiling (default = 20000)
min.fold.change	Minimum fold change for filter (default = 3)
min.delta	Minimum delta for filter (default = 100)
num.outliers.to.exclude	Number of experiments to exclude (max & min) before applying variation filter (default = 0)
row.normalization	turn on row normalization • no (0) (default) • yes (1)
row.sampling.rate	Value for uniform probability filter (default = 1) Note: this parameter is only applied if threshold.and.filter is set to no.
threshold.for.removing.rows	Threshold value for removing rows
num.columns.above.threshold	Remove row if n columns not >= given threshold
log2.transform	Apply log2 transform after all other preprocessing steps • no (default) • yes
output.file.format	Output file format • gct • res • same as input (default)
output.file (required)	Output file base name

Return Value:

The filtered, preprocessed output file

Platform dependencies:

Task type:	Preprocess & Utilities
CPU type:	any
OS:	any
Java JVM level:	1.6
Language:	Java

PreprocessDataset Version Notes

Version	**Description**
4.0	The log2.transform.flag was added in version 4.0. This parameter allows data to be log-base-2 transformed after all other preprocessing steps have been performed. In addition, the sigma parameter was deprecated in version 4.0. Change parameters names for clarity.