ExpressionFileCreator Documentation

Module name: ExpressionFileCreator
Description: Creates a res or gct file from a set of Affymetrix CEL files
Author: Joshua Gould, gp-help@broad.mit.edu

Summary

The ExpressionFileCreator module creates an expression dataset from a set of individual Affymetrix CEL files. The conversion is done using either the standard Affymetrix probe modeling algorithm MAS5 or the RMA algorithm. The result is a matrix containing one intensity value per probe set, in the GCT file format described at http://www.broad.mit.edu/genepattern/tutorial/index?gp_tutorial_fileformats.html.

Samples can be annotated using the clm file format. A clm file allows you to change the name of the samples in the expression matrix, reorder the sample names, and create a class label file in the cls format, also described on the Web page above.

By default, sample names are taken from the CEL file names contained in the zip file. A clm file allows you to specify the sample names explicitly. Additionally, the columns in the expression matrix are reordered so that they are in the same order as the sample names appear in the clm file. For example, the zip file my.cel.files.zip contains the files scan1.cel, scan2.cel, and scan3.cel. The clm file could contain the following text:

scan3.cel	sample3	tumor
scan1.cel	sample1	tumor
scan2.cel	sample2	normal

The column names in the expression matrix would be: sample3, sample1, sample2.

Notes:

- The results obtained from running MAS5 differ slightly from the official implementation of this algorithm.
- The input file can be a zip of CEL files or a zip of zipped CEL files.

References:

1. Affymetrix. Affymetrix Microarray Suite User Guide. Affymetrix, Santa Clara, CA, version 5 edition, 2001.
2. Rafael A. Irizarry, Bridget Hobbs, Francois Collin, Yasmin D. Beazer-Barclay, Kristen J. Antonellis, Uwe Scherf, and Terence P. Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics, 2003b. To appear.

Parameters:

Name	Description	Options
input filename	A zip file containing CEL files or zipped CEL files	
method	The method to use (MAS5 or RMA)	RMA;MAS5
quantile normalization	(RMA only) Whether to normalize data	yes;no

	using quantile normalization	
background correct	(RMA only) Whether to background correct using RMA background correction	yes;no
compute present absent calls	(MAS5 only) Whether to compute Present/Absent calls	yes=yes (create res file);no=no (create gct file)
normalization	The normalization method to apply after expression values are computed	none;mean scaling;median scaling
reference.scan.name	(mean or median scaling only) The scan name to be used as a reference for normalization. Leave blank to use median scan as reference	
scale value	(MAS5 only) Value to which all arrays will be scaled.	
clm.filename	tab-delimited text file containing one scan name, sample name, and class name per line	
output file	The base name of the output file	

Return Value:

1. gct or res file
2. cls file if clm.filename is supplied

Platform dependencies:

Task type:	Preprocess & Utility
CPU type:	any
OS:	any
Language:	R