SNPFileCreator Documentation

Module name: SNPFileCreator
Description: Creates a SNP file from a set of Affymetrix SNP chip CEL files
Author: David Twomey (Broad Institute), gp-help@broad.mit.edu

Summary

The SNPFileCreator module creates a GenePattern .snp file from a set of individual CEL files generated using an Affymetrix SNP chip. The conversion is done using one of four modeling algorithms: Average Difference, PM/MM Difference Model (dChip), Median Probe, or Trimmed Mean. The result is a matrix containing intensity values per probe set in the .snp file format. Chromosome and Physical Location columns in the output file are populated based on the Human Genome of May 2004 (hg17).

Source CEL Files

GenePattern modules run on the GenePattern server. Typically, input files are transferred from your file server to the GenePattern server for processing. However, high-density SNP arrays generate massive amounts of data and transferring large amounts of data between servers can be time consuming. To address this issue, SNPFileCreator provides two ways for you to specify your CEL files:

- **Network directory name**: If your CEL files are in a network directory that can be accessed by the GenePattern server, enter the directory name in the *network directory name* parameter. The SNPFileCreator module reads the CEL files from the networked directory, avoiding the time consuming process of transferring the file to the GenePattern server.

  When you enter the name of the directory, use the file specification format that you would use to access that directory from the GenePattern server. For example, if your GenePattern server is running a UNIX server, you might enter /xchip/genome/data/test; if your GenePattern server is running a Windows server, you might enter j:\data\test or \\gensvr\xchip\genome\data\test.

  One way to confirm that you are using a valid network directory specification is to view the directory from the command line of the machine that is running your GenePattern server.

- **Zip filename**: If you have a zip file that contains your CEL files, select the zip file using the Browse button next to the *zip filename* parameter. The SNPFileCreator module transfers the zip file from your file server to the GenePattern server and begins processing. This is the way most GenePattern modules work with input files.

The GenePattern team recommends using network file specifications for large SNP files.

Output Files

Use the *output file* parameter to specify the name of the .snp file to be created. To avoid confusion, always use the .snp file extension. If you enter just the file name, SNPFileCreator creates the file on the GenePattern server; this is the default behavior for most GenePattern modules. If you enter the full path and file name for the output file, SNPFileCreator creates the file in the specified directory.

When you create GenePattern pipelines, you often want to specify the output file of one module as the input file of another module. Typically, you do this when you add a task to the pipeline by clicking the *Use output from previous task* check box next to the input file name (see Working with Pipelines in the *GenePattern Web Client Guide*). For this method of chaining modules to work, the output files must be created on the GenePattern server. If you add SNPFileCreator to a GenePattern pipeline and you specify the full path and file name of the .snp output file, to use that .snp file in a subsequent module of the pipeline, you must specify its full path and file name.

If you create the output file on the GenePattern server, the server can track the location of the output file, and a GenePattern pipeline can use the output file as an input file for a subsequent task. If you specify the full path and file name of the output file, the server writes the file to that location, but does not store that location; therefore, a GenePattern pipeline cannot use that output file as an input file for a subsequent task.

References

1. Cheng Li and Wing Hung Wong (2001) Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. Genome Biology 2:research0032.1-0032.11, doi:10.1186/gb-2001-2-8-research0032 (http://genomebiology.com/2001/2/8/research/0032.1).

Parameters

Name	Description
chip type	Affymetrix SNP chip used to generate the CEL files. Options are: • 50K Hind (default) • 50K Xba • 250K Sty • 250K NSP
network directory name	Name of the networked directory that contains the CEL files and, optionally, the associated .txt genotype call files. The directory must be accessible to the GenePattern server. Specify either *network directory name* or a *zip filename*. If you do not include the genotype call files, all calls in the output file are set to No.
zip filename	Name of the zip file that contains the CEL files and, optionally, the associated .txt genotype call files. Specify either *network directory name* or a *zip filename*. If you do not include the genotype call files, all calls in the output file are set to No.

normalization method	Normalization method: • Invariant Set Normalization (dChip) (default). Use the *reference method* parameter to select a reference. • Quantile Normalization (dChip). Uses the median column as a reference. The two methods are fairly equivalent; Quantile Normalization is faster.
reference method	Reference for Invariant Set Normalization: • Median (default) • Reference Supplied. Use the *reference filename* parameter to specify the reference file.
reference filename	Name of the reference file for normalization. Used only when *normalization method* is Invariant Set Normalization and *reference method* is Reference Supplied.
model method	Method used to determine the intensity value for each SNP based on the intensity levels of the probes in each probe set: • Average Difference • PM/MM Difference Model (dChip) (default) • Median Probe • Trimmed Mean
allele specific	Intensity values per probe set to include in the output file: • Non Allele-Specific (default). Determines one intensity value per probe set. The output file contains two columns per sample: intensity value and call. • Allele-Specific. Determines an intensity value for each allele per probe set. The output file contains three columns per sample: intensity value for allele A, intensity value for allele B, and call. **Note**: Not all SNP modules accept allele-specific snp files. Check the documentation of the modules that you are interested in using before creating your snp files.
sort snp file	Sort the SNPs in the output file: • Sort (default). Sort SNPs by chromosome and physical location. Probe sets that do not have location information are omitted from the output file. **Note**: The SNPViewer module requires that the SNPs be sorted. • Do no sort.
output file	Name of the output file, including the .snp extension. If you specify a full path name, SNPFileCreator creates the file in the specified directory. If you specify just the file name, SNPFileCreator creates the file on the GenePattern server. For more information, see the “Output File” section above.

Return Value

1. .snp file (raw intensity value per probe)

Platform Dependencies

Task type:	SNP Analysis
CPU type:	any
OS:	any
Java JVM level:	1.5
Language:	R