

IlluminaExpressionFileCreator Documentation

Module Name: IlluminaExpressionFileCreator

Description: Creates GCT files from a set of Illumina expression IDAT files.

Author: David Eby (contacts: Chris Williams, Ted Liefeld),

gp-help@broadinstitute.org

Summary

The IlluminaExpressionFileCreator module converts raw Illumina BeadArray data to the GenePattern GCT file format.

The module extracts the mean value for each probe from a set of Illumina expression IDAT files supplied in a single ZIP archive. Two GCT files are created, one containing the values for the gene probes and the other containing the values for the control probes. Optionally, the results can be collapsed by probe set to a single value per gene for each sample.

Illumina BeadStudio v3.1.3 with GeneExpression plugin v3.4.0 and GenomeStudio 2010.1 with GeneExpression plugin v1.0 were used to generate reference data.

Usage

Required Files

The following files are required input for this module:

- ZIP archive of Illumina IDAT files. The module only supports ZIP archives that contain IDAT files of the same array type.
- An array annotation manifest file, corresponding to the Illumina IDAT files. Some
 manifest files are available through GenePattern, and can be selected from a
 drop-down list. If your manifest file is not in this list, you can provide a manifest
 file via upload or URL. Manifest files are available from the Illumina website at
 http://www.switchtoi.com/annotationfiles.ilmn. This manifest must be uploaded
 in tab-delimited (TXT) format, not the alternate binary (BGX) format.

Optional Parameters

Background subtraction: The module can optionally perform background subtraction on the values prior to creating the gene probe GCT output file. The background value is calculated as the mean of the negative control signal values. This number is subtracted from all gene probe values, but not from the control probe values.



Collapse Mode: Choose whether to collapse probes into a single value. Where multiple probes correspond to the same gene, the results for each sample can optionally be collapsed to a single value. The available options for *collapse mode* are:

- none: do not collapse values; this is the default
- max: use the maximum of all the probe values for each gene
- median: find and use the median of all the probe values for each gene

When the probes are not collapsed, each value in the GCT output file will be listed by probe ID with the gene name as the description. When probes are collapsed, the values will be listed by gene name and the description will be taken from either the Gene_Title field of the CHIP file (if present) or else from the Definition field of the manifest.

The GCT output for the control probes will always be listed by probe ID, using the Reporter_Group_Name field from the manifest as the description.

Optional Files

- <u>CHIP file</u>: By supplying an optional CHIP file, probes can be mapped to gene names for the purposes of collapsing probes or looking up descriptions. If no CHIP file is provided, the module will use the mapping provided in the annotation manifest. You can select a CHIP file from the drop-down list, or provide a CHIP file via upload or URL.
- <u>CLM file</u>: By default, sample names are derived from the IDAT file names in the ZIP archive. Samples can be annotated by providing a CLM file. The CLM file allows the user to specify a set of IDAT files and sample names to be used in creating the GCT file. If a CLM file is specified, only the samples listed in that file will be included in the GCT output and a <u>CLS</u> file will be output. Any IDAT files in the input ZIP that are absent from the CLM file will be ignored by the module.

The left-to-right order of the sample columns in the GCT output will match the top-to-bottom order of the samples in the CLM. With no CLM present, the sample order may be unpredictable: they will be ordered as they are encountered while unpacking the ZIP.

References

- Bead Studio User Guide, version 3.1.3, Illumina.
- Bead Studio Gene Expression Module User Guide, version 3.4.0, Illumina
- Genome Studio Gene Expression Module User Guide, version 1.0, Illumina
- BeadStudio Normalization Algorithms for Gene Expression Data, Illumina technical note



Parameters

Name	Description
idat zip (required)	The ZIP archive containing Illumina expression IDAT files. Any additional non-IDAT files in the ZIP will be ignored. The module assumes that all IDAT files are of the same array type.
manifest custom manifest (required)	The Illumina annotation manifest for these IDAT files. Select a manifest file from the drop-down list, or provide a custom manifest file via upload or URL. The file provided must be in tab-delimited TXT format.
output file (required)	The GCT output files. The module will use the name of the input ZIP file as the name of the output files by default. Two GCT files are produced: • a file that contains the values for the gene probes (name: <base file="" name=""/> .gct) • a file that contains the values for the control probes (name: <base file="" name=""/> -controls.gct) Optionally, there may also be a CLS file if a CLM file is specified that contains class names (name: <base file="" name=""/> .cls).
background subtraction mode (optional)	Allows the user to choose to subtract background (the mean of the negative control signal values) from probe values; "false" is default.
collapse mode (optional)	Allows the user to choose the mode for collapsing probe sets; possible values are "none" (default), "max", and "median".
chip custom chip (optional)	The CHIP file that maps probes to gene names. Select a CHIP file from the drop-down list, or provide a custom CHIP file via upload or URL. The format is described at http://www.genepattern.org/tutorial/gp_fileformats.html#chip .
clm (optional)	The CLM file that maps IDAT file names to sample names. Provide via upload or URL. The format is described at http://www.genepattern.org/tutorial/gp_fileformats.html#clm .



Input Files

- 1. IDAT Illumina expression files in a single ZIP archive (required) IDAT files must be of the same array type.
- Illumina manifest file in TXT format (required)
 Files available with the module can be selected from the drop-down list or manifest files can be uploaded from file or URL. Other manifest files are available from the Illumina website.
- CHIP file (optional)
 Maps probes to gene names. Files available with the module can be selected from the drop-down list or CHIP files can be uploaded from file or URL.
- 4. CLM file (optional) Map IDAT file names to sample names, specifying a set of IDAT files and sample names to be used in creating the GCT file. If a CLM file is provided, only the samples listed in that file will be included in the GCT output and a CLS file will be output. Any IDAT files in the input ZIP that are absent from the CLM file will be ignored by the module.

Output Files

- Two GCT files are created
 One GCT file contains the values for the gene probes (name: <base file name>.gct)
 and the other GCT file contains the values for the control probes (name: <base file
 name>-controls.gct).
- CLS file (optional)
 A CLS file is generated if a CLM file containing class names (the third column of the file) is specified (name: <base file name>.cls). This CLS file contains this class information, and, along with the GCT files, can be used as input to other GenePattern modules.

Platform Dependencies

Module type: Preprocess & Utilities

CPU type: Any
OS: Any
Language: Java