



ExpressionFileCreator Documentation

Description: Creates a RES or GCT file from a set of Affymetrix CEL files
Author: Joshua Gould, gp-help@broadinstitute.org
Last Updated: April 26, 2012

Summary

The ExpressionFileCreator module creates an expression dataset from a set of individual Affymetrix CEL files. The conversion is done using the standard Affymetrix probe modeling algorithm MAS5, the RMA algorithm, the GCRMA algorithm, or the dChip algorithm. The result is a matrix containing one intensity value per probe set, in the GCT or RES file format described at http://genepattern.org/tutorial/gp_fileformats.html.

Samples can be annotated by specifying a clm file. A clm file allows you to change the name of the samples in the expression matrix, reorder the columns, select a subset of the scans in the input zip file, 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 scan names appear in the clm file. For example, the input zip file contains the files scan1.cel, scan2.cel, and scan3.cel. The clm file could contain the following text:

| | | |
|-------|---------|--------|
| scan3 | sample3 | tumor |
| scan1 | sample1 | tumor |
| scan2 | sample2 | normal |

The column names in the expression matrix would be: sample3, sample1, sample2. Additionally, only scan names in the clm file will be used to construct the GCT or RES file; scans not present in the clm file will be ignored.

Notes

- The MAS5 and dChip algorithms are based on their Bioconductor implementations. Therefore the results obtained from these algorithms will differ slightly from their official implementations.
- The GCRMA and RMA algorithms produce values that are in log2 but ExpressionFileCreator removes the log2 transformation before generating the result file.
- ST 1.1 and ST exon arrays are not currently supported.

Arrays supported

For a list of arrays supported by R2.8 please see

<http://bioconductor.org/packages/2.3/data/annotation/>

Common Errors:

Please visit the GenePattern FAQ for any errors you may encounter:

<http://www.broadinstitute.org/cancer/software/genepattern/doc/faq>

GenePattern

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.
3. Li, C. and Wong, W.H. (2001) Genome Biology 2, 1–11.
4. Li, C. and Wong, W.H. (2001) Proc. Natl. Acad. Sci USA 98, 31–36.

Parameters

| Name | Description |
|------------------------------|--|
| input file | A zip file of CEL files |
| method | The method to use Note: Because MAS5 requires MM probes it cannot be used with ST arrays. dchip is also an invalid method for ST arrays. |
| quantile normalization | (GCRMA and RMA only) Whether to normalize data using quantile normalization |
| background correct | (RMA only) Whether to background correct using RMA background correction |
| compute present absent calls | Whether to compute Present/Absent calls |
| normalization method | (MAS5 only) The normalization method to apply after expression values are computed. The column having the median of the means is used as the reference unless the parameter value to scale to is given. |
| value to scale to | (median/mean scaling only) The value to scale to. |
| clm file | A tab-delimited text file containing one scan, sample, and class per line |
| annotate probes | Whether to annotate probes with the gene symbol and description. NOTE: It is possible that no annotations may be found. |
| cdf file | Custom CDF file. Leave blank to use default internally provided CDF file. Note: If you specify a custom CDF file, the .res output file may be sorted differently depending on whether it is generated on Windows versus Mac or Linux. This is due to the C file parsing library used by this module. |
| output file | The base name of the output file |

Output Files:

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

GenePattern

Platform Dependencies

Module type: Preprocess & Utilities
CPU type: any
OS: any
Language: R

GenePattern Module Version Notes

| Version | Description |
|---------|---|
| v.10 | ExpressionFileCreator module v.10 contains a fix for a bug with annotating probes when some annotations are missing |
| v.11 | ExpressionFileCreator module v.11 now links to the latest (as of Feb 2012) Affymetrix CSV annotations. |