MapChipFeaturesGeneral Documentation

Description: Changes (maps) the features (gene names) of a data set.

Author: GenePattern Team, gp-help@broadinstitute.org

Summary

This module takes a RES file and produces an output RES file with the same numeric data but with the features mapped to a different probe, feature, or gene set.

The module replaces the set of features, gene names, or accessions in the file with a different set. The alternative set is listed in a "lookup table," which is a tab-delimited text file provided by the user. The module can be used, for example, to convert the probe names in a RES file that was originally generated from one chip to the equivalent probe names from another chip.

Lookup table

The lookup table can have two, three, or four tab-delimited columns, as specified in the *table format* parameter.

- Two columns: one column contains the feature names/gene accessions/gene names that are in the file to be converted, and the other column contains the names/accessions to be substituted for them; each row is a one-to-one name correspondence (specify *2 column* in the *table format* parameter)
- Three columns: the first column contains the probe set identifiers, the second column contains the names/accessions to be substituted for them, and the third column has the gene title of the probe set; this format is similar to the CHIP file format, and only the first two columns are used by the module (specify *2 column* in the *table format* parameter)
- Four columns: the first and third columns are accession-to-accession columns (as in the two column format), and the other two columns contain annotations; only the first and third columns are used by the module (specify *4-column* in the *table format* parameter)

The first two rows of the file are documentation headers that are ignored by the module, and the content of the table should start on the third row.

The direction of the lookup (*direct* or *reverse*) can be specified with the *direction* parameter. For example, assume you have a RES file containing ChipA names and a lookup table that contains ChipA information in column 1 and ChipB information in column 2. If you do a "direct" lookup, the module replaces ChipA features with ChipB features. If, instead, the lookup file has ChipB in column 1 and ChipA in column 2 and you wanted ChipA features replaced with ChipB, you would use the "reverse" lookup.

GenePattern

Parameters

Name	Description
input filename	The name of the input gene expression file (RES file).
lookup table filename	Name of the lookup table file with the feature mapping (tab-delimited TXT); if a four-column table is provided, only columns 1 and 3 are used.
table format	The table format (*2 column* or *4-column*). If you have a 3-column file, select *2 column*; only columns 1 and 2 will therefore be used by the module.
direction	Direction of lookup (*direct* or *reverse*).
output prefix	The prefix to the name of the output file; default uses the input file name as the base, and appends "fea" before the file extension (RES file).

Input Files

1. RES file
2. The lookup table, which contains two header lines (these are ignored by the module, but are used for descriptive purposes) and is in two-, three-, or four-column format. The structure of each is shown below.

   Two Column:
   Header 1
   Header 2
   Feature1 Feature1_Mapping
   Feature2 Feature2_Mapping
   Feature3 Feature3_Mapping
   ...

   Three Column (module ignores third column):
   Header 1
   Header 2
   Feature1 Feature1_Mapping Feature1_Description
   Feature2 Feature2_Mapping Feature2_Description
   Feature3 Feature3_Mapping Feature3_Description
   ...

Four Column (module ignores second and fourth columns):

Header 1
Header 2

Feature1	Feature1_Description	Feature1_Mapping	Feature1_Mapping_Description
Feature2	Feature2_Description	Feature2_Mapping	Feature2_Mapping_Description
Feature3	Feature3_Description	Feature3_Mapping	Feature3_Mapping_Description

...

Note that in the case of a three-column table, only columns 1 and 2 are used by the module, and in the case of a four-column table, only columns 1 and 3 are used.

Here is an example of a lookup table in two-column format:

Affy 6800 vs U95 lookup table

6800 Probe Set	U95 Probe Set
A28102_at	31726_at
AB000114_at	41031_at
AB000115_at	36927_at
AB000220_at	376_at
AB000220_at	377_g_at

…

This example file would produce an output file with the same numeric data but with the features mapped into the 6800 chip set instead of the original U95 set.

A good source of three-column CHIP files is the GSEA FTP site:
ftp://ftp.broadinstitute.org/pub/gsea/annotations

Output Files

1. RES file

   The module returns an output RES file with the same name as the input data file, but with the suffix “fea” added before the file extension name. Thus, if the input file name is "foo.res," the output file will be called "foo.fea.res."

Platform Dependencies

Module type:	Preprocess & Utility
CPU type:	any
OS:	any
Language:	Perl

GenePattern

GenePattern Module Version Notes

Version	Release Date	Description
2	7 June 2013	Fixed a broken script reference.
1	13 May 2004	Initial version.