GenePattern

GISTIC Documentation

Description:	Genomic Identification of Significant Targets in Cancer
Author:	Jen Dobson, Rameen Beroukhim, Gad Getz
Date:	2 June 2008
Release:	1.0

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

- Beroukhim R, Getz G, et al. (2007). “Assessing the significance of chromosomal abberations in cancer: Methodology and application to glioma.” *Proc Natl Acad Sci,* 104:20007-20012.

Input Parameters

Name	**Description**
refgene file	The cytoband file to use in the analysis. Allowed values: {Human Hg18, Human Hg17, Human Hg16}. (Default=Human Hg16)
amplifications threshold	Threshold for copy number amplifications. Regions with a log2 ratio above this value are considered amplified. (Default=0.1)
deletions	Threshold for copy number deletions. Regions with a log2

threshold	ratio below the negative of this value are considered deletions. (Default=0.1)
join segment size	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)
qv thresh	Threshhold for q-values. Regions with q-values below this number are considered significant. (Default=0.25)
extension	Extension to append to all output files.
remove x	Flag indicating whether to remove data from the X-chromosome before analysis. Allowed values= {1,0}. (Default=1(yes))
seg file	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.
markers file	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. If not already, markers are sorted by genomic position.
array list file	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	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.

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.

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)

Example Segmentation File

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
(2) Chromosome
(3) Marker Position (in bases)

Example Markers File

3. 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.

Example Array List File

4. 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*

Example CNV File

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.

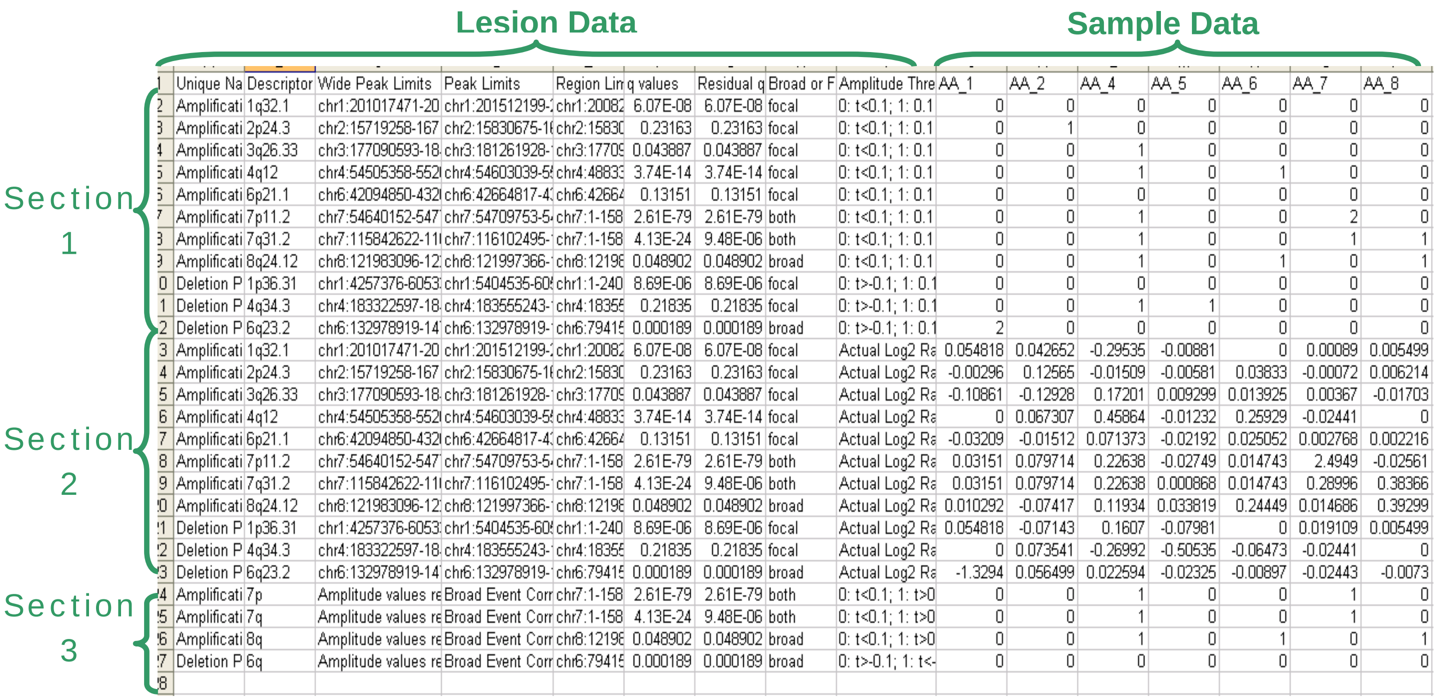
(5) *Region Limits:* The boundaries of the entire significant region of amplification or deletion.
(6) *q-values:* The q-value of the peak region.
(7) *Residual q-values:* The q-value of the peak region after removing (“peeling off”) amplifications or deletions that overlap other more significant peak 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.



1	Unique Na	Descriptor	Wide Peak Limits	Peak Limits	Region Lin	q values	Residual q	Broad or F	Amplitude Thre	AA_1	AA_2	AA_4	AA_5	AA_6	AA_7	AA_8
2	Amplificati	1q32.1	chr1:201017471-20	chr1:201512199-2	chr1:20082	6.07E-08	6.07E-08	focal	0: t<0.1; 1: 0.1	0	0	0	0	0	0	0
3	Amplificati	2p24.3	chr2:15719258-167	chr2:15830675-16	chr2:15830	0.23163	0.23163	focal	0: t<0.1; 1: 0.1	0	1	0	0	0	0	0
4	Amplificati	3q26.33	chr3:177090593-18	chr3:181261928-1	chr3:17709	0.043887	0.043887	focal	0: t<0.1; 1: 0.1	0	0	1	0	0	0	0
5	Amplificati	4q12	chr4:54505358-552	chr4:54603039-55	chr4:48833	3.74E-14	3.74E-14	focal	0: t<0.1; 1: 0.1	0	0	1	0	1	0	0
6	Amplificati	6p21.1	chr6:42094850-432	chr6:42664817-43	chr6:42664	0.13151	0.13151	focal	0: t<0.1; 1: 0.1	0	0	0	0	0	0	0
7	Amplificati	7p11.2	chr7:54640152-547	chr7:54709753-54	chr7:1-158	2.61E-79	2.61E-79	both	0: t<0.1; 1: 0.1	0	0	1	0	0	2	0
8	Amplificati	7q31.2	chr7:115842622-11	chr7:116102495-1	chr7:1-158	4.13E-24	9.48E-06	both	0: t<0.1; 1: 0.1	0	0	1	0	0	1	1
9	Amplificati	8q24.12	chr8:121983096-12	chr8:121997366-1	chr8:12198	0.048902	0.048902	broad	0: t<0.1; 1: 0.1	0	0	1	0	1	0	1
10	Deletion P	1p36.31	chr1:4257376-6053	chr1:5404535-605	chr1:1-240	8.69E-06	8.69E-06	focal	0: t>-0.1; 1: 0.1	0	0	0	0	0	0	0
11	Deletion P	4q34.3	chr4:183322597-18	chr4:183555243-1	chr4:18355	0.21835	0.21835	focal	0: t>-0.1; 1: 0.1	0	0	1	1	0	0	0
12	Deletion P	6q23.2	chr6:132978919-14	chr6:132978919-1	chr6:79415	0.000189	0.000189	broad	0: t>-0.1; 1: 0.1	2	0	0	0	0	0	0
13	Amplificati	1q32.1	chr1:201017471-20	chr1:201512199-2	chr1:20082	6.07E-08	6.07E-08	focal	Actual Log2 Ra	0.054818	0.042652	-0.29535	-0.00881	0	0.00089	0.005499
14	Amplificati	2p24.3	chr2:15719258-167	chr2:15830675-16	chr2:15830	0.23163	0.23163	focal	Actual Log2 Ra	-0.00296	0.12565	-0.01509	-0.00581	0.03833	-0.00072	0.006214
15	Amplificati	3q26.33	chr3:177090593-18	chr3:181261928-1	chr3:17709	0.043887	0.043887	focal	Actual Log2 Ra	-0.10861	-0.12928	0.17201	0.009299	0.013925	0.00367	-0.01703
16	Amplificati	4q12	chr4:54505358-552	chr4:54603039-55	chr4:48833	3.74E-14	3.74E-14	focal	Actual Log2 Ra	0	0.067307	0.45864	-0.01232	0.25929	-0.02441	0
17	Amplificati	6p21.1	chr6:42094850-432	chr6:42664817-43	chr6:42664	0.13151	0.13151	focal	Actual Log2 Ra	-0.03209	-0.01512	0.071373	-0.02192	0.025052	0.002768	0.002216
18	Amplificati	7p11.2	chr7:54640152-547	chr7:54709753-54	chr7:1-158	2.61E-79	2.61E-79	both	Actual Log2 Ra	0.03151	0.079714	0.22638	-0.02749	0.014743	2.4949	-0.02561
19	Amplificati	7q31.2	chr7:115842622-11	chr7:116102495-1	chr7:1-158	4.13E-24	9.48E-06	both	Actual Log2 Ra	0.03151	0.079714	0.22638	0.000868	0.014743	0.28996	0.38366
20	Amplificati	8q24.12	chr8:121983096-12	chr8:121997366-1	chr8:12198	0.048902	0.048902	broad	Actual Log2 Ra	0.010292	-0.07417	0.11934	0.033819	0.24449	0.014686	0.39299
21	Deletion P	1p36.31	chr1:4257376-6053	chr1:5404535-605	chr1:1-240	8.69E-06	8.69E-06	focal	Actual Log2 Ra	0.054818	-0.07143	0.1607	-0.07981	0	0.019109	0.005499
22	Deletion P	4q34.3	chr4:183322597-18	chr4:183555243-1	chr4:18355	0.21835	0.21835	focal	Actual Log2 Ra	0	0.073541	-0.26992	-0.50535	-0.06473	-0.02441	0
23	Deletion P	6q23.2	chr6:132978919-14	chr6:132978919-1	chr6:79415	0.000189	0.000189	broad	Actual Log2 Ra	-1.3294	0.056499	0.022594	-0.02325	-0.00897	-0.02443	-0.0073
24	Amplificati	7p	Amplitude values re	Broad Event Corr	chr7:1-158	2.61E-79	2.61E-79	both	0: t<0.1; 1: t>0	0	0	1	0	0	1	0
25	Amplificati	7q	Amplitude values re	Broad Event Corr	chr7:1-158	4.13E-24	9.48E-06	both	0: t<0.1; 1: t>0	0	0	1	0	0	1	0
26	Amplificati	8q	Amplitude values re	Broad Event Corr	chr8:12198	0.048902	0.048902	broad	0: t<0.1; 1: t>0	0	0	1	0	1	0	1
27	Deletion P	6q	Amplitude values re	Broad Event Corr	chr6:79415	0.000189	0.000189	broad	0: t>-0.1; 1: t<-	0	0	0	0	0	0	0
28																

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:

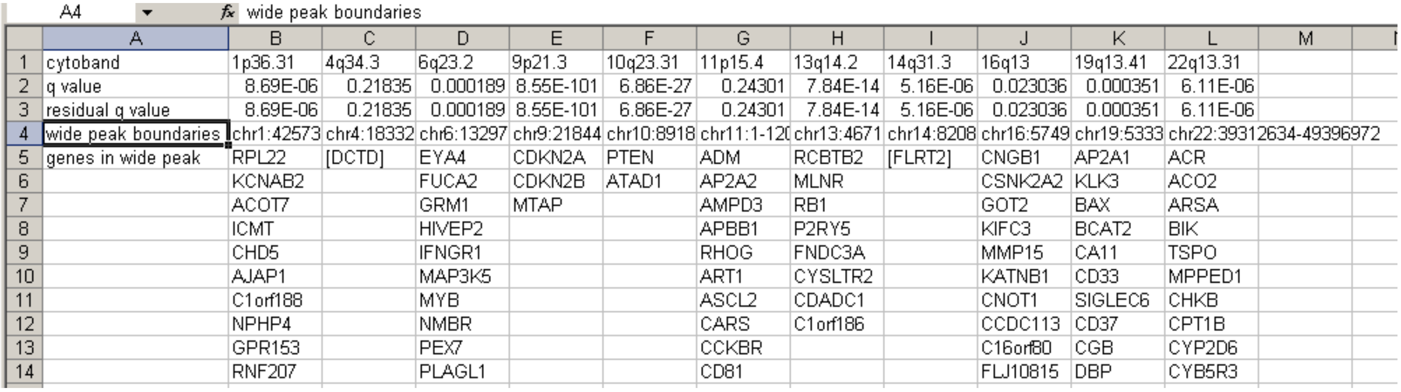
(1) *cytoband*
(2) *q-value*
(3) *residual q-value*
(4) *wide peak boundaries*

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.

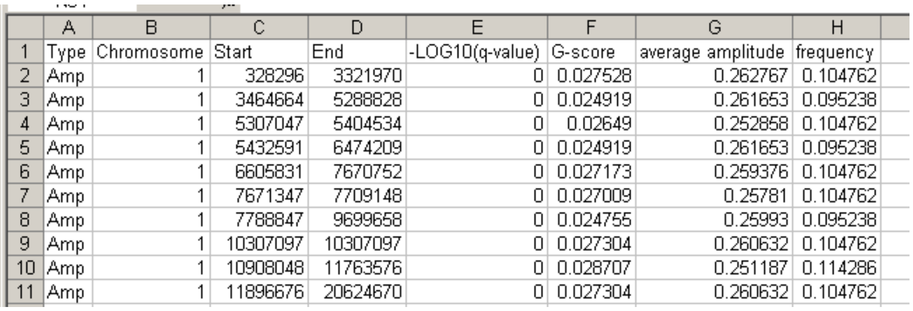


A4 wide peak boundaries

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	cytoband	1p36.31	4q34.3	6q23.2	9p21.3	10q23.31	11p15.4	13q14.2	14q31.3	16q13	19q13.41	22q13.31	
2	q value	8.69E-06	0.21835	0.000189	8.55E-101	6.86E-27	0.24301	7.84E-14	5.16E-06	0.023036	0.000351	6.11E-06	
3	residual q value	8.69E-06	0.21835	0.000189	8.55E-101	6.86E-27	0.24301	7.84E-14	5.16E-06	0.023036	0.000351	6.11E-06	
4	wide peak boundaries	chr1:42573	chr4:18332	chr6:13297	chr9:21844	chr10:8918	chr11:1-120	chr13:4671	chr14:8208	chr16:5749	chr19:5333	chr22:39312634-49396972	
5	genes in wide peak	RPL22	[DCTD]	EYA4	CDKN2A	PTEN	ADM	RCBTB2	[FLRT2]	CNGB1	AP2A1	ACR	
6		KCNAB2		FUCA2	CDKN2B	ATAD1	AP2A2	MLNR		CSNK2A2	KLK3	ACO2	
7		ACOT7		GRM1	MTAP		AMPD3	RB1		GOT2	BAX	ARSA	
8		ICMT		HIVEP2			APBB1	P2RY5		KIFC3	BCAT2	BIK	
9		CHD5		IFNGR1			RHOG	FNDC3A		MMP15	CA11	TSPO	
10		AJAP1		MAP3K5			ART1	CYSLTR2		KATNB1	CD33	MPPED1	
11		C1orf188		MYB			ASCL2	CDADC1		CNOT1	SIGLEC6	CHKB	
12		NPHP4		NMBR			CARS	C1orf186		CCDC113	CD37	CPT1B	
13		GPR153		PEX7			CCKBR			C16orf80	CGB	CYP2D6	
14		RNF207		PLAGL1			CD81			FLJ10815	DBP	CYB5R3	

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

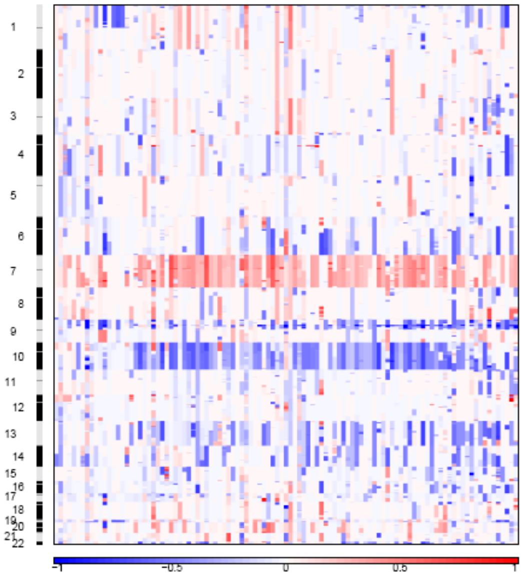
The scores file lists the q-values [presented as -log10(q)], G-scores, average amplitudes among aberrant samples, and frequency of aberration, across the genome for both amplifications and deletions. The scores file is viewable with the Integrative Genomics Viewer (IGV).



	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	3321970	0	0.027528	0.262767	0.104762
3	Amp	1	3464664	5288828	0	0.024919	0.261653	0.095238
4	Amp	1	5307047	5404534	0	0.02649	0.252858	0.104762
5	Amp	1	5432591	6474209	0	0.024919	0.261653	0.095238
6	Amp	1	6605831	7670752	0	0.027173	0.259376	0.104762
7	Amp	1	7671347	7709148	0	0.027009	0.25781	0.104762
8	Amp	1	7788847	9699658	0	0.024755	0.25993	0.095238
9	Amp	1	10307097	10307097	0	0.027304	0.260632	0.104762
10	Amp	1	10908048	11763576	0	0.028707	0.251187	0.114286
11	Amp	1	11896676	20624670	0	0.027304	0.260632	0.104762

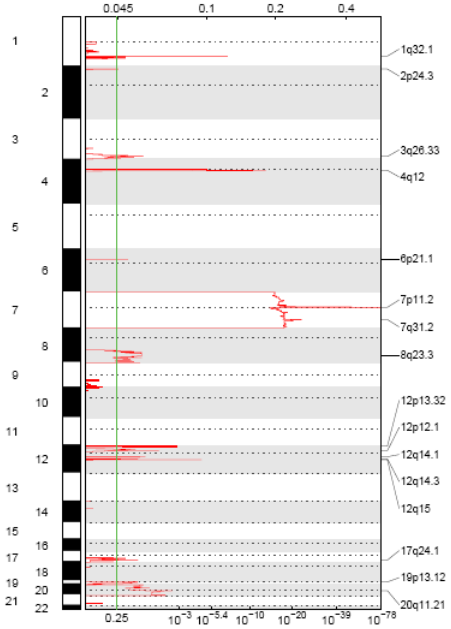
5. Raw Copy Number (Raw_copy_number.xx.pdf)

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



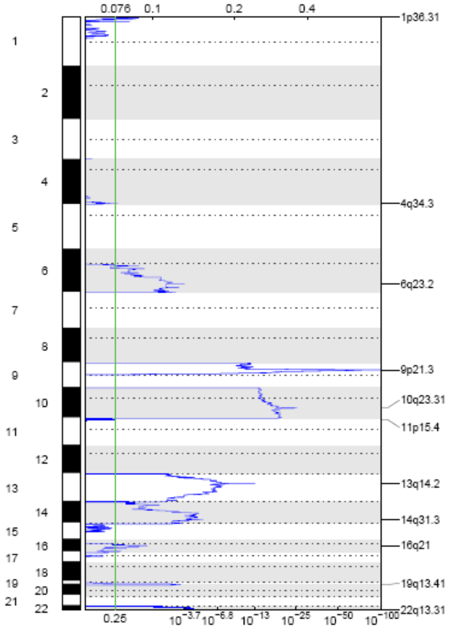
7. Amplification GISTIC plot (amplification.xx.pdf)

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.



8. Deletion GISTIC plot (deletion.xx.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.



Platform Dependencies

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