SnoopCGH express manual

Introduction

SnoopCGH is a java desktop application for visualising and exploring comparative genomic hybridization (CGH) data. The software allows the user to interactively analyse several sets of data simultaneously. The input is based on a tab-, space- or comma-delimited format, containing series of logarithmic intensity values corresponding to one or more comparisons or samples.

SnoopCGH provides CGH plots with unlimited zoom (in both axes) that can be explored interactively with the mouse. The use of multiple layers, that can be stacked and combined, facilitates the visualization of the data. It is possible to apply several layers to a plot in order to filter the CGH ratios or perform statistical analysis in regions of interest. Analysis methods have been implemented and enable the rapid visualisation and dissection of putative structural variants (SVs).

In particular, data are smoothed using an algorithm based on Haar wavelets [1], and islands of potential SVs are estimated using SW-Array [2]. Other powerful feature of SnoopCGH is its ability to interface with downloadable annotation files (e.g. embl) from genomic browsers, that include information on gene names and genomic features (e.g. GC content). The user has a visual representation of the annotations at the foot of the plot and can easily access detailed textual information. Direct links to the main genomic browsers are incorporated.

***Note**: This is a express manual, for a detailed introduction to SnoopCGH check the **Visual Introduction** pdf file.*

Running SnoopCGH

The software has been developed in Java so it should run in any platform supported by the Java Virtual Machine without problems. SnoopCGH is distributed as a JAR executable file. If you have the java JDK properly installed (v1.6 preferred) , just type from a console: *java –jar SnoopCGH.jar.*

Special Memory Requirements

If you are working with big datasets, sometimes it is necessary to start SnoopCGH specifying the required amount of memory (if not you will obtain a memory error if you try to open a file). You can specify the initial amount of memory assigned to SnoopCGH with the –Xms option. The maximum amount SnoopCGH could get is indicated with the –Xmx option.

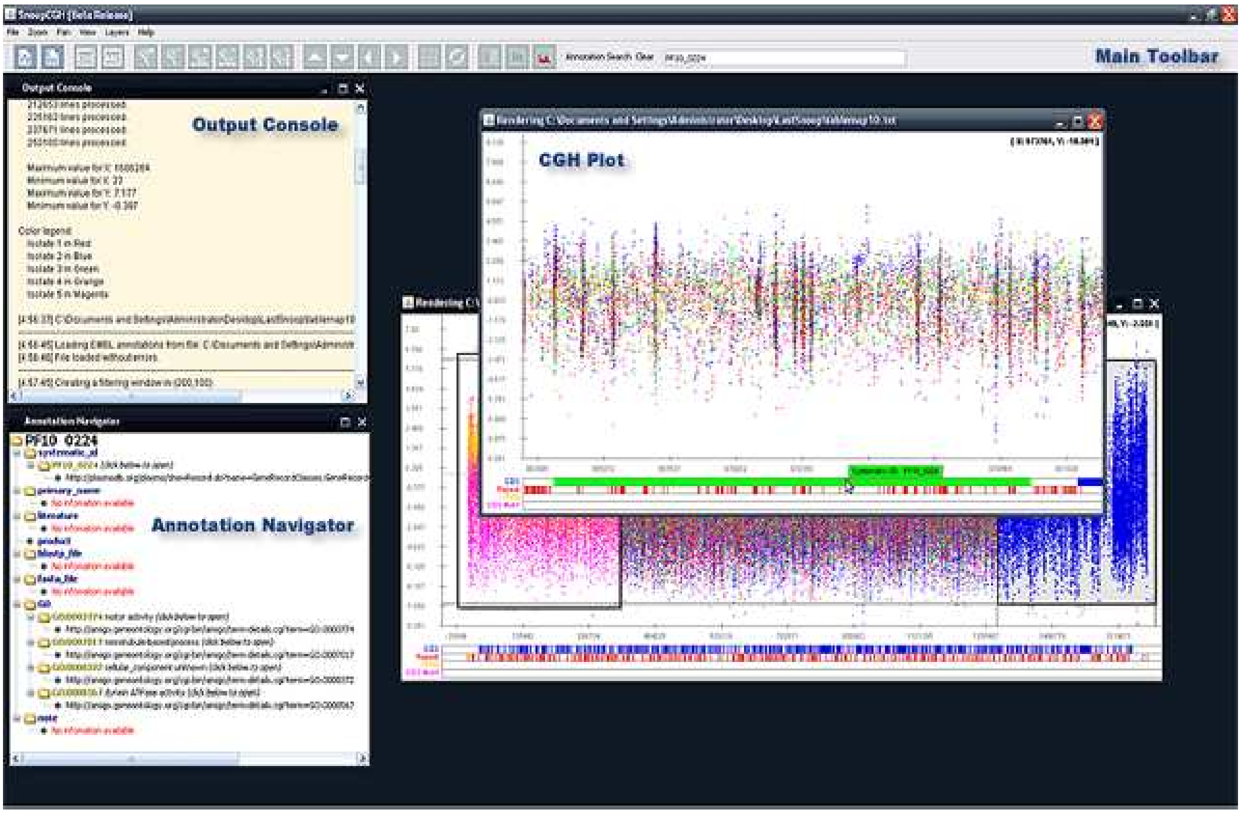
For instance: *java -jar -Xms128m -Xmx256m Snoop.jar* will set the initial amount of memory to 128MB and the maximum to 256MB (notice the m after the numbers).

Launchers

You can use the .bat (windows) or .sh (linux) launchers to run SnoopCGH with the default amount of memory (128/256 MB).

Main Components

The following capture shows the main components of SnoopCGH.

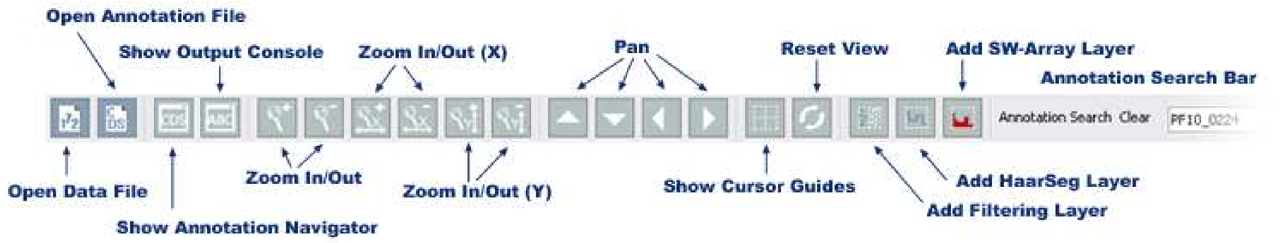

SnoopCGH (Beta Release)
File Zoom Pan View Layers Help
Annotation Search Clear PF10_0224
Main Toolbar
Output Console
Annotation Navigator
CGH Plot
Rendering C:\Documents and Settings\Administrator\Desktop\LastSnoop\tablemap10.txt
Rendering C:\

```
212653 lines processed
225162 lines processed
237671 lines processed
250180 lines processed

Maximum value for X: 1508284
Minimum value for X: 23
Maximum value for Y: 7.137
Minimum value for Y: -0.367

Color legend:
  Isolate 1 in Red
  Isolate 2 in Blue
  Isolate 3 in Green
  Isolate 4 in Orange
  Isolate 5 in Magenta

[4:56:37] C:\Documents and Settings\Administrator\Desktop\LastSnoop\tablemap10
[4:56:45] Loading EMBL annotations from file: C:\Documents and Settings\Administr
[4:56:46] File loaded without errors
[4:57:46] Creating a filtering window in (200,100)
```

```
PF10_0224
  systematic_id
    PF10_0224 (click below to open)
      http://plasmodb.org/plasmo/showRecord.do?name=GeneRecordClasses.GeneRecord
  primary_name
    No information available
  literature
    No information available
  product
  blastp_file
    No information available
  fasta_file
    No information available
  GO
    GO:0003774 motor activity (click below to open)
      http://amigo.geneontology.org/cgi-bin/amigo/term-details.cgi?term=GO:0003774
    GO:0007017 microtubule-based process (click below to open)
      http://amigo.geneontology.org/cgi-bin/amigo/term-details.cgi?term=GO:0007017
    GO:0008372 cellular_component unknown (click below to open)
      http://amigo.geneontology.org/cgi-bin/amigo/term-details.cgi?term=GO:0008372
    GO:0008567 dynein ATPase activity (click below to open)
      http://amigo.geneontology.org/cgi-bin/amigo/term-details.cgi?term=GO:0008567
  note
    No information available
```

- The **main toolbar** gives access to all the software features.
- The **output console** shows information related with the last operations performed (progress, analysis numeric values, errors, etc.).
- The **annotation navigator** window shows the details of the last annotation feature selected.
- A **CGH plot** is a window that allows the user to explore the data. Each opened data file is presented in a different plot.

Main Toolbar



Open Data File: Opens a data file in a new window. The file should be a plain text file with tabs, blank spaces or commas as separators. Each line has to contain the values for each probe, being the columns:

Probe_id chromosome_number start_position end_position ratio_value_1 ratio_value_2 ... ratio_value_n

Note: A description of the ongoing data loading process will be shown in the output console window (if the output is active). The messages in this window should be checked if any problem appears when loading a file.

Open Annotation File: Opens and associates an EMBL annotation data file with the current active plot. After loading an annotation file, the annotations will be shown below the CGH plot.

Note: Only one annotation file can be associated with each plot, however it is possible to open the same data file twice (in two different windows) and then associate each window with a different embl file.

Show Output Console: Shows the output console (brings it to the front).

Show Annotation Navigator: Shows the annotation detail windows (brings it to the front).

Zoom: The zoom operations can be performed using the mouse wheel in combination with some keys:

- ***Zoom In*** *(mouse wheel up or Ctrl + num pad +)*
- ***Zoom Out*** *(mouse wheel down or Ctrl + num pad -)*
- ***Zoom In X*** *(mouse wheel up + Ctrl)*
- ***Zoom Out X*** *(mouse wheel down + Ctrl)*
- ***Zoom In Y*** *(mouse wheel up + Shift)*
- ***Zoom Out Y*** *(mouse wheel down + Shift)*

Pan: These buttons control the pan operations over the plot (up, down, left and right). However, the plot can be moved just dragging it with the mouse in any direction.

Show Cursor Guides: Shows a pair of guides from the cursor to the axes.

Reset View: Resets the zoom and pan levels of the selected plot window.

Add Filtering Layer: Adds a new filtering layer. The layer will show only the selected set of isolates.

Add Segmentation Layer: Adds a segmentation layer based on the HaarSeg segmentation algorithm.

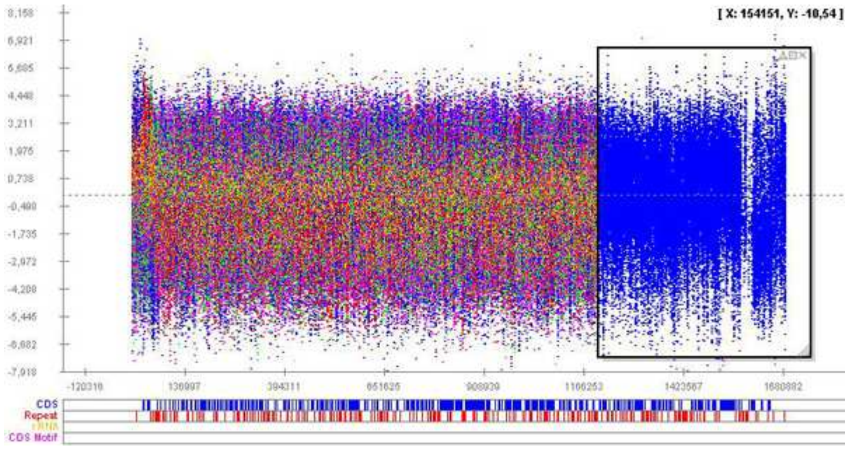
Add SW-Array layer: Adds an analysis layer based on the SW-Array algorithm.

Annotation Search: It is possible to search using the annotations of the active plot (using systematic ids, primary names, literature terms, etc.). After any match the annotation feature will be centred in the window. To move to the next match, just click again the search button.

Plot Window

The plot window represents each isolate with a different colour (check the output console after loading a data file for the correspondence). The plot can be moved with the mouse in any direction. While the mouse wheel controls the zoom level. The X-axis represents genomic locations while the Y-axis represents the log ratio-values. The legend in the top right corner indicates the exact point of the mouse cursor.

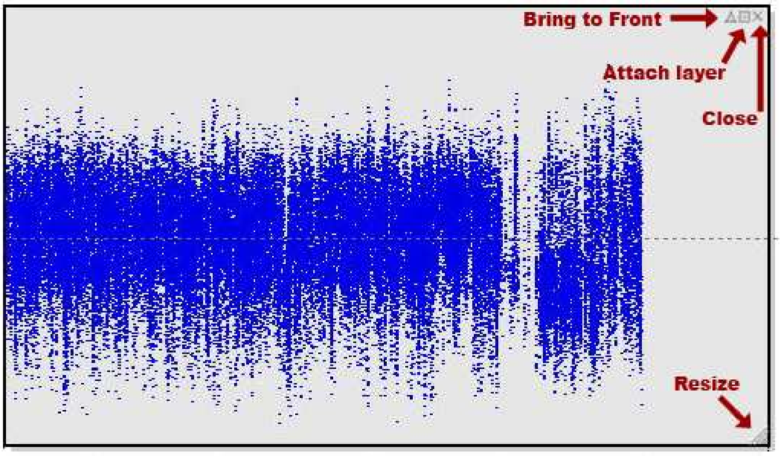
Note: The zoom behaviour can be modified by pressing shift/control while moving the mouse wheel. Pressing shift will affect only the X-axis, pressing control will affect only the Y-axis.



When an annotation file has been loaded, the cursor can be placed over any feature bar to obtain the related systematic id. Clicking on it will update its details in the annotation navigator window and the double click will centre the featured region in the plot window.

Layers

It is possible to add layers to a plot in order to filter the data or perform some analysis. To move a layer just drag it with the mouse (the layer will become active when the mouse is placed over it).



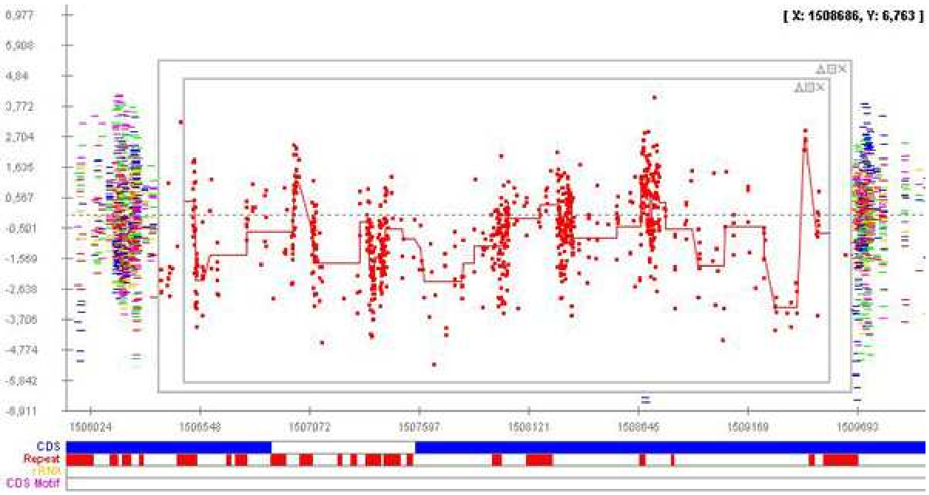
Basic controls in a layer:

- **Close** (x): Remove the layer.
- **Bring to front** (triangle): Moves the layer to the top.
- **Attach** the layer to the background (square): it is not possible to move the layer until this button is clicked again.
- **Resizing** corner: Resizes the layer.

HaarSeg Segmentation

SnoopCGH implements the HaarSeg algorithm to segment CGH data. A segmentation layer can be added for any of the isolates in the plot and combined with a filtering layer to analyse the distribution of points and peaks. The main configuration parameters for this algorithm are:

- **FDR q value**: False discovery rate threshold (0 < q < 0.5)
- **Start Level**: The detail sub-band from which we start to detect peaks. The higher this value is, the less sensitive we are to short segments. This value is interpreted as 2^n.
- **End Level**: The detail sub-band until which we detect peaks. The higher this value is, the more sensitive we are to large trends in the data. This value does not indicate the largest possible segment that can be detected. This value is interpreted as 2^n.



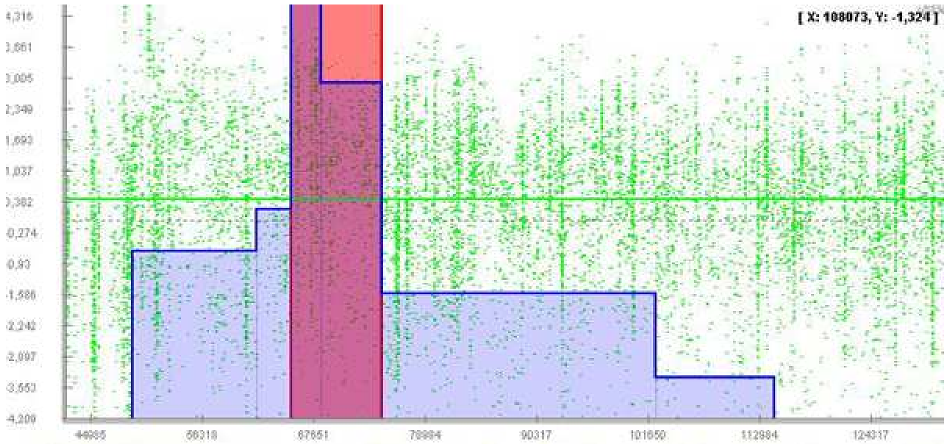
In addition, you can perform a pre-filtering analysis in order to remove outliers based on a Z-score test using the median absolute deviation (MAD). (http://en.wikipedia.org/wiki/Standard_score, http://en.wikipedia.org/wiki/Median_absolute_deviation)

For a detailed explanation of this algorithm please check [1].

SW-Array Analysis

SnoopCGH implements the SW-Array algorithm to automatically detect CN variations in the CGH data. Notice this algorithm could take a long time to be executed over large data sets. The graphical components in the SW-Array layer are:

- **Threshold**: A green line indicating the threshold value used to discard outliers.
- **Robustness**: Semi-transparent blue regions. The robustness value for a region is proportional to its height in the layer (100% if it is filled from bottom to top).
- **Candidate Islands:** Semi-transparent red bars. The regions that are candidates for CN variations.



The main configuration parameters are:

- **Permutations**: The number of permutations for the significance test.
- **Max. Islands**: The maximum number of islands to be considered.
- **Robustness Iterations**: The number of times the algorithm will be executed with different threshold values in order to calculate the robustness.
- **Minimum Robustness**: The minimum robustness an island must have to be considered in the layer as a candidate for CN variations.
- **Minimum Significance (lower than)**: The significance limit value an island must have to be considered in the layer as a candidate for CN variations.
- **Detect Polysomy/Deletions**: Establishes if we are looking for CN increments or decrements.
- **Alpha value:** Weight value for the threshold calculation.

For a detailed explanation of this algorithm please check [2].

References

[1] Ben-Yaacov E, Eldar YC. *A fast and flexible method for the segmentation of aCGH data*. Bioinformatics 2008; 24:1139-45.

[2] Price TS et al. *SW-ARRAY: a dynamic programming solution for the identification of copy-number changes in genomic DNA using array comparative genome hybridization data*. Nucleic Acids Res 2005; 33:3455-64.