GISTIC2 Documentation

Description: Genomic Identification of Significant Targets in Cancer

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

Release 2.0

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.

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;In Press.

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
base.dir (required)	-b	The directory in which all output files will be saved.
amplifications. threshold	-ta	Threshold for copy number amplifications. Regions with a log2 ratio above this value are considered amplified. (Default: 0.1)

deletions. threshold	-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	-js	Smallest number of markers to allow in segments from the segmented data. Segments that contain fewer than this number of markers are joined to the neighboring segment that is closest in copy number. (Default: 4)
gv.thresh	-gvt	Significance threshold for q-values. Regions with q-values below this number are considered significant. (Default: 0.25)
extension	-ext	Extension to append to all output files. (Default: ’’, no extension)
remove.x	-rex	Flag indicating whether to remove data from the X chromosome before analysis. Allowed values are {1,0}. (Default: 1, remove X)
cap_val	-cap	Minimum and maximum cap values on analyzed data. Regions with a log2 ratio greater than the cap are set to the cap value; regions with a log2 ratio less than -cap value are set to -cap. Values must be positive. (Default: 1.5)
run.broad. analysis	-broad	Flag indicating that an additional broad-level analysis should be performed. Allowed values are {1,0}. (Default: 0, no broad analysis).
broad.length. cutoff	-brlen	Threshold used to distinguish broad form focal events, given in units of fraction of chromosome arm. (Default: 0.98)
use.two.sided	-twosides	Flag indicating that a two-dimensional quadrant figure should be created as part of a broad analysis. Allowed values are {1,0}. (Default: 0, no figure).

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)
resolution	-res	Resolution used to create the empirical distributions used to estimate background probabilities. Lower values generate more accurate results at a cost of greater computation time. (Default: 0.05)
conf.level	-conf	Confidence level used to calculate the region containing a driver. (Default: 0.75)
do.gene.gistic	-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. Allowed values are {1,0}. (Default: 0, no gene GISTIC).
do.arbitration	-arb	Flag for using the arbitrated peel-off algorithm when resolving the significance of overlapping peaks. Allowed values are {1,0}. (Default: 1, use arbitrated peel-off)
peak.types	-peak_type	Method for evaluating the significance of peaks, either *robust* (Default) or *loo* for leave-one-out.
save.disk.space	-smalldisk	Flag indicating that large MATLAB$^{\text{TM}}$ objects should not be saved to disk. Allowed values are {1,0}. (Default: 0, save large).
use.segarray	-smallmem	Flag indicating that the SegArray memory compression scheme should be used to reduce the memory requirements of the computation for large data sets. Computation is somewhat slower with memory compression enabled. Allowed values are {1,0}. (Default: 1, compress memory)
write.gene.files	-savegene	Flag indicating that gene files should be saved. Allowed values are {1,0}. (Default: 0, don't save gene files)

save.seg.data	-saveseg	Flag indicating that the segmented data used as input for the GISTIC analysis should be saved. Allowed values are {1,0}. (Default: 1, save segmented input data)
write.data.files	-savedata	Flag indicating that native MATLAB^{TM} output files should be saved in addition to text data. Allowed values are {1,0}. (Default: 1, save MATLAB^{TM} files)
verbosity	-v	Integer value indicating the level of verbosity to use in the program execution log. Suggested values are {0,10,20,30}. 0 sets no verbosity; 30 sets high level of verbosity. (Default: 0)

Input Files

1. Segmentation File (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)]

2. Markers File (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)

3. 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^{TM} and are not viewable with a text editor. The GISTIC 2.0 release has three reference genomes located in the *refgenefiles* directory: *hg16.mat*, *hg17.mat*, and *hg18.mat*. *(Bobbie) strongly suggest that we add hg19 as well and make this a drop down, which option for custom – I have the hg19.mat)*

4. Array List File (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 (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_file.txt)

   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 Log Value*. 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 Log2 Ratio." The second section exactly reproduces the first section, except that here the exact log2 ratios are provided rather than zeroes, ones, and twos.
The final section is similar to the first section, except that here only broad events (called "broad") and independently significant broad and focal events (called "both") 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.
**This image is not in the 2.0 doc, but I think it's very helpful – Steve, given that nothing seems to have changed in the description of this file –is this image still valid?Actually this doesn't match the examples from the 2.0 Doc – I'll need to redo this for Judy – remind me(Bobbie).*

2. Amplification Genes File (Amp_genes.txt)

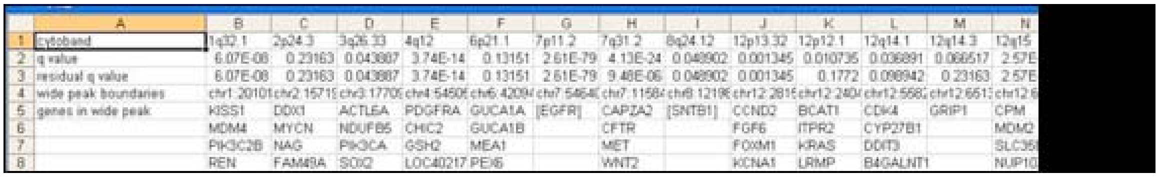
The amp genes file contains one column for each amplification identified in the GISTIC analysis. The first four rows are:

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.txt)

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.
This is how it appeared, copied from the doc… I (Bobbie) can redo this for better resolution, but it's a place holder for now.

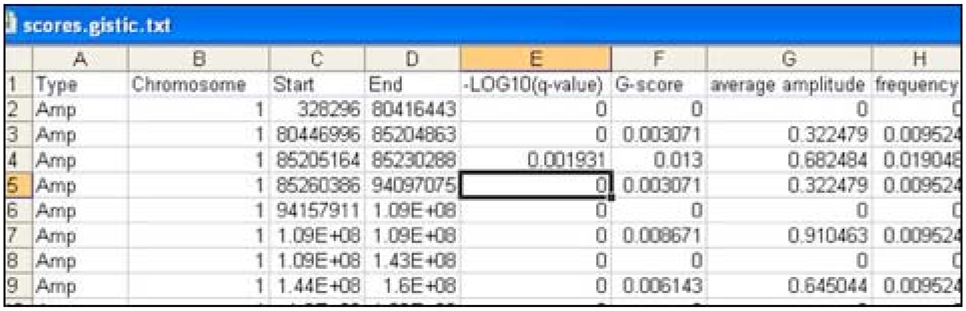


	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	cytoband	1q32.1	2p24.3	3q26.33	4q12	6p21.1	7p11.2	7q31.2	8q24.12	12p13.32	12p12.1	12q14.1	12q14.3	12q15
2	q value	6.07E-08	0.23163	0.043887	3.74E-14	0.13151	2.61E-79	4.13E-24	0.048902	0.001345	0.010735	0.036891	0.066517	2.57E
3	residual q value	6.07E-08	0.23163	0.043887	3.74E-14	0.13151	2.61E-79	9.48E-06	0.048902	0.001345	0.1772	0.098942	0.23163	2.57E
4	wide peak boundaries	chr1:20101	chr2:15715	chr3:17706	chr4:54506	chr6:42094	chr7:54640	chr7:11584	chr8:12196	chr12:2815	chr12:2404	chr12:5682	chr12:6513	chr12:6
5	genes in wide peak	KISS1	DDX1	ACTL6A	PDGFRA	GUCA1A	[EGFR]	CAPZA2	[SNTB1]	CCND2	BCAT1	CDK4	GRIP1	CPM
6		MDM4	MYCN	NDUFB5	CHIC2	GUCA1B		CFTR		FGF6	ITPR2	CYP27B1		MDM2
7		PIK3C2B	NAG	PIK3CA	GSH2	MEA1		MET		FOXM1	KRAS	DDIT3		SLC35
8		REN	FAM49A	SOX2	LOC40217	PEX6		WNT2		KCNA1	LRMP	B4GALNT1		NUP10

4. Gistic Scores File (scores.gistic.txt)

   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).

   Judy – I'll get you a better image.

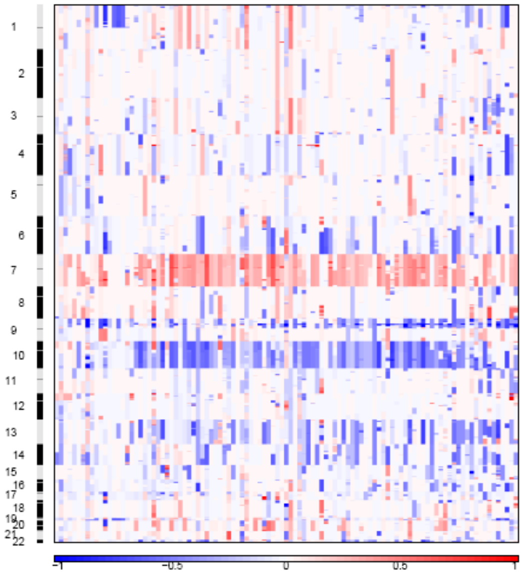


scores.gistic.txt

	A	B	C	D	E	F	G	H
1	Type	Chromosome	Start	End	-LOG10(q-value)	G-score	average amplitude	frequency
2	Amp	1	328296	80416443	0	0	0	0
3	Amp	1	80446996	85204863	0	0.003071	0.322479	0.009524
4	Amp	1	85205164	85230288	0.001931	0.013	0.682484	0.019048
5	Amp	1	85260386	94097075	0	0.003071	0.322479	0.009524
6	Amp	1	94157911	1.09E+08	0	0	0	0
7	Amp	1	1.09E+08	1.09E+08	0	0.008671	0.910463	0.009524
8	Amp	1	1.09E+08	1.43E+08	0	0	0	0
9	Amp	1	1.44E+08	1.6E+08	0	0.006143	0.645044	0.009524

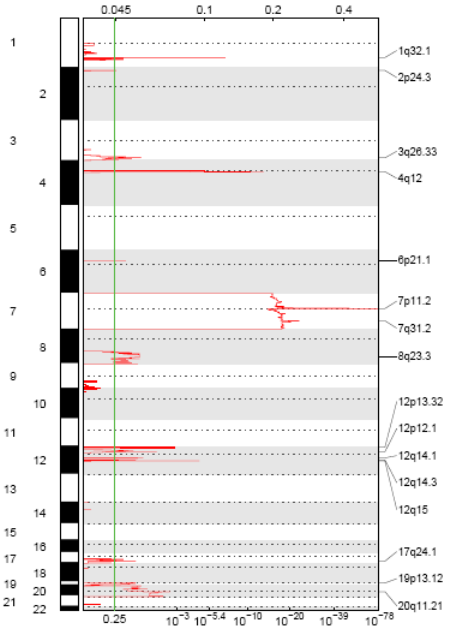
5. Segmented Copy Number (xx.segmented_copy_number.pdf*)*

   The segmented copy number PDF file is a heat map image of the segmented copy number profiles in the input data.



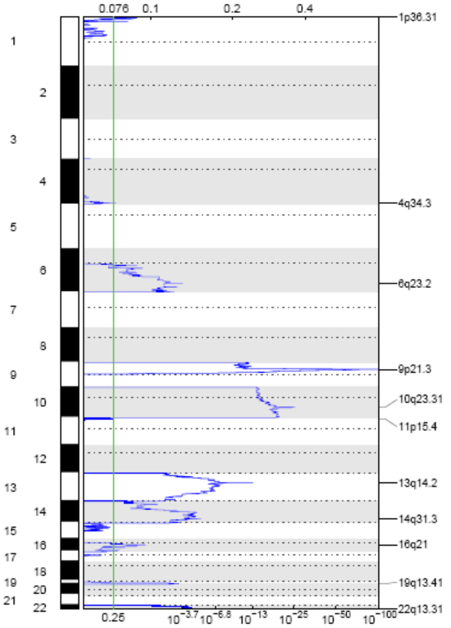
6. Amplification GISTIC plot (xx.amplification.pdf)

   The amplification PDF is a plot of the G-scores (top) and q-values (bottom) with respect to amplifications for all markers over the entire region analyzed.



7. Deletion GISTIC plot (xx.deletion. pdf)

   The deletion PDF is a plot of the G-scores (top) and q-values (bottom) with respect to deletions for all markers over the entire region analyzed.



Troubleshooting

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

Example Data

- Example segmentation file [TO BE LINKED: currently at /xchip/sqa/Modules/GISTIC/GISTIC2.0_stdAlone/examplefiles]
- Example markers file
- Example array list file
- Example CNV file

Platform Dependencies

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