GenePattern

ExpressionFileCreator Documentation

Description: Creates a RES or GCT file from a ZIP archive of Affymetrix CEL files.

Author: Joshua Gould, gp-help@broadinstitute.org

Summary

The ExpressionFileCreator module creates a gene expression dataset from a ZIP archive containing individual Affymetrix CEL files. The conversion is done using one of the following algorithms:

- the standard Affymetrix probe modeling algorithm MAS5
- the RMA (Robust Multi-array Average) algorithm
- the GCRMA algorithm (link to PDF)
- the dChip algorithm

The result is a matrix containing one intensity value per probe set, in the GCT or RES file format.

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.

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.

Requirements

ExpressionFileCreator requires R 2.15.2 with the following packages:

- boot_1.3-7
- class_7.3-5
- cluster_1.14.3
- foreign_0.8-51
- KernSmooth_2.23-8
- lattice_0.20-10
- MASS_7.3-22
- Matrix_1.0-9
- mgcv_1.7-21
- nlme_3.1-105
- nnet_7.3-5
- rpart_3.1-55
- spatial_7.3-5
- BiocGenerics_0.4.0

- DBI_0.2-5
- RSQLite_0.11.2
- IRanges_1.16.2
- Biobase_2.18.0
- AnnotationDbi_1.20.1
- zlibbioc_1.4.0
- affyio_1.26.0
- preprocessCore_1.20.0
- affy_1.36.0
- Biostrings_2.26.2
- gcrma_2.30.0
- makecdfenv_1.36.0

Each of these R packages has been bundled into a GenePattern patch and will be automatically downloaded and installed when the module is installed. This process will take some time due to the size and number of these packages, so be patient during installation. R2.15.2 must be installed and configured independently.

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.
- The underlying Affymetrix R package used by ExpressionFileCreator v12 fixes a bug in the dChip algorithm implementation. Unfortunately, this means that dChip expression files created with previous versions are not directly comparable with newly created dChip files. **It is our strong recommendation that you discard older dChip results and re-create the expression files with the new version.**

Arrays supported

For a list of arrays supported by R2.15 please see http://bioconductor.org/packages/2.10/data/annotation/

Alternatively, you can provide a CDF with your job to process other array types.

Common Errors:

Check the GenePattern FAQ regarding errors you may encounter: http://www.broadinstitute.org/cancer/software/genepattern/doc/faq

References

Affymetrix. *Affymetrix Microarray Suite User Guide*, version 5. Santa Clara, CA:Affymetrix, 2001.

Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*. 2003;4:249-264.

Li C, Wong WH. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc Natl Acad Sci USA*. 2001;98:31-36.

Li C, Wong WH. Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. *Genome Biology*. 2011;2:research0032-research0032.11.

Parameters

Name	Description
input file (required)	A ZIP file of CEL files.
method (required)	The method to use to generate the GCT expression file. (default: RMA) 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 (link to PDF). (default: yes)
background correct	(RMA only) Whether to background correct using RMA background correction. (default: yes)
compute present absent calls	Whether to compute Present/Absent calls. If you do compute them, you will generate a RES file. If you do not compute them, you will generate a GCT file. (default: no)
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. Options include: • linear fit • mean scaling • median scaling (default) • none • quantile normalization
value to scale to	(median/mean scaling only) The value to which the median/mean scaling normalization should scale.

clm file	A tab-delimited text file containing one scan, sample, and class per line.
annotate probes (required)	Whether to annotate probes with the gene symbol and description. NOTE: It is possible that no annotations may be found. (default: yes)
cdf file	Custom chip definition file (CDF). 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 (required)	The base name of the output file.

Output Files

1. GCT file (if present/absent calls are NOT computed) or RES file (if present/absent calls ARE computed)
2. CLS file (if a CLM file is supplied)

Platform Dependencies

Module type:	Preprocess & Utilities
CPU type:	Any
OS:	Any
Language:	R 2.15

GenePattern Module Version Notes

Version	**Release Date**	**Description**
10		Fixed a bug with annotating probes when some annotations are missing.
11	4/30/2012	Linked to the latest (as of Feb 2012) Affymetrix CSV annotations.
12	3/15/2013	Updated to use R 2.15.2 and recent versions of all packages. Fixes a bug in handling CLM files with many CEL files. Fixes a bug in the ‘dChip’ implementation.