GenePattern

GISTIC2 Documentation

Description: Genomic Identification of Significant Targets in Cancer

Author: Gad Getz, Rameen Beroukhim, Craig Mermel, Steve Schumacher, and Jen Dobson, gistic-help@broadinstitute.org

Release 2.0.16

Summary

The GISTIC module identifies regions of the genome that are significantly amplified or deleted across a set of samples. Each aberration is assigned a G-score that considers the amplitude of the aberration as well as the frequency of its occurrence across samples. False Discovery Rate q-values are then calculated for the aberrant regions, and regions with q-values below a user-defined threshold are considered significant.

For each significant region, a “peak region” is identified, which is the part of the aberrant region with greatest amplitude and frequency of alteration. In addition, a “wide peak” is determined using a leave-one-out algorithm to allow for errors in the boundaries in a single sample. The “wide peak” boundaries are more robust for identifying the most likely gene targets in the region.

Each significantly aberrant region is also tested to determine whether it results primarily from broad events (longer than half a chromosome arm), focal events, or significant levels of both. The GISTIC module reports the genomic locations and calculated q-values for the aberrant regions. It identifies the samples that exhibit each significant amplification or deletion, and it lists genes found in each “wide peak” region.

Note: The GISTIC module is memory-intensive.

References

Mermel C, Schumacher S, et al. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biology*. 2011;12:R41.

Beroukhim R, Mermel C, et al. The landscape of somatic copy-number alteration across human cancers. *Nature*. 2010;463:899-905.

Parameters

Name	Option String	Description
refgene.file (required)	-refgene	The reference file including cytoband and gene location information. (Default: hg19)

seg.file (required)	-seg	The segmentation file contains the segmented data for all the samples identified by GLAD, CBS, or some other segmentation algorithm. (See GLAD file format for more information.) It is a six-column, tab-delimited file with an optional first line identifying the columns. Positions are in base pair units.
markers.file (required)	-mk	The markers file identifies the marker names and positions of the markers in the original dataset before segmentation. It is a three-column, tab-delimited file with an optional header. Markers are sorted by genomic position if the file is not already in that order.
array.list.file	-alf	The array list file is an optional file identifying the subset of samples to be used in the analysis. It is a one-column file with an optional header. The sample identifiers listed in the array list file must match the sample names given in the segmentation file.
cnv.file	-cnv	Copy number variant file. There are two options for the CNV file. The first option allows CNVs to be identified by marker name. The second option allows the CNVs to be identified by genomic location. CNVs and CNV-like artifacts can be seen in segmented normal data. Use this optional input file to exclude all regions seen as significant in the segmented normal data from the GISTIC analysis, because they will cause false peaks.
gene.gistic (required)	-genegistic	Flag indicating that the gene GISTIC algorithm should be used to calculate the significance of deletions at a gene level instead of a marker level. (Default: yes)
amplifications. threshold (required)	-ta	Threshold for copy number amplifications. Regions with a log2 ratio above this value are considered amplified. (Default: 0.1)
deletions. threshold (required)	-td	Threshold for copy number deletions. Regions with a log2 ratio below the negative of this value are considered deletions. (Default: 0.1)

join.segment.size (required)	-js	Smallest number of markers to allow in segments from the segmented data. Segments that contain a number of markers less than or equal to this number are joined to the adjacent segment, closest in copy number. (Default: 4)
qv.thresh (required)	-qvt	Significance threshold for q-values. Regions with q-values below this number are considered significant. (Default: 0.25)
remove.x (required)	-rx	Flag indicating whether to remove data from the X chromosome before analysis. (Default: yes)
confidence.level (required)	-conf	Confidence level used to calculate the region containing a driver. (Default: 0.75)
run.broad.analysis (required)	-broad	Flag indicating whether an additional broad-level analysis should be performed. (Default: no).
broad.length.cutoff (required)	-brlen	Threshold used to distinguish broad from focal events, given in units of fraction of chromosome arm. (Default: 0.98)
max.sample.segs	-maxseg	Maximum number of segments allowed for a sample in the input data. Samples with more segments than this threshold are excluded from the analysis. (Default: 2500)
arm.peel	-armpeel	Whether to perform arm level peel off, which helps separate peaks and clean up noise. (Default: no)
output.prefix (required)	-fname	The prefix for the output file name.

Input Files

1. Reference Genome File (-refgene) (REQUIRED)

   The reference genome file contains information about the location of genes and cytobands on a given build of the genome. Reference genome files are created in

MATLAB™ and are not viewable with a text editor. The GISTIC 2.0 release includes the following reference genomes: *hg16.mat*, *hg17.mat*, *hg18.mat*, and *hg19.mat).*

2. Segmentation File (-seg) (REQUIRED)

   The segmentation file contains the segmented data for all the samples identified by GLAD, CBS, or some other segmentation algorithm. (See GLAD file format in the GenePattern file formats documentation.) It is a six column, tab-delimited file with an optional first line identifying the columns. Positions are in base pair units. Seg.CN values should be log transformed; if not, GISTIC will automatically log transform the values. The column headers are:

   1. *Sample* (sample name)
   2. *Chromosome* (chromosome number)
   3. *Start Position* (segment start position, in bases)
   4. *End Position* (segment end position, in bases)
   5. *Num markers* (number of markers in segment)
   6. *Seg.CN* (log2() -1 of copy number)]

3. Markers File (-mk) (REQUIRED)

   The markers file identifies the marker names and positions of the markers in the original dataset (before segmentation). It is a three-column, tab-delimited file with an optional header. The column headers are:

   1. *Marker Name* (marker name)
   2. *Chromosome* (chromosome number)
   3. *Marker Position* (in bases)

4. Array List File (-alf) (OPTIONAL)

   The array list file is an optional file identifying the subset of samples to be used in the analysis. It is a one column file with an optional header (*array*). The sample identifiers listed in the array list file must match the sample names given in the segmentation file.

5. CNV File (-cnv) (OPTIONAL)

   There are two options for the CNV file. The first option allows CNVs to be identified by marker name. The second option allows the CNVs to be identified by genomic location.
   Option #1: A two-column, tab-delimited file with an optional header row. The marker names given in this file must match the marker names given in the markers file. The CNV identifiers are for user use and can be arbitrary. The column headers are:

   1. *Marker Name*
   2. *CNV Identifier*

   Option #2: A 6-column, tab-delimited file with an optional header row. The *CNV Identifier*, *Narrow Region Start*, and *Narrow Region End* are for user use and can be arbitrary. The column headers are:

   1. *CNV Identifier*
   2. *Chromosome*
   3. *Narrow Region Start*
   4. *Narrow Region End*
   5. *Wide Region Start*
   6. *Wide Region End*

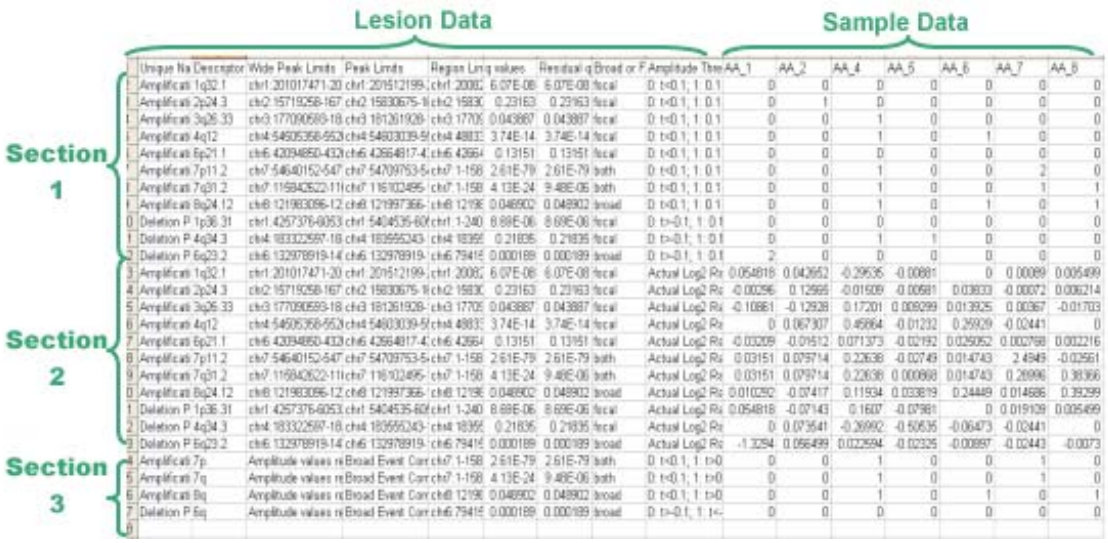
Output Files

1. All Lesions File (all_lesions.conf_*XX*.txt, where *XX* is the confidence level)

   The all lesions file summarizes the results from the GISTIC run. It contains data about the significant regions of amplification and deletion as well as which samples are amplified or deleted in each of these regions. The identified regions are listed down the first column, and the samples are listed across the first row, starting in column 10.
   Region Data
   Columns 1-9 present the data about the significant regions as follows:

   1. *Unique Name:* A name assigned to identify the region.
   2. *Descriptor:* The genomic descriptor of that region.
   3. *Wide Peak Limits:* The "wide peak" boundaries most likely to contain the targeted genes. These are listed in genomic coordinates and marker (or probe) indices.
   4. *Peak Limits:* The boundaries of the region of maximal amplification or deletion.
   5. *Region Limits:* The boundaries of the entire significant region of amplification or deletion.
   6. *q-values:* The q-value of the peak region.
   7. *Residual q-values:* The q-value of the peak region after removing ("peeling off") amplifications or deletions that overlap other more significant peak regions in the same chromosome.
   8. *Broad or Focal:* Identifies whether the region reaches significance due primarily to broad events (called "broad"), focal events (called "focal"), or independently significant broad and focal events (called "both").
   9. *Amplitude Threshold:* Key giving the meaning of values in the subsequent columns associated with each sample.

   Sample Data
   Each of the analyzed samples is represented in one of the columns following the lesion data (columns 10 through end). The data contained in these columns varies slightly by section of the file.
   The first section can be identified by the key given in column 9 – it starts in row 2 and continues until the row that reads *Actual Copy Change Given*. This section contains summarized data for each sample. A '0' indicates that the copy number of the sample was not amplified or deleted beyond the threshold amount in that peak region. A '1' indicates that the sample had low-level copy number aberrations (exceeding the low threshold indicated in column 9), and a '2' indicates that the sample had high-level copy number aberrations (exceeding the high threshold indicated in column 9).
   The second section can be identified as the rows in which column 9 reads *Actual Copy Change Given*. The second section exactly reproduces the first section, except that here the actual changes in copy number are provided rather than zeroes, ones, and twos.
   The final section is similar to the first section, except that here only broad events are included. A "1" in the samples columns (columns 10+) indicates that the median copy number of the sample across the entire significant region exceeded the threshold given in column 9. That is, it indicates whether the sample had a geographically extended event, rather than a focal amplification or deletion covering little more than the peak region.



2. Amplification Genes File (Amp_genes.conf_*XX*.txt, where *XX* is the confidence level)

   Tables of amplification peaks, followed by the genes contained in them, organized in "ragged columns." The amp genes file contains one column for each amplification peak identified in the GISTIC analysis. The first four rows are:

   1. *cytoband*
   2. *q-value*
   3. *residual q-value*
   4. *wide peak boundaries*

   These rows identify the lesion in the same way as the all lesions file.
   The remaining rows list the genes contained in each wide peak. For peaks that contain no genes, the nearest gene is listed in brackets.

3. Deletion Genes File (Del_genes.conf_*XX*.txt, where *XX* is the confidence level)

   Tables of deletion peaks, followed by the genes contained in them, organized in "ragged columns." The del genes file contains one column for each deletion identified in the GISTIC analysis. The file format for the del genes file is identical to the format for the amp genes file.

	A	B	C	D	E	F	G	H	I	J	K	L
1	cytoband	9p21.3	9p21.3	1p36.31	10q23.31	9p24.3	13q14.2	10q23.31	10q26.13	19q13.41	14q31.3	4q34.3
2	q value	2.11E-92	1.23E-42	8.15E-06	1.13E-06	0.0004056	0.00061778	2.29E-05	0.046345	0.10999	0.15544	0.16647
3	residual q va	1.45E-79	7.44E-09	8.15E-06	0.00033684	0.0004056	0.00061778	0.045641	0.046345	0.10999	0.15544	0.16647
4	wide peak bc	chr9:219485	chr9:245441	chr1:425680	chr10:89350	chr9:1-2695	chr13:46638	chr10:90243	chr10:11270	chr19:46920	chr14:56162	chr4:181134
5	genes in wid	CDKN2A	[ELAVL2]	RPL22	PTEN	DMRT1	RCBTB2	ACTA2	ACADSB	A1BG	ACTN1	SLC25A4
6		CDKN2B		KCNAB2		FOXD4	RB1	STAMBPL1	ADAM8	AP2A1	ACYP1	CASP3
7				ACOT7		SMARCA2	P2RY5	ANKRD22	ADRB1	APOC1	ARG2	DCTD
8				ICMT		VLDLR	FNDC3A		BNIP3	APOC2	ZFP36L1	F11
9				CHD5		DMRT2	CYSLTR2		CASP7	APOC4	ENTPD5	ACSL1

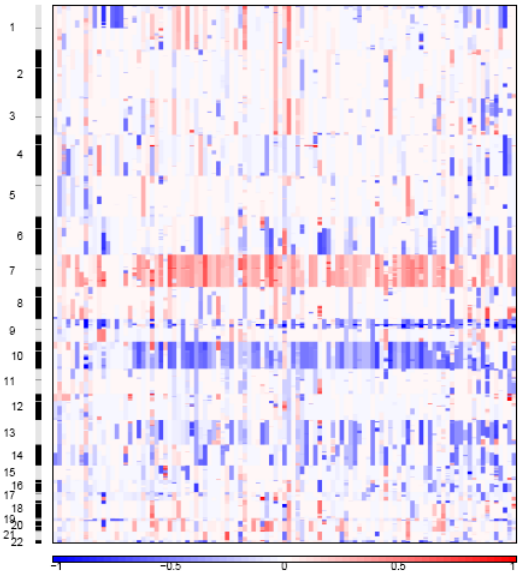
4. GISTIC Scores File (scores.gistic)

   A table of segmented q-values, scores, and amplification/deletion frequencies for the sample set. The scores file lists the q-values [presented as -log10(q)], G-scores, average amplitudes among aberrant samples, and frequency of aberration, across the genome for both amplifications and deletions. The scores file is viewable with the Integrative Genomics Viewer (IGV).

	A	B	C	D	E	F	G	H
1	Type	Chromosom	Start	End	#NAME?	G-score	average amp	frequency
2	Amp	1	328296	3321970	0	0.045449	0.353461	0.133333
3	Amp	1	3464664	5288828	0	0.035435	0.348485	0.12381
4	Amp	1	5307047	5404534	0	0.039989	0.340892	0.133333
5	Amp	1	5432591	6474209	0	0.035435	0.348485	0.12381
6	Amp	1	6605831	7709148	0	0.042375	0.349052	0.133333
7	Amp	1	7788847	9699658	0	0.035435	0.3465	0.12381
8	Amp	1	10307097	10307097	0	0.045198	0.368215	0.12381

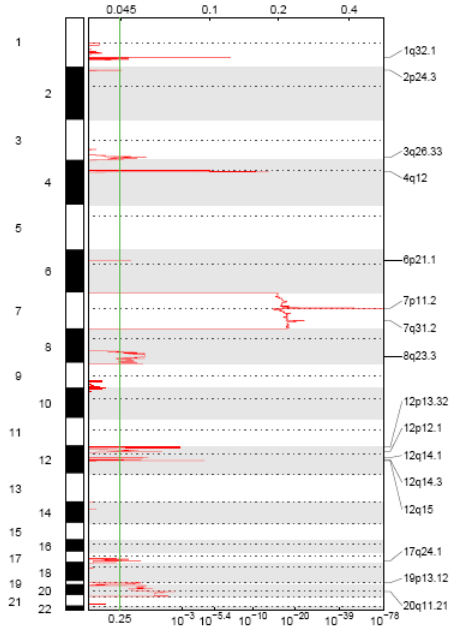
5. Segmented Copy Number (raw_copy_number.pdf and raw_copy_number.png)

   The segmented copy number file (both PDF and PNG) is a heat map image of the segmented copy number profiles in the input data.



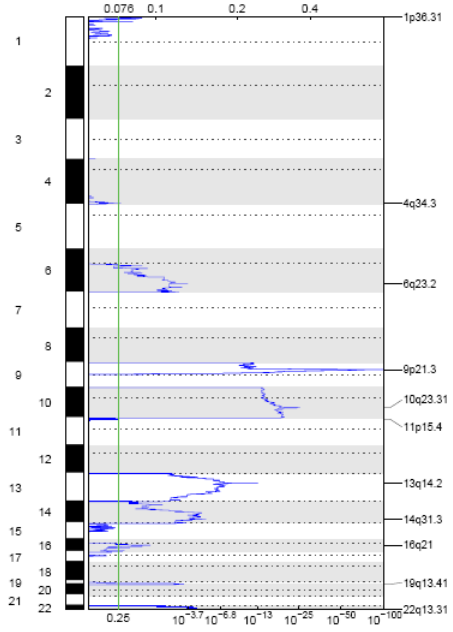
6. Amplification GISTIC plot (amp_qplot.pdf and amp_qplot.png)

   The amplification plot (in both PDF and PNG format) shows the G-scores (top) and q-values (bottom) with respect to amplifications for all markers over the entire region analyzed.



7. Deletion GISTIC plot (del_qplot.pdf and del_quplot.png)

   The deletion plot (in both PDF and PNG format) shows the G-scores (top) and q-values (bottom) with respect to deletions for all markers over the entire region analyzed.



8. all_thresholded.by_genes.txt

   The table in this file is obtained by applying both low- and high-level thresholds to the gene copy levels of all the samples. The entries with value +/- 2 exceed the high-level thresholds for amps/dels, and those with +/- 1 exceed the low-level thresholds but not the high-level thresholds. The low-level thresholds are just the

'amplifications_threshold' and 'deletions_threshold' noise threshold input values (typically 0.1 or 0.3) and are the same for every threshold.

By contrast, the high-level amplification (or deletion) thresholds are calculated on a sample-by-sample basis and are based on the maximum (or minimum) median arm-level amplification (or deletion) copy number found in the sample. The idea, for deletions anyway, is that this level is a good approximation for hemizygous given the purity and ploidy of the sample. The actual cutoffs used for each sample can be found in a table in the sample_cutoffs.txt file.

9. Other result files include:
    - regions_track.conf_XX.bed
    - broad_significance_results.txt (only output if run.broad.analysis is set to "yes")
    - broad_values_by_arm.txt (only output if run.broad.analysis is set to "yes")
    - freqarms_vs_ngenes.pdf (only output if run.broad.analysis is set to "yes")
    - arraylistfile.txt (only output if an array.list.file is provided as input)
    - all_data_by_genes.txt
    - broad_data_by_genes.txt
    - focal_data_by_genes.txt
    - sample_cutoffs.txt
    - amp_qplot.v2.pdf and amp_qplot.v2.ps (do not contain gene labels)
    - del_qplot.v2.pdf and del_qplot.v2.ps (do not contain gene labels)

Troubleshooting

Please see the GenePattern FAQ (http://www.broadinstitute.org/cancer/software/genepattern/doc/faq) for assistance with specific errors.

Example Data

- Example segmentation file (the segmentation file contains segmented data for all the samples identified by some segmentation algorithm)
- Example markers file (the markers file identifies the marker names and positions of the markers in the original dataset before segmentation)
- Example array list file (the array list file is an optional file identifying the subset of samples to be used in the analysis)
- Example CNV file (the optional CNV file identifies CNVs by either marker name or genomic location)

Platform Dependencies

Module type:	SNP Analysis
CPU type:	x86
OS:	64-bit Linux
Language:	MATLAB

GenePattern Module Version Notes

Version	Description
v.5	GISTIC module v.5 contains the update to GISTIC 2.0.16. There are extensive changes to the algorithms and result files, from GISTIC 1.0. See Mermel et al (2011) for more information about the update.
v. 6	Added description of the all_thresholded.by_genes.txt output file to the documentation.