GenePattern

SNPMultipleSampleAnalysis Documentation

Module name:	SNPMultipleSampleAnalysis
Description:	Analysis of copy number aberration data for the identification of concordant aberration across multiple samples.
Author:	Mitchell Guttman (Broad Institute) gp-help@broad.mit.edu
Date:	05/29/07
Release:	1.1

Summary:
SNPMultipleSampleAnalysis (SNP-MSA) identifies regions of concordance across a class of samples. This module is an implementation of the Multiple Sample Analysis (MSA) algorithm. MSA identifies regions of significant concordant structure across multiple samples in a statistically rigorous, multiple testing corrected, non-parametric framework. MSA is used to specifically look for concordant copy number aberration or copy number variation in a group of samples. It has been tested on Microarray based Comparative Genomic Hybridization (aCGH) and Single Nucleotide Polymorphism (SNP) chips.

References:
Manuscript under review visit www.cbil.upenn.edu/MSA for update and referencing

Parameters:

Name	Description
input.filename	The input file to analyze (.cn, .xcn, .snp, .txt)
save.dir	Directory to save to
permutations	Number of permutations to use
num.tests	Number of MSA Tests to use
num.positions.bin	Number of positions per bin
resolution	The size of the plotted chromosome diagram. The smaller the resolution the more information is plotted and the smaller the aberrations that can be detected. However, smaller resolutions require more memory. Practically, the default value of 250,000 can be used for most arrays with a smaller 50,000-100,000 resolution for the Affy 250k.
cytoband.filename	Cytoband file to use
analyze.by	Analyze by arms or chromosomes
chromosomes	Comma delimited list of chromosomes to analyze (chr1, chr2, chr22) Running all chromosomes simultaneously can take a long time for large arrays. It is advised that you break up genome runs into smaller batches.
save.images	Saves images to the specified directory (if not specified it will not save images)
significance.level	The value at which a region is called significant
run.name	The name of the job

Minimum Required Files:

In order to run SNP-MSA the minimum required files needed for analysis is a data file and a Cytoband file. The default Cytoband file (Human Hg18) is provided with the package and will be used unless otherwise specified.

Experiment File:

The input file can be a .cn, .xcn, or log ratio file. The .xcn and .cn formats are described on the GenePattern website under file formats. If these formats are used then the file extensions will be read to determine the file format. If the file extension ends in .snp it will be read as a .xcn file.

Alternatively, if the file extension is any other string then it will be read as a log ratio file. This file format is a tab-delimited file containing the SNP ID its chromosome and position and the log value of test to reference for each experiment. The first line is a header line. The input file looks like:

ID	Chromosome	Start	Exp0	Exp1	Exp2	Exp3	Exp4	Exp5	Exp6
SNP0	1	0	-2.00561	-0.30581	0.382339	-1.4119	0.459363	0.464569	1.138328
SNP1	1	2460000	1.124908	0.935039	1.48414	-0.53269	2.819285	1.41112	1.264908
SNP2	1	4920000	-0.5868	0.302385	-0.32557	1.011326	-2.50569	1.234803	1.172301
SNP3	1	7380000	1.053332	-0.03961	0.129179	-0.52697	-0.99172	1.947016	-0.58448
SNP4	1	9840000	-1.49996	0.336802	-0.50181	0.426686	-0.2287	0.385609	0.629297
SNP5	1	12300000	1.940396	-0.5761	-2.33183	1.445521	0.756655	0.544598	0.980779
SNP6	1	14760000	-0.24479	-0.0231	-0.20732	-0.61816	-0.50545	0.171149	0.18168
SNP7	1	17220000	-0.49603	1.286371	-0.43136	1.73766	-0.43588	0.487942	-1.36223
SNP8	1	19680000	-0.61494	0.315023	1.023693	-0.26873	0.896739	-0.94198	0.195279
SNP9	1	22140000	1.058912	-0.72203	-0.33089	1.33298	-0.03638	0.366303	0.394091
SNP10	1	24600000	-0.6968	-0.84599	-0.78316	0.075815	-0.47813	-0.65001	-1.46863
SNP11	1	27060000	0.022487	0.941432	-0.10405	-0.69255	-0.87946	-0.11485	0.434379
SNP12	1	29520000	1.410611	-0.59974	0.797803	0.504457	-0.28944	1.514422	0.716182
SNP13	1	31980000	0.764305	1.530071	-0.65763	1.092064	-1.57382	0.75783	0.082763
SNP14	1	34440000	-1.14251	2.153274	-0.52284	1.063675	-0.26182	-1.35722	-2.18321
SNP15	1	36900000	-0.39911	1.624002	-0.89159	-0.23847	1.273964	-0.24617	-0.95801
SNP16	1	39360000	-1.21675	1.827362	-1.00983	0.999052	-1.4738	2.31696	0.205977
SNP17	1	41820000	0.812655	-0.14936	0.747986	-0.90268	0.318832	0.473987	0.361568

It is similar to the .cn format except the values are log ratios rather than copy numbers. The relationship between the copy number values and the log ratio are specified as $CN = 2^{\log(T/R)+1}$.

The experiment names that are used in all output files are taken from the header lines of the input file.

Missing Values:

Currently missing values in the input file will cause SNP-MSA to generate an error. In order to avoid this problem replace all missing values with flags. A flag in SNP-MSA is defined as a -999 entry in a cell. Any value less than -999 will be read as a flag as well. In the .cn and .xcn files any value less than 0 is read as a flag. If a value needs to be skipped for any reason these flags will force this skip. A GenePattern module will be written to fill in these flags for all missing values. The MSA algorithm deals with missing values full details can be found in the technical specifications at www.cbil.upenn.edu/MSA.

Return Value:

<chromosome>HM.jpg	For each chromosome a HM.jpg file will be output which represents a plot of the raw data and the MSA confidence at each location plotted along the chromosomal ideogram
<chromosome>MSAMV.jpg	A MSAMV.jpg plots the MSA Merged View which shows each region of concordant aberrations and the single samples that contributed significance.
<chromosome>freq.jpg	A freq.jpg file plots the frequency of significant aberration at each location along the chromosomal idiogram
<run.name>MSAHeatMapFile.hmv	Output of MSA with raw data values and amplification and deletion confidence.
<run.name>MSAConfidenceFile.txt	A text file representing the confidence at each location of the genome.
<run.name>MSAConfidenceFile.txt.MSASSV	A text file representing the single sample values at each location of the genome

Output File Descriptions:

MSA Confidence File

This file contains p-values for each location of the genome. The first column is the location followed by the multiple testing corrected p-value and then the confidence. The confidence is simply (1-pvalue).

MSA Single Sample Values

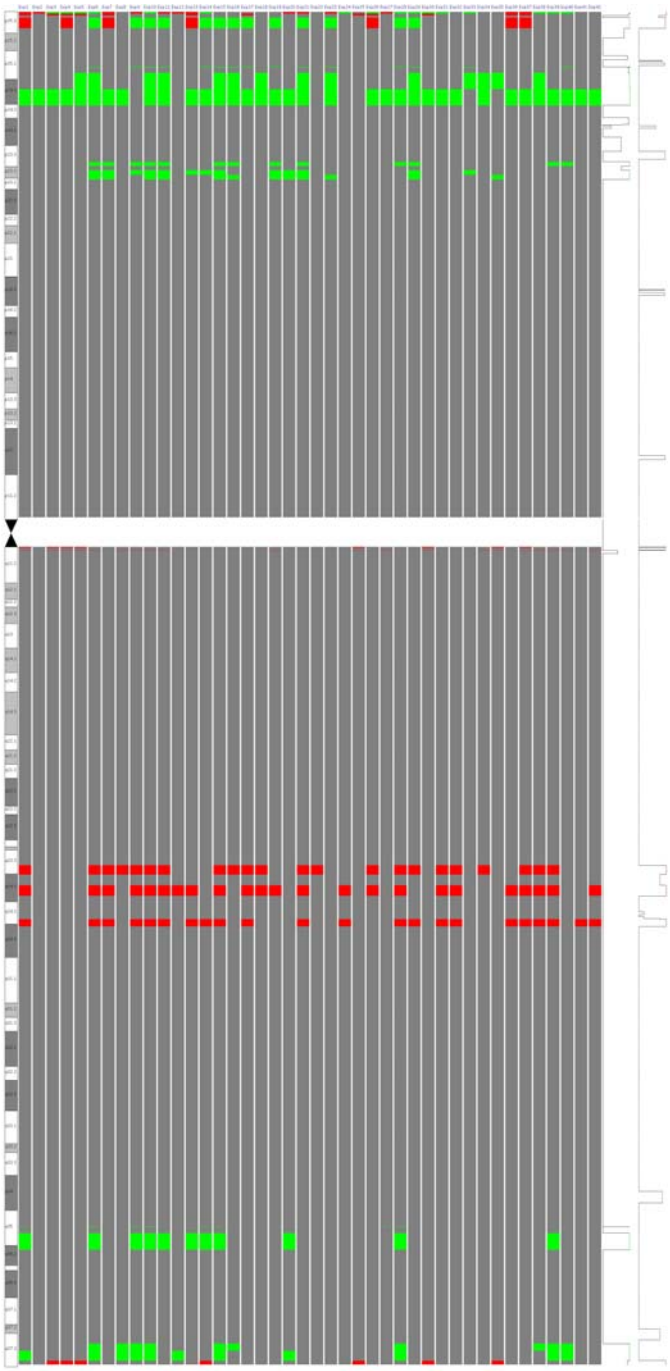
This is a text file that contains each location followed by the confidence and then a call for each sample. This file should only be referenced for a p-value less than a given significance level. The calls are meaningless for p-values greater than the p-value.

HeatMap File

This is text file containing the raw data for each location along with its associated amplification and deletion p-values.

GenePattern

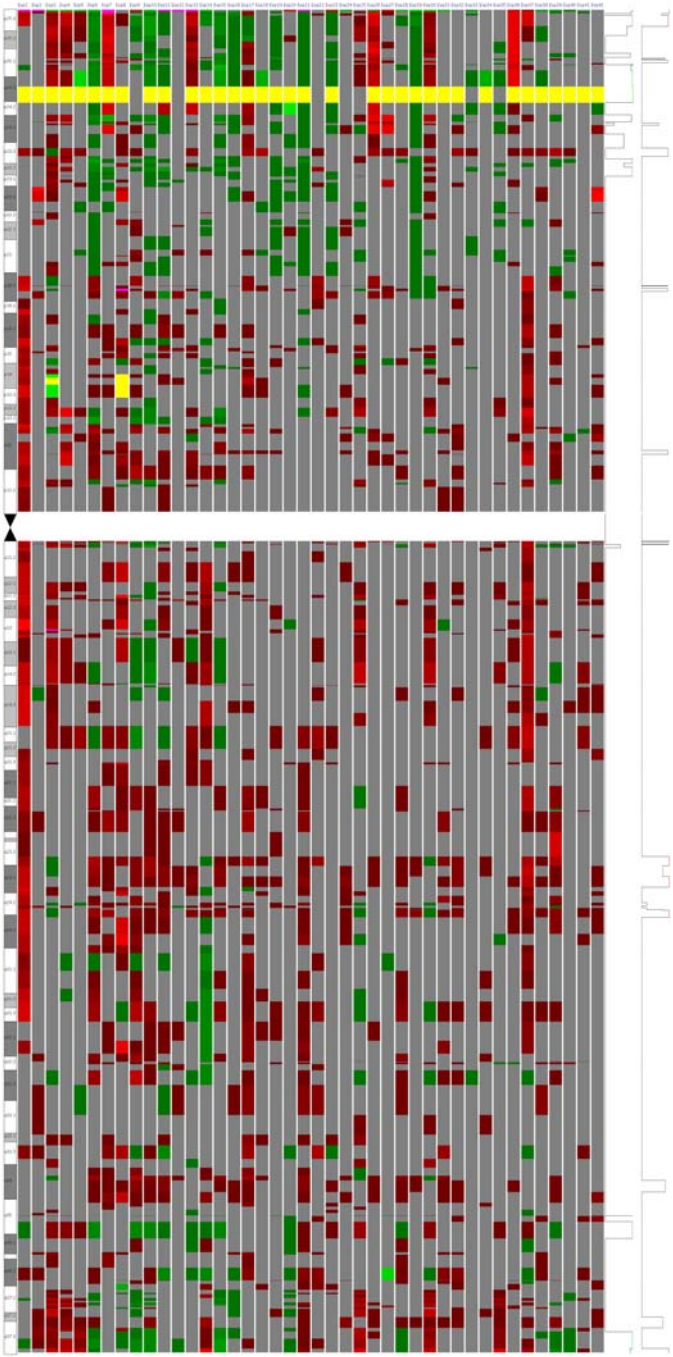
MSA Merged View Files (<chromosome>MSAMV.jpg)



Only regions that are significant at the significance level specified are used to determine single sample calls. Given an aberrant region, a sample is colored green if it contributed a gain to the multiple sample significance. Similarly, a sample is colored red if it contributed a loss to the multiple sample significance.

To the right a line graph is plotted that tracks the p-value for each location. If the p-value is significant at a given region then the line is colored green or red based on gain or loss significance.

Pseudo-HeatMap Files (<chromosome>HM.jpg)



This image represents a pseudo-heatmap of the aberrations in the data. This is plotted regardless of significance and is intended to allow the user to determine approximate levels of

aberrations. They are colored based on the dynamic range of aberrations tested over by MSA. A line graph is plotted that tracks the p-value by location.

Frequency Plots

These represent the frequency of aberration for each significant location. It is only plotted if a region is significant at the specified significance level. There are 3 dotted lines on the right and the left. The right represents gain and the left loss. The tick lines represent 25%, 75%, and 100% frequency.

Note:
An internalFiles directory is created to hold intermediate results. This directory appears in the list of result files, but can be disregarded.

Platform dependencies:

Task type:	SNP Analysis
CPU type:	any
OS:	any
Java JVM level:	n/a
Language:	Java
Support files:	n/a