CNApp

Copy Number Alterations integrative analysis

User's Guide

1. Input data formats and supported platforms

CNApp supports segmented genomic data from any segmentation algorithm, and works for both Illumina and Affymetrix SNP arrays, as well as for next generation sequencing data. Human genome builds hg19 and hg38 are accepted.

Minimum input file (MIF) should be tab- or comma-separated and **must** include sample name (*ID*), chromosome (*chr*), start (*loc.start*) and end (*loc.end*) positions, and the log2 ratio of the copy-number amplitude (*seg.mean*) for each segment. when available, the input file will also include sample purity estimations (*purity*) and BAF values (*BAF*), which correct the accuracy of CNA calls and provide copy number neutral loss-of-heterozygosity (CN-LOH) events. **File HEADER and COLUMNs ORDER are mandatory**.

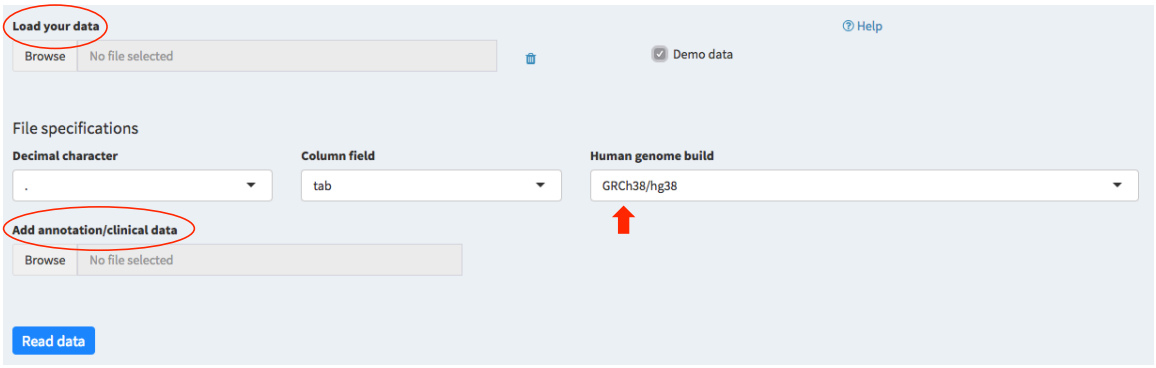
Example of MIF:

```
ID  chr*    loc.start   loc.end seg.mean
TCGA-2Y-A9GT    chr2    101196995   102933932   0.3482
TCGA-2Y-A9GT    chr3    57302395    59093641    0.362
TCGA-2Y-A9GT    chr5    23361094    23361904    -1.4161
TCGA-BC-A112    chr5    8752690 8833772 0.3972
TCGA-BC-A112    chr7    7180530 7180603 1.2037

(*) chromosome information can be loaded either with or without 'chr' prefix
```

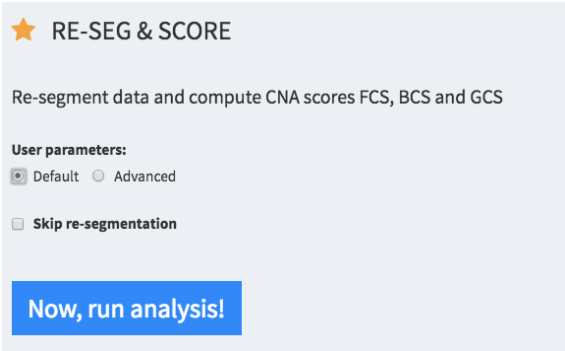
Annotation/clinical data can be included in the MIF as additional columns (tagged in every segment from each sample), or described in another tab- or comma-separated file by specifying new variables to every sample. **File HEADER is mandatory and the first column should be the sample name (*ID*) with the same code used in the segmented file**.

Input files with the segmented data and the annotation/clinical data –if available- can be directly uploaded from your computer. *Human genome build* should be specified here.



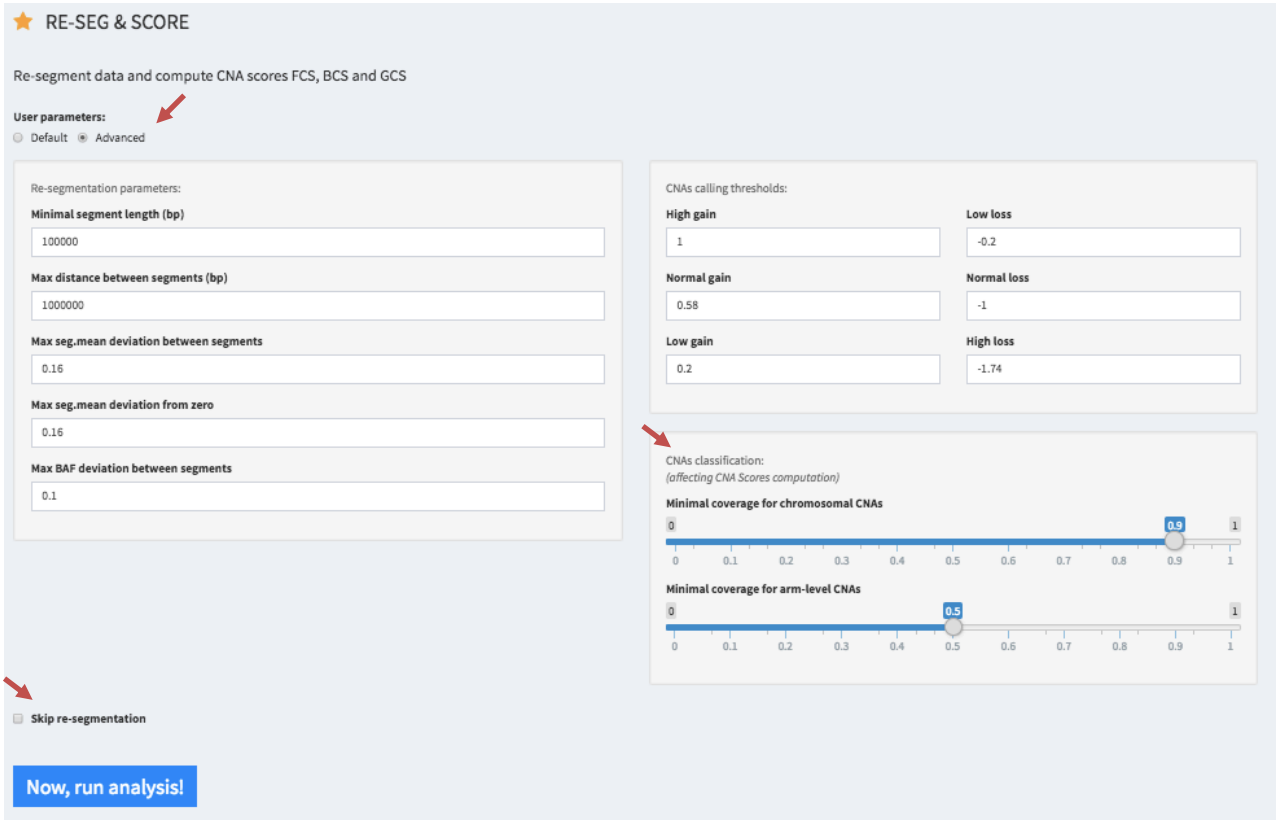
2. Re-Seg & Scores

First, CNApp applies a re-segmentation approach aiming at correcting potential background noise and amplitude divergence due to technical variability, as well as to re-adjust copy number thresholds using the estimated purity (if provided). Users can customize the parameters (*Advanced* option) or use the default ones.



Re-segmentation step can be skipped. If so, the scores are computed considering segments from the input file and default thresholds for classifying broad and focal events (see below).

Alternatively, the user can just customize broad and focal thresholds and skip the re-segmentation by modifying the *CNAs classification* parameter in *Advanced* option and then clicking *Skip re-segmentation*.



- ***Re-segmentation parameters*** **[default values]**:

- *minimum segment length* [100000 bp]

  Minimum length of a copy number segment to be included in the analysis. Segments below this value will be automatically discarded.

- *maximum distance between segments* [1000000 bp]

  Maximum distance allowed between segments to be merged. Segments above this value will not be merged even accomplishing the remaining required parameters.

- *maximum amplitude (seg.mean) deviation between segments* [0.16]

  Maximum difference allowed in *seg.mean* values between adjacent segments.

- *maximum amplitude (seg.mean) deviation from segment to zero* [0.16]

  Maximum difference allowed in *seg.mean* values different from 0.

- *maximum BAF deviation between segments* [0.1]

  Maximum deviation allowed in BAF values between adjacent segments. If BAF is not provided, this parameter is not considered.

- ***CNAs calling thresholds*** **(default values)**:

CNA level	Log2ratio values	Number of copies
High-level gain	1	≥ 4 copies
Medium-level gain	0.58	[3 – 4) copies
Low-level gain	0.2	[2.3 – 3] copies
Low-level loss	-0.2	(1 – 1.7] copies
Medium-level loss	-1	(0.6 – 1] copies
High-level loss	-1.74	≤ 0.6 copies

If the user wishes to change default values, the relation between the *seg.mean* and the estimated number of copies can be calculated as:

```
seg.mean = log2(number of copies/2)
```

- ***CNAs classification, affecting CNA scores computation*** **[default values]**:

- *minimal coverage for chromosomal CNAs* [0.9]

  considering the relative length of a segment to the whole-chromosome, the segment will be tagged as chromosomal if covers ≥[value] of the chromosome

- *minimal coverage for arm-level events* [0.5]

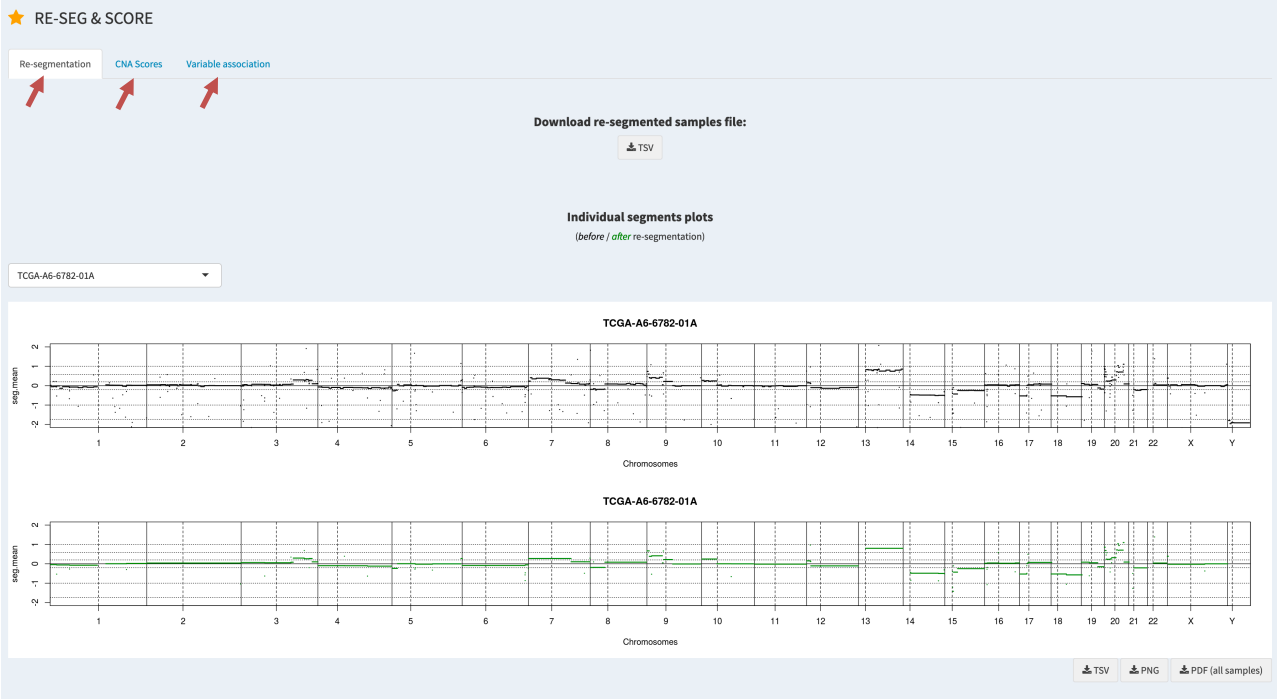
considering the relative length of a segment to a chromosome-arm, the segment will be tagged as arm-level if covers ≥[value] of the arm

Results from this step are presented in three different tabs:

2a. Re-segmentation:

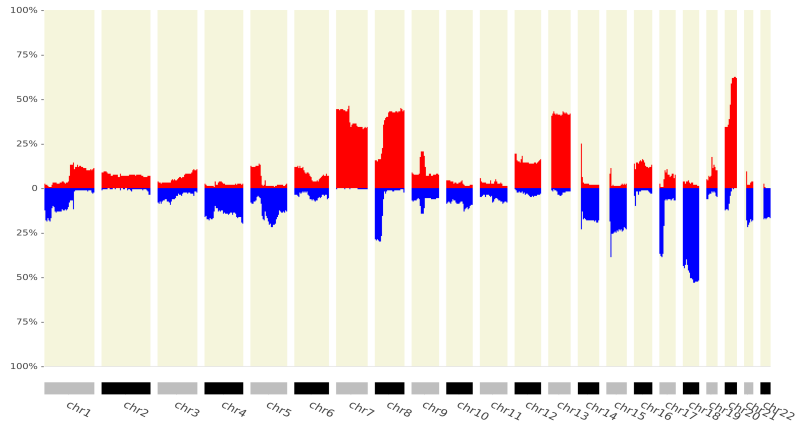
Re-segmented data of all samples can be downloaded in a TSV file. This file contains the new segmented data and also includes those segments that have not been classified by CNApp neither as gain, loss or CN-LOH. This unclassified status is consequence of presenting *seg.mean* and BAF values lower than the defined thresholds. Unclassified segments are just informative and are NOT considered for computing the CNA scores.

Individual segmented plots showing copy number segments before and after the re-segmentation are displayed and can be downloaded for each sample (PNG) or for all samples (PDF).

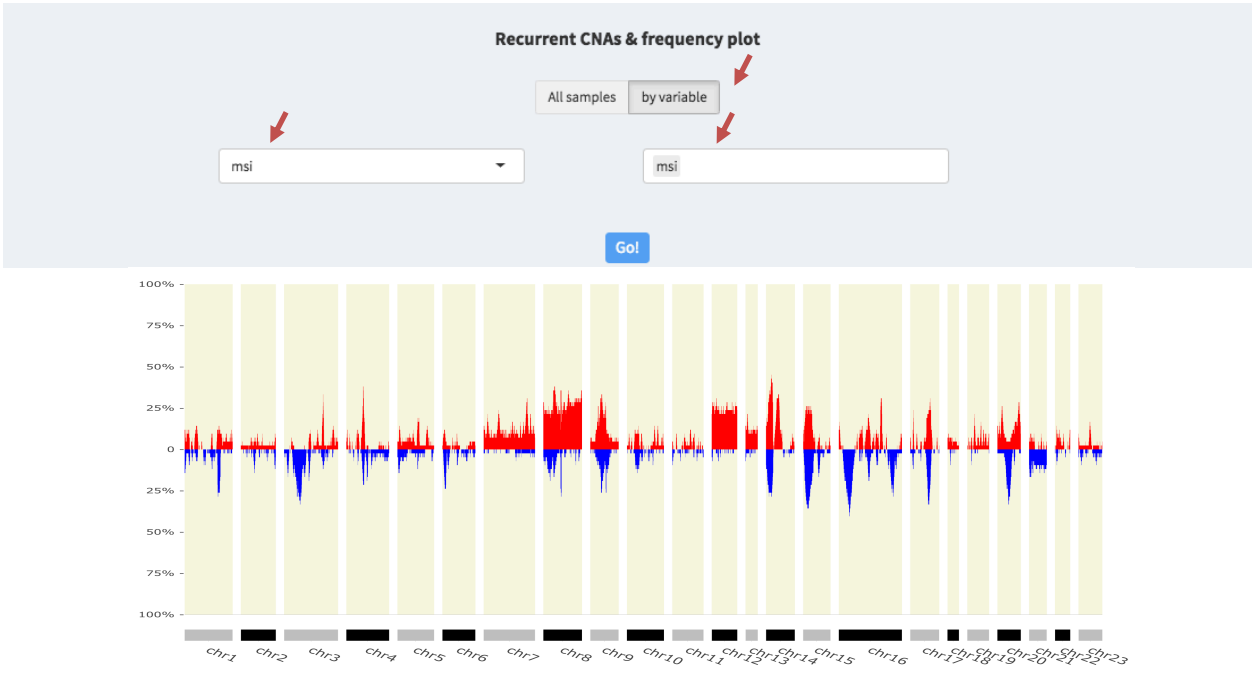


Example of re-segmentation for an individual sample

Genome-wide Frequency plots showing the overall frequencies of the CNAs in the cohort (*all samples*) or in a subset of samples defined by any annotated variable provided (*by variable*) can be also displayed and downloaded.



Genome-wide frequency plot all samples (demo data)



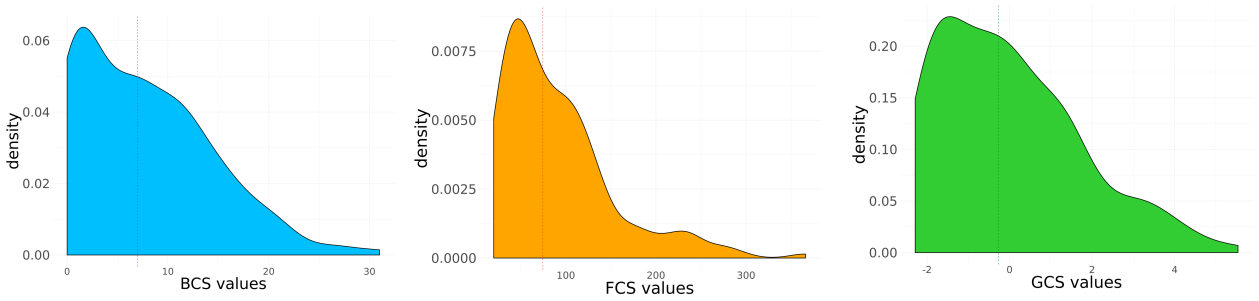
Genome-wide frequency plot samples with microsatellite instability (MSI) (demo data)

A table (TSV) including the percentages of gains and losses of the different genomic ranges can also be obtained. If the dataset includes ≤ 150 samples, minimum shared regions and frequencies are calculated

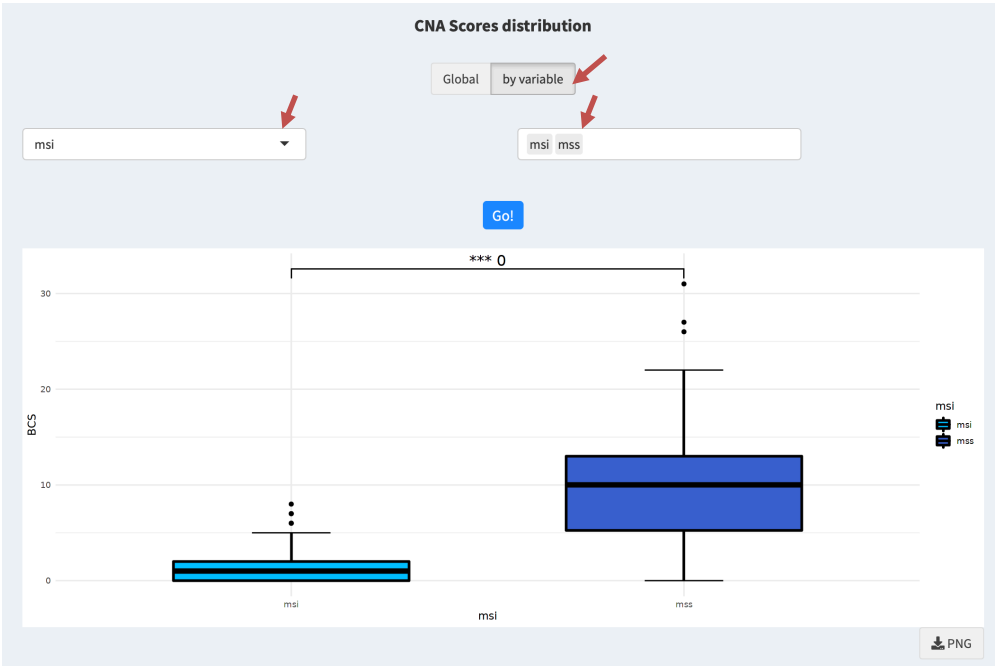
with the GenomicRanges R package; if the dataset comprises > 150 samples, genomic windows of 5-Mb are generated by default and the frequencies of *gains* and *losses* from the selected samples are plotted.

2b. CNA scores:

Plots showing the CNA scores distribution of each score (BCS, FCS and GCS) can be displayed and downloaded (*Global*). Additionally, it is possible to compare the distribution of the scores between different groups, which can be selected from the annotated variables provided (*by variable*). In such case, a **Student's t-test** is performed and *p-values* are displayed in the plots.

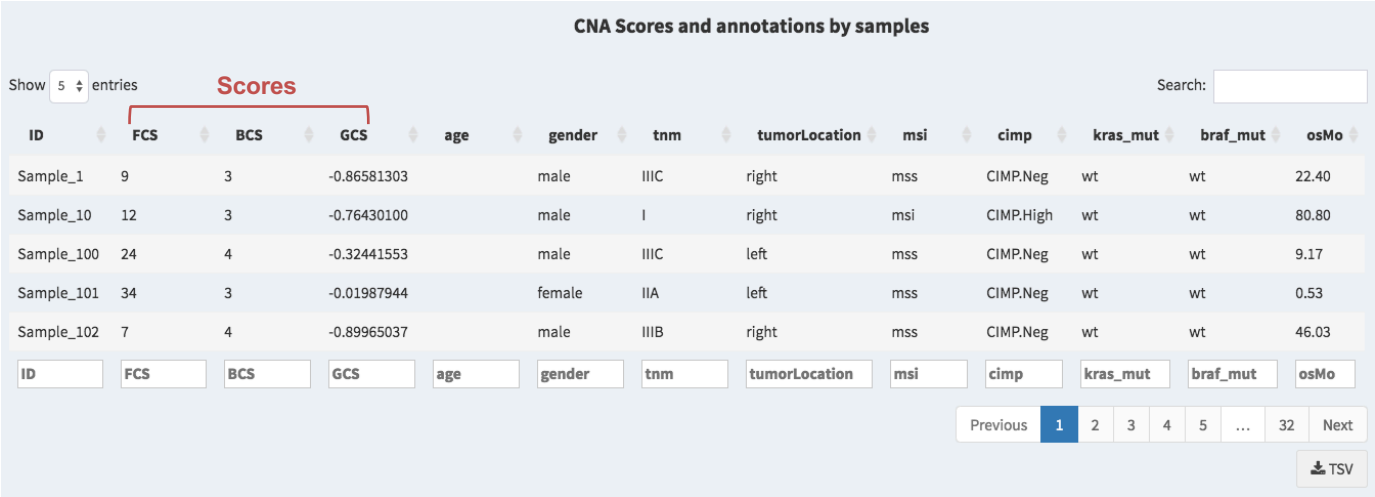


Global distribution of the CNA scores in demo data (BCS, FCS and GCS)



Comparison of the distribution of the BCS between MSI and MSS samples (demo data)

At the same time, a TSV file with the CNA Scores calculated for each sample can be downloaded in a data frame also containing the information provided in the annotated/clinical data.



2c. Variable association:

Annotated variables from input file are statistically associated with CNA Scores computed by CNApp in Re-Seg&Scores part. Different association tests are applied according to variable class (i.e. categoric or numerical), as described in the following table. Both parametric and non-parametric tests are computed to present *p-values*, in order to assess statistical significance for each association.

		Parametric	*Non-parametric*
Categoric	*n* = 2	Student's T-test	Wilcoxon signed-rank test
	n > 2	ANOVA	ANOVA: Kruskal test
Numerical		Pearson correlation	Spearman correlation

n = groups defined by annotation variable

Example of the results generated from non-parametric tests using the demo data. *P-values* showing the association between the annotated variables (rows) and scores (columns) are shown.

*ns = no significant; ** significant; *** highly significant*

Non-parametric tests *(p-values)*

Show 25 entries

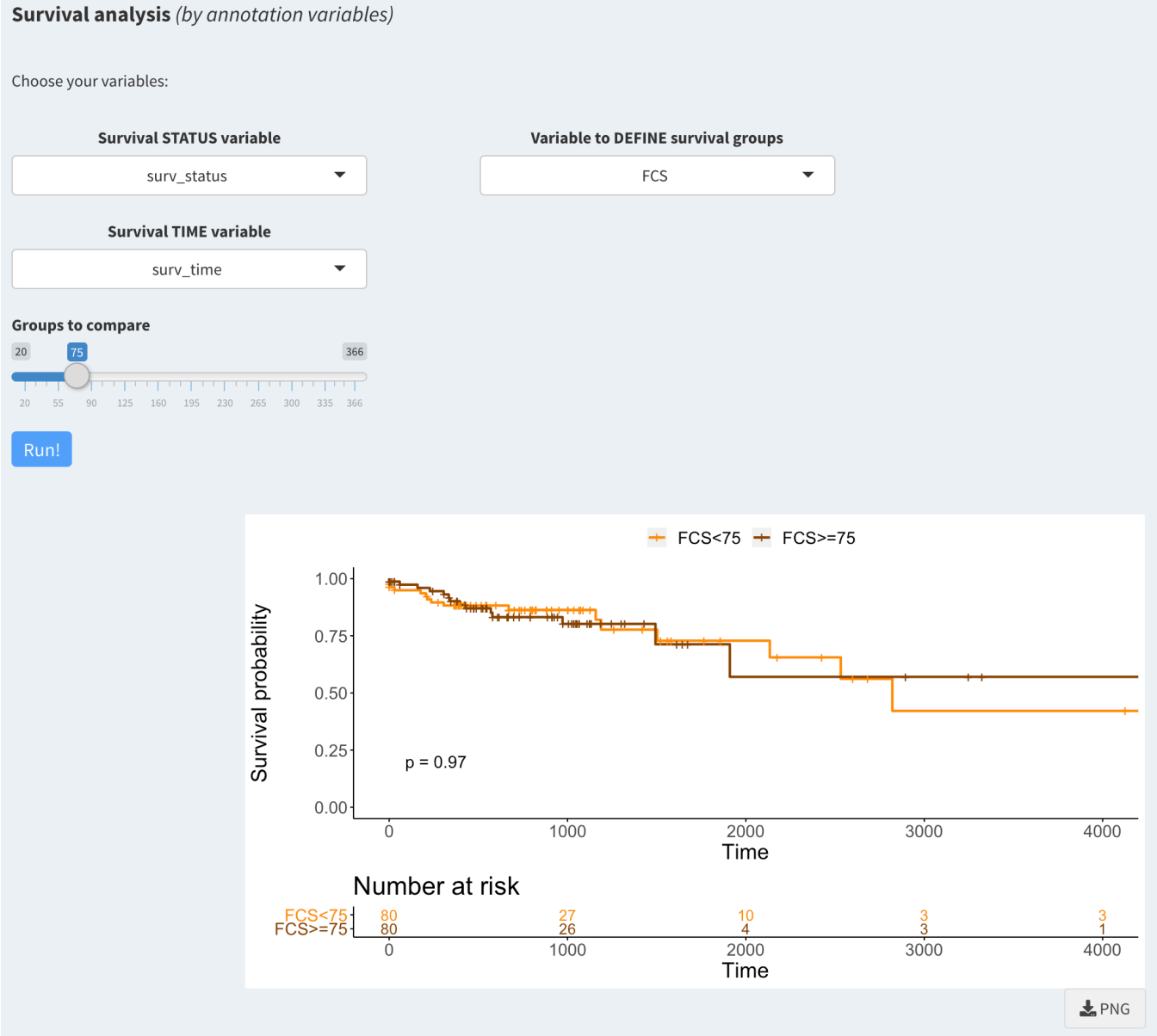
Variable	FCS	BCS	GCS
age			
gender	0.26779 (ns)	0.00093 ***	0.15253 (ns)
tnm	0.15514 (ns)	0.3833 (ns)	0.17468 (ns)
tumorLocation	0.00504 **	0.0359 *	0.00264 **
msi	0.00208 **	0.00357 **	0.00113 **
cimp	0.20031 (ns)	0.14651 (ns)	0.18549 (ns)
kras_mut	0.82898 (ns)	0.84692 (ns)	0.86367 (ns)
braf_mut	0.12453 (ns)	0.2694 (ns)	0.17697 (ns)
osMo	0.51252 (ns)	0.81601 (ns)	0.61806 (ns)

Previous 1 Next

TSV

2d. Survival analysis:

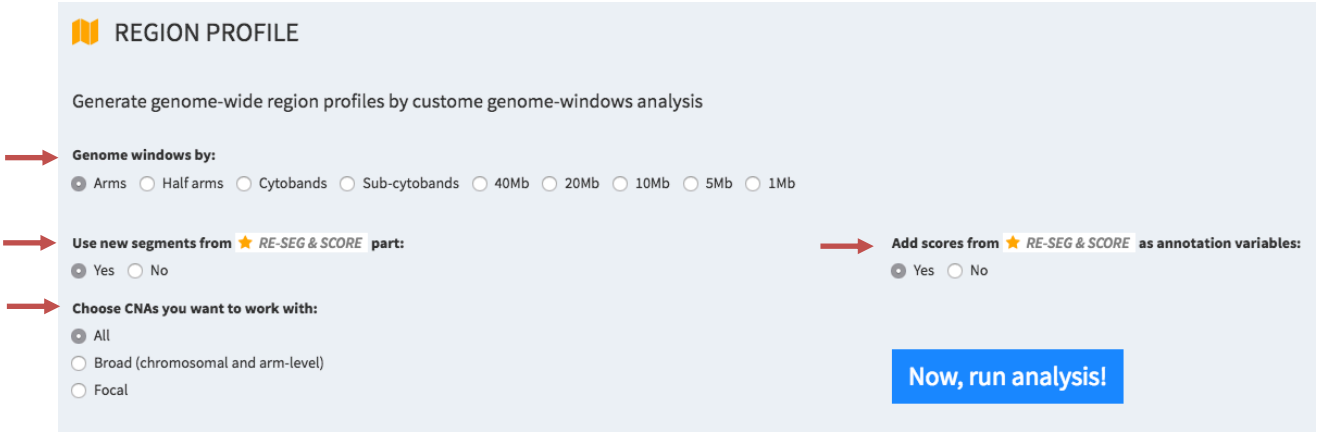
By uploading survival annotation into the dataset CNApp can compute Kaplan-Meyer curves plots. Suvival STATUS and TIME column annotations are necessary in order to complete this analysis. A third annotation variable to define sample groups can be selected by the user (e.g. CNA scores values or any other annotation column provided by the user). The following screen shot shows survival plot analysis with samples groups defined according to the threshold FCS=75 (survival plot can be downloaded as PNG image):



3. Region Profile

- ***CN PROFILES***

The genomic region profiling generates genome-wide profiles using re-segmented data obtained from the first step or uploaded segmented data without re-segmentation to allow sample-to-sample correlations. To conduct this, re-segmented CNA profiles are transformed into genome region profiles defined by a user-selected window (i.e., chromosome arms, half-arms, cytobands, sub-cytobands or 40-1 Mb windows). Length-relative means are computed for each window by considering amplitude values from those segments included in the specific window.

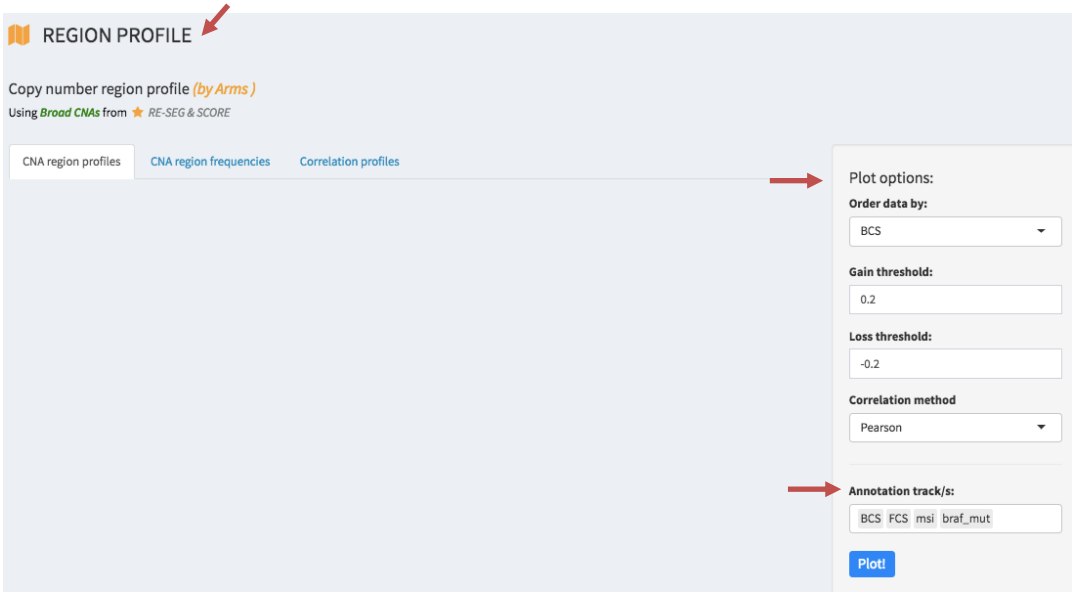


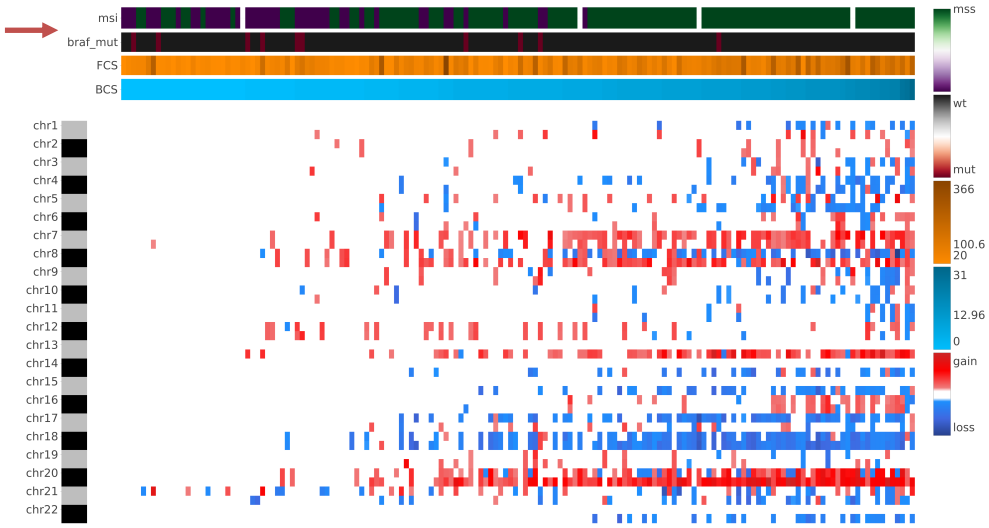
By using the classification of segments in broad or focal from *Re-Seg & Score*, all re-segmented data, focal only or broad only segments can be selected for this analysis. Similarly, the scores computed in the previous step can be included as annotated variables.

Results from are also presented in three different tabs:

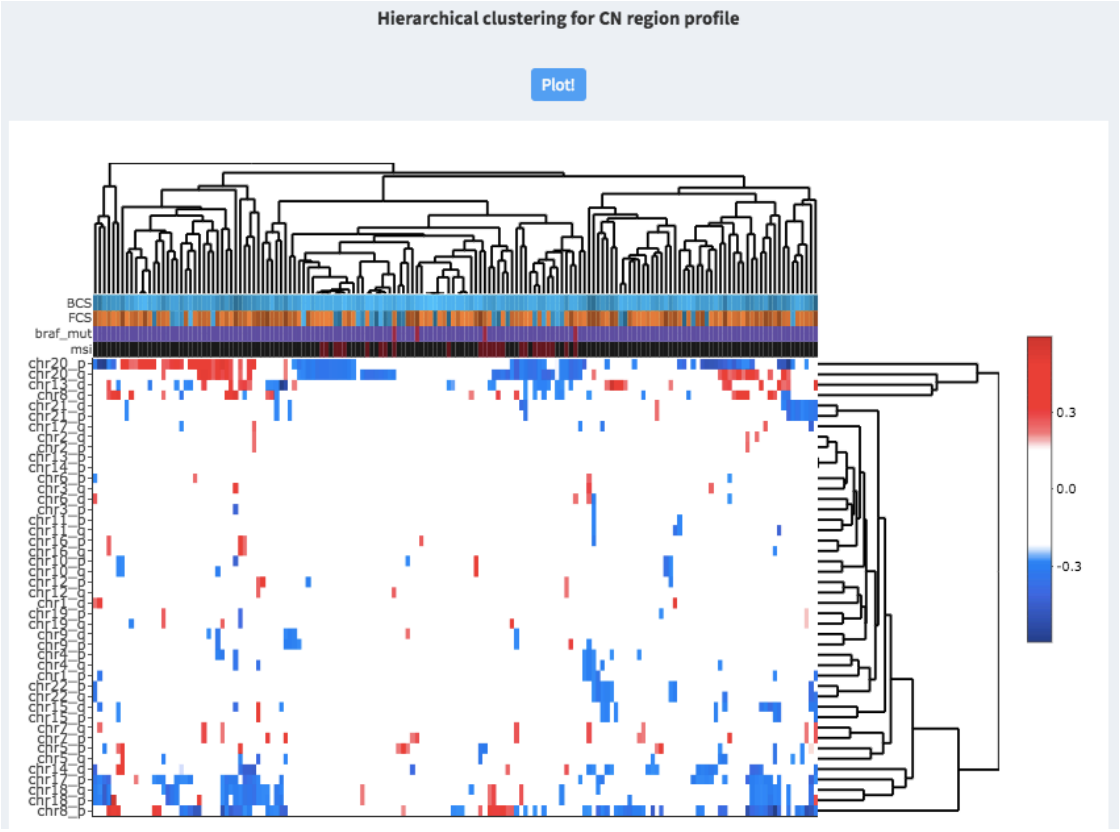
3a. CNA region profiles:

To visualize the selected copy number genome profiles a Heatmap is generated. By default, the established low-copy gains and losses thresholds (i.e., |0.2|) are used as cutoffs to classify genome regions and to calculate their frequencies. However, if the user wishes to display alterations of higher amplitude, these thresholds can be modified in the right panel. Samples can be ordered by any of the annotated variables provided and up to 6 annotation tracks (from these annotated variables) can be added to the heatmap and plotted simultaneously allowing for visual comparison and correlation between CNA profiles and different variables. A table (TSV) showing the average of *seg.mean* for selected genomic windows and for each sample can be also downloaded.





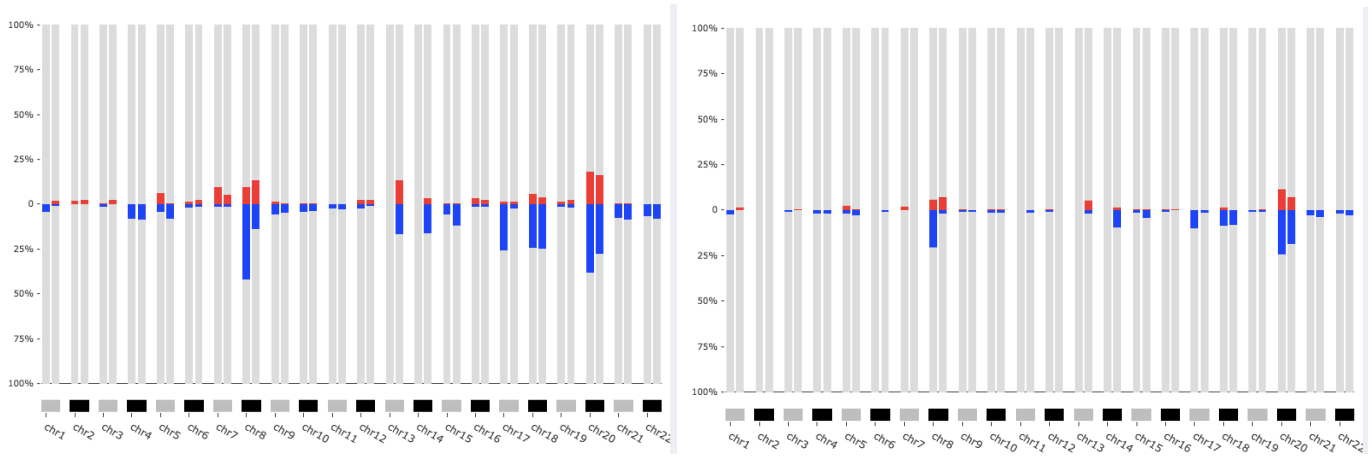
Optionally, the user can generate and download an unsupervised hierarchical clustering for the selected CN region profile.



Finally, the list of genes and their coordinates included in the each selected genomic window can be downloaded as a TSV file.

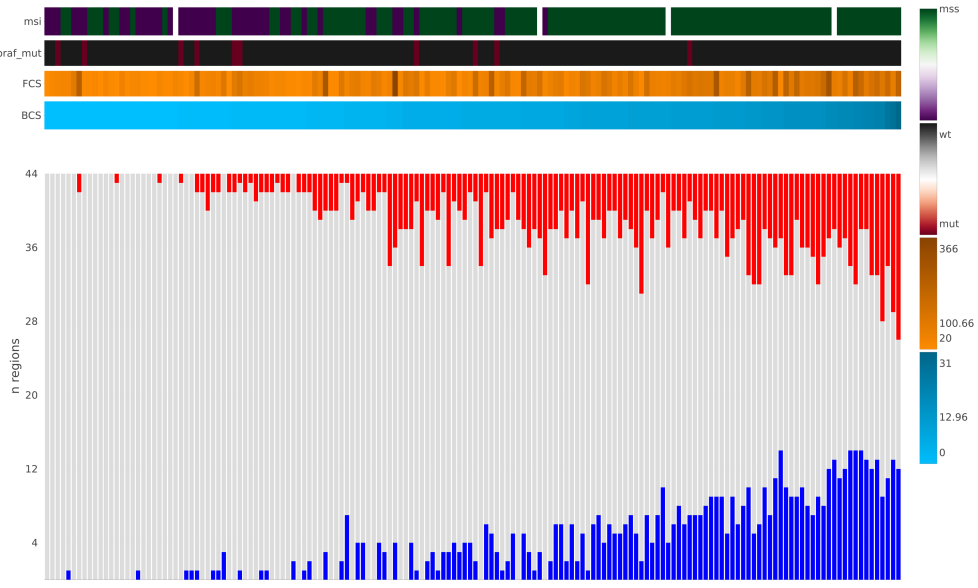
3b. CNA region frequencies:

A genome-wide frequency plot of the selected genomic-windows is displayed and can be downloaded. This information can be also obtained in a table (TSV) showing the % of samples with *gains, losses,* and *normal* status (no alteration). CNA cutoff can be also modified by the user.



Genome-wide frequency plot with default CNA thresholds vs frequency plot with CNA thresholds set at 0.3 (demo data)

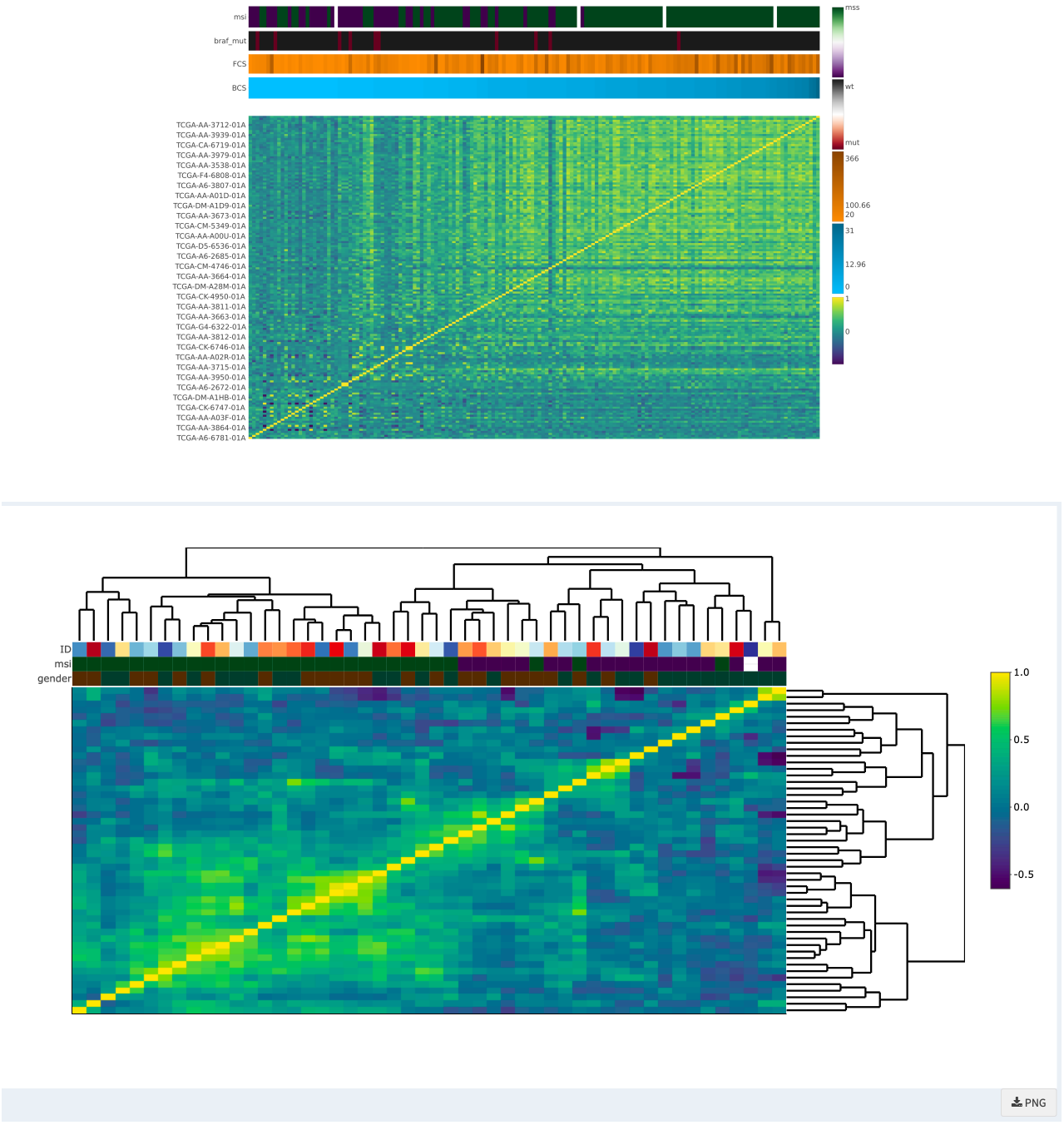
A stacked bar plot showing the counts of gains and losses for each sample and genomic region is also displayed and can be downloaded as a plot or table (TSV).



In this case, samples can be ordered (right panel) by any of the annotated variables provided by the user and up to 6 annotation tracks can be added and plotted simultaneously.

3c. Correlation profiles:

Sample-to-sample correlation plot are generated upon choosing the variable of interest. A hierarchical clustering for samples correlation is optional.



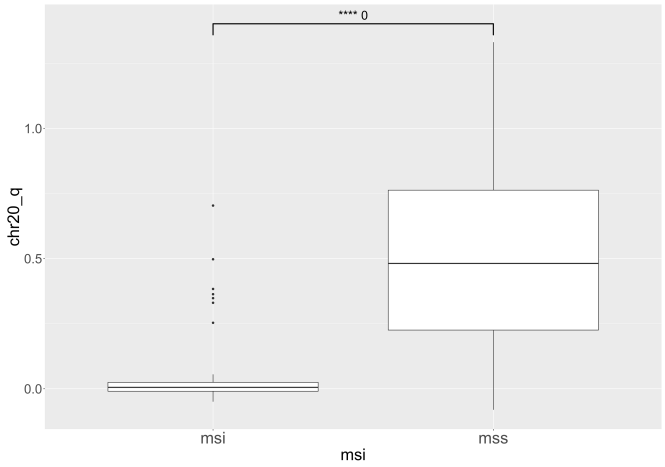
Samples correlation heatmap and hierarchical clustering by clinical variable (MSI status)

- ***DESCRIPTIVE REGIONS***

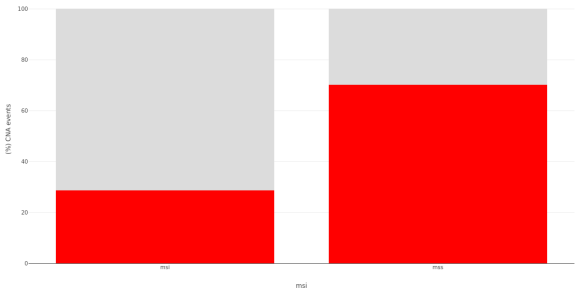
Assessing differentially altered regions between sample groups might unveil the biological significance of specific genomic events. CNApp interrogates descriptive regions associated with any sample-specific

annotation variable provided in the input file. **Student's t-test** or **Fisher's test** are applied when considering CNAs as continuous alterations (*seg.mean* amplitude values) or as categorical events (presence of gains and losses), respectively. The user can select the groups to compare in the right panel. Default statistical significance is set to adjusted p-value lower than 0.1 but it can be manually modified.

A plot showing regions with significant differences between the selected groups is displayed and can be downloaded (also as a table TSV). To visualize the sense of the difference, boxplots are displayed for the region of interest.



Boxplot showing chromosome arm 20q region values across MSI and MSS sample groups.



Barplot showing chromosome arm 20q region presencde across MSI and MSS sample groups.

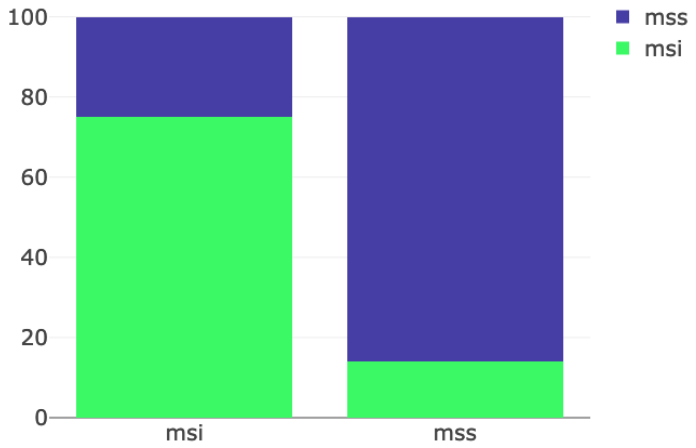
The list of genes and their coordinates included in the region can be downloaded as a TSV file.

4. Classifier Model

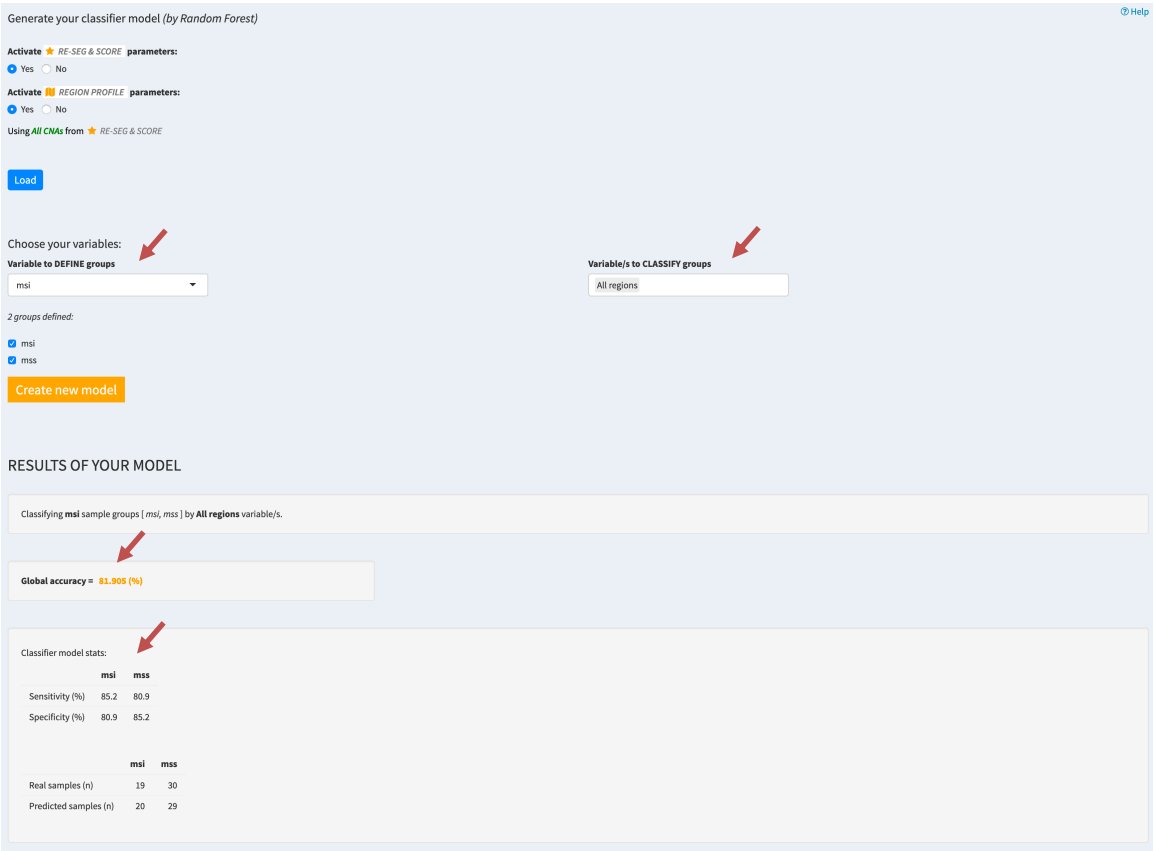
Machine learning-based classifier models are used by choosing a variable to define sample groups and one or multiple classifier variables. Annotation variables from the input file are loaded. If *Re-Seg & Score*

and/or *Region profile* sections have been previously completed, the user can upload data from these sections.

Predictions for performance are generated and global accuracies are computed along with sensitivity and specificity values by sample group. A summary of the data distribution and plots for real and model-predicted groups are visualized.



Predicted-tagged samples between real clinical sample groups (MSI status)



Results of the classifier model indicating sensitivity and specificity