SelectFileMatrix Documentation

Description: Select a matrix from a delimited file.

Author: David Eby (Broad Institute), gp-help@broadinstitute.org

Summary

This module takes a delimited text file as input and creates a subset matrix file based on user-specified parameters. It can be used to extract data from a tab-, space-, or comma-separated input file in multiple ways, from a single cell or an individual row/column to a rectangular matrix or a discontinuous selection of partial rows and columns.

There are many ways to specify your matrix using boundary indices (start and end), individual index selections, selection ranges, or any of these in combination. Use *start.row* and *end.row* to specify the bounds of a row selection and *start.column* and *end.column* to do the same for columns. The *row.select* and *column.select* parameters can be used to specify rows and columns individually and in ranges. If the *output.rejects* parameter is set to **true**, the module will also produce a file containing any items not selected.

The ordering of the data in the output file will match its relative ordering in the input file; the selection process will not reorder the data. Likewise, the ordering of unselected data in the *rejects* file will also match its relative ordering in the input file.

For example, take the following GCT file containing five samples and assume that you want to extract only the expression values for samples 1,2,3 and 5. You can specify this by using *start.row=4* and *column.select=3-5,7*. In this case it is not necessary to specify *end.row* as it will default to the last row in the file.

Input GCT file:

#1.2						
6	5					
Name	Description	Sample1	Sample2	Sample3	Sample4	Sample5
AFFX-BioB-5_at	AFFX-BioB-5_at	-243.0	-130.0	-256.0	-62.0	86
AFFX-BioB-M_at	AFFX-BioB-M_at	-218.0	-177.0	-249.0	-23.0	-36.0
AFFX-BioB-3_at	AFFX-BioB-3_at	-163.0	-28.0	-410.0	-7.0	-141.0
AFFX-BioC-5_at	AFFX-BioC-5_at	182.0	266.0	24.0	142.0	252.0
AFFX-BioC-3_at	AFFX-BioC-3_at	-289.0	-170.0	-535.0	-233.0	-201.0

Output file for *start.row=4* and *column.select=3-5,7*:

-243.0	-130.0	-256.0	86
-218.0	-177.0	-249.0	-36.0
-163.0	-28.0	-410.0	-141.0
182.0	266.0	24.0	252.0
-289.0	-170.0	-535.0	-201.0

If the boundary selectors are specified, they will be in effect even if mixed with individual or range selections. This can be used to further restrict a matrix without changing the existing select specifier. With the input file above, for example, adding *start.column=3* and *end.column=6* will select the values or sample 2 and 3 only.

Output file for *start.row=4*, *column.select=3-5,7, start.column=3,* and *end.column=6*:

-130.0	-256.0
-177.0	-249.0
-28.0	-410.0
266.0	24.0
-170.0	-535.0

Note that for certain input file formats the output file produced may not be valid in that format. With a RES file, for example, if the header lines are not included in the selection then the output file will not be a RES file. Even if they are included, the embedded row count will be incorrect if other rows are removed. This module makes no attempt to adjust output to match the input format; its functionality is strictly limited to matrix selection. For such needs, use *SelectFeaturesRows* or *SelectFeaturesColumns* instead.

Parameters

Name	Description
input file	Input file from which to pull the matrix. Must be a delimited text file.
output file base name	Base name for the output file. The module will add an appropriate extension, either .csv for a comma-delimited file or .txt for a space- or tab-delimited file.
start row	Starting row of the matrix (line number within the file). Rows are indexed beginning with 1. If no value is provided, the matrix will start at the beginning of the file.

end row	Ending row of the matrix (line number within the file). Rows are indexed beginning with 1. If no value is provided, the matrix will extend to the end of the file.
start column	Starting column of the matrix. Columns are indexed beginning with 1. If no value is provided, the matrix will start with the first column in the file.
end column	Ending column of the matrix. Columns are indexed beginning with 1. If no value is provided, the matrix will extend to the last column in the file.
delimiter	The character for the column delimiter. Choices are *tab*, *comma*, and *space*. The same delimiter will be used for both the input file and the output file. Default: *tab* Each occurrence of the delimiter character indicates a column separator and so multiple consecutive characters will be treated as a series of empty columns. Two values separated by five spaces, for example, will be identified as six columns (four empty) rather than two.
column select	Select specific columns to include in the matrix (e.g., "1,2,4-6"). Note that start/end column parameters may limit this selection.
row select	Select specific rows to include in the matrix (e.g., "3,5,7-12"). Note that start/end row parameters may limit this selection.
output rejects	Produce an output file containing the items not meeting the selection criteria. The file will be named <output.file.base.name>.REJECT.<ext> with the same extension as for the output file.

Output Files

1. <output.file.base.name>.<ext>

   The output file containing only the specified matrix (rows and columns).

2. <output.file.base.name>.REJECT.<ext>

   Optionally, an output file containing any items not selected for inclusion in the output file.

Platform Dependencies

Module type: Preprocess & Utilities

CPU type: any

OS: any

Language: Java

GenePattern Module Version Notes

Version	Description
1	Initial release