CNVCat

CGHPro Edition

Version 2.0

Entwickelt von:

Katrin Tebel [tebel@molgen.mpg.de]

Reinhard Ullmann [ullmann@molgen.mpg.de]

Introduction

CNVCat has been developed to detect and visualize DNA copy number variants (CNV) within a set of array CGH data by user defined parameters.

The application supports the following features:

- Detection of CNV within your data -“Extract CNV (CNV calling)”
- Grouping of experiments to highlight similarities and differences -“Define groups and colours”
- Comparison of CNV between different cases or groups of cases CNV in distinct view modes -“Explore CNV ”
- Export lists of CNV, export chromosomal or genomic view as images

Dependencies

CNVCat as part of the Software Package CGHPRO depends on the CGHPro database (based on MySQL). CGHPRO is a program for the primary analysis of array CGH data [1].

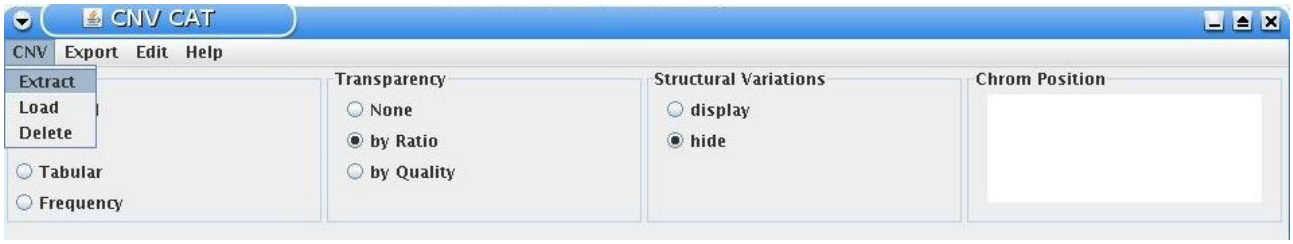
Before using CNVCat array CGH data have to be imported into the database following the workflow as described in the CGHPRO manual.

Extract CNV (CNV calling)

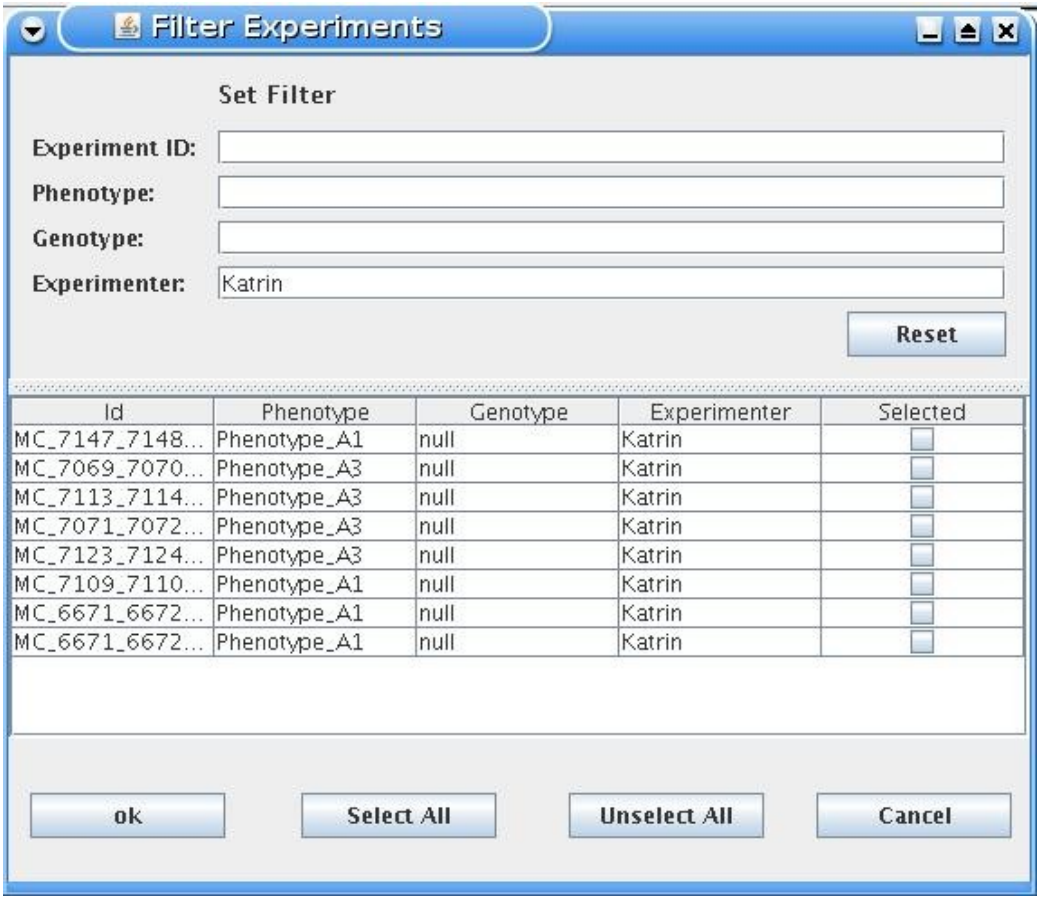
This feature is dedicated to the identification of CNV in single cases or groups of cases by custom defined parameters (CNV calling).

The procedure to detect CNV within an experiment basically scans all datapoints along the genomic location to find consecutive regions exceeding a certain threshold.

To run the extraction procedure select the menu item "Extract" from the menu "CNV".



The dialogue "Filter Experiments" will appear.



This dialogue allows to search for experiments from the CGHPro database.

To get a certain experiment use the filter fields above the list. Enter the full text (i.e. experiment identifier) or only a matching part and hit enter. If no filter should be applied leave all fields empty and press enter. All experiments matching the search criteria will be listed.

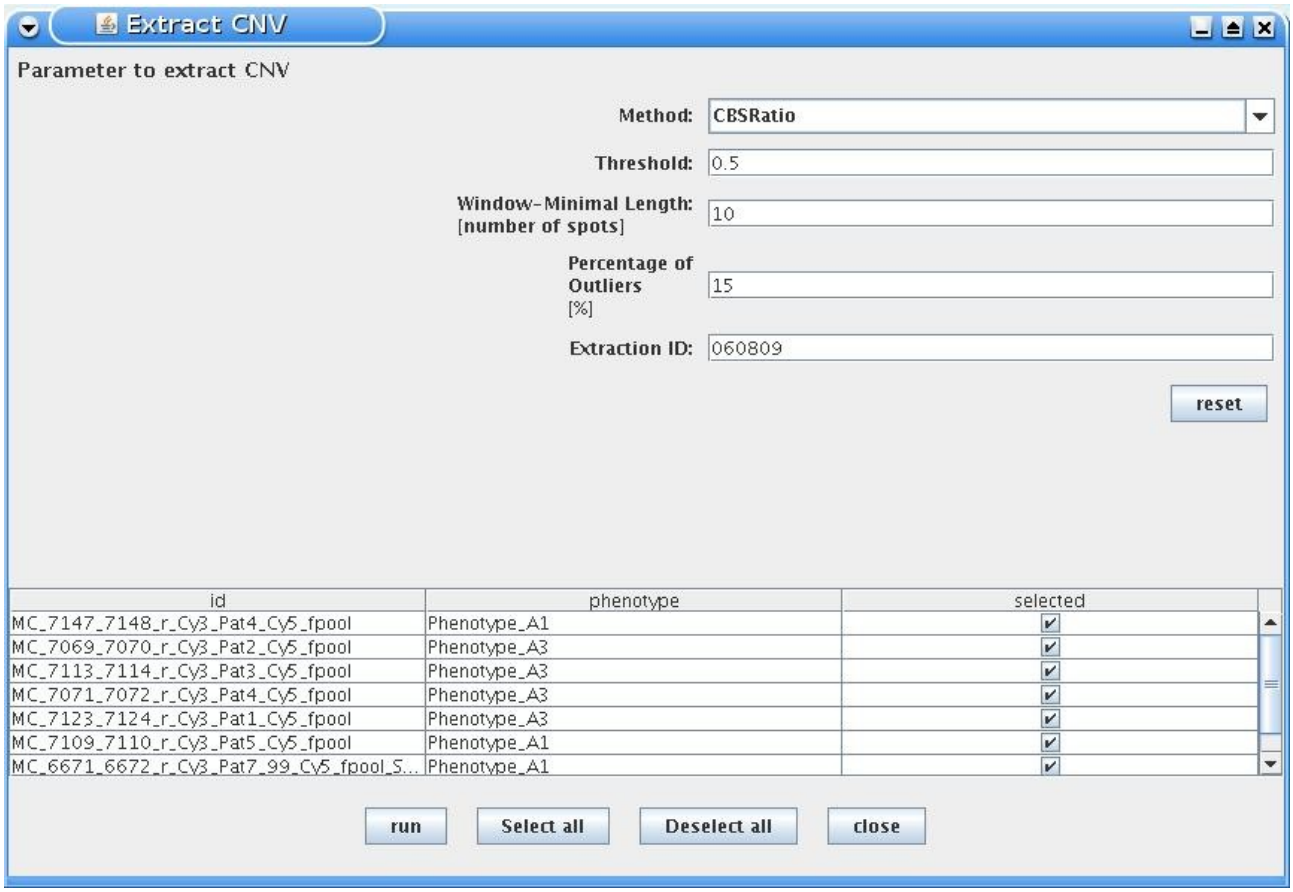
For each experiment the experiment identifier, the corresponding phenotype, genotype and the experimenter as stored in the CGHPRO database are presented.

With *"*Reset*"* the filter restriction will be removed.

CNV calling will be performed only for those cases which have been selected either by manually setting the hook or by pressing the "Select All" button.

Setting extraction parameters

Sensitivity and specificity of CNV calling is heavily dependent on the extraction parameters used. CNVCat offers the user the chance to set various parameters to adapt to the quality and biological question.



Extraction Parameters:

Method:

- Ratio – the (normalized) ratios will be used to extract CNV.
- Ratio CBS –if stored in the CGHPRO database, ratios smoothed by circular binary segmentation [2] will be used. For more information about CBS see our CGHPro manual.(REF)

CBS ratios spread less than the non smoothed ratios, this has an impact on the parameter "Threshold".

Threshold:

Defines the ratio value that classifies data points as aberrant.
The threshold value is used for ratio shifts in both directions, gains and losses.

Window – Minimal Length [number of spots]:

Defines the minimum number of consecutive data points with ratios beyond the threshold necessary to be recognized as an aberration. If the number of consecutive aberrant spots doesn't reach this size they will not be interpreted as CNV.
A CNV may encompass outliers, but those will not count for the window length.

The bigger the window the less resolution and sensitivity the CNV calling will have!

Percentage of outliers [%]

Within a potential CNV some data points can deviate from the common trend. These outliers can have ratios below the threshold or even ratio shifts in the opposite direction.
This parameter defines the percentage of tolerated outliers within a potential CNV. Those percentage is consecutive calculated (based on the current number of aberrant data points and outliers). If the percentage of outliers exceeds the given value the already detected aberrant datapoints are interpreted as CNV if the region already fulfils the criteria of minimal window length. Otherwise the recent aberrant region is skipped and the scanning for the next aberrant data points is proceeded.

If the percentage of tolerated outliers is too low, a continues CNV could be split apart. On the other side, too much tolerance could lead to a high false positive rate. There is also the risk that e.g. a small deletion in between two duplications could be missed!

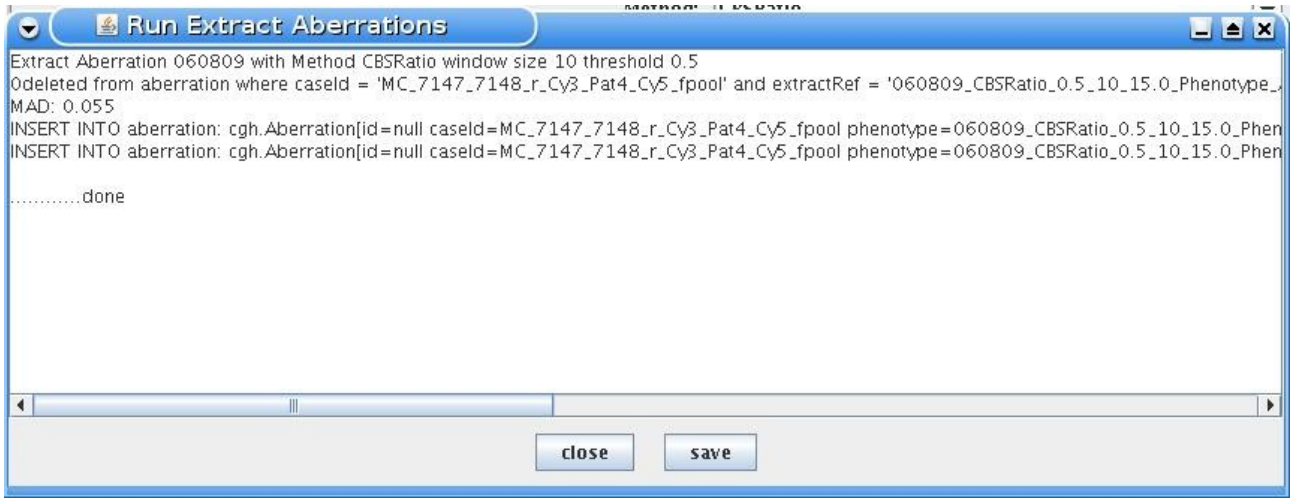
Extraction ID:

This label helps to identify the set of extracted CNV later. A useful setting could be the current date. There is no need to include extraction parameters within the ID as these values are stored automatically.

With “run” the extraction procedure starts for all selected experiments. When the extraction procedure is finished the program returns to this dialogue, so that you can run the extraction for a different subset of the cases or re-run with modified parameters.

During the extraction a window appears that reports the current state of the calculation progress. At the end the report can be stored into a separate file.

All CNV are deposited at the internal database by the program automatically and can be exported into a csv-file via the export function (see “Export CSV File”).



Structure of CNV data

For each CNV the following data are stored inside the database and can be exported into a csv-file via the export function (see “Export CSV File”). Those data are later visible at the main display view along with the plotted CNV (how to load CNV see “Loading CNV sets“).

Structure of CNV data:

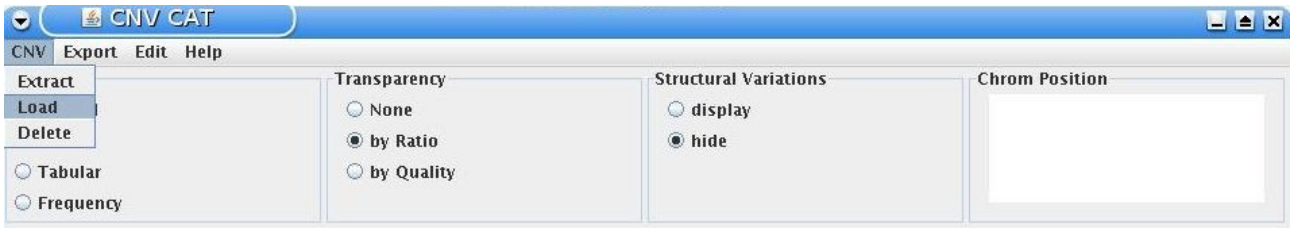
(labelling used by the application at the main display view is put in brackets):

- chip or case identifier (Case ID)
- Parameter for CNV calling (Parameter[Extraction])
  All user defined parameter are collected into one text string. This string has the following structure:

  ExtractionId_Method_Threshold_WindowMinimalLength_PercentageOfOutliers_Phenotype
- chromosome, start and end position of the CNV (Loc Start, Loc End)
- number of spots (Count), first and last spot identifier (StartClone, EndClone).
- average ratio (Ratio)
  The mean ratio of all aberrant spots within the CNV.
- quality score ratio (Quality)
  The averaged ratios are qualified by a value representing the deviation within the data.
  If available, the MAD as “median averaged deviation” , calculated during the data smoothing by CBS, is taken to qualify the CNV.
  Otherwise the MAD as “median absolute deviation” of the ratios is taken.
  The quality score is calculated for each CNV as the (averaged) ratio devided by the MAD.
- type of aberration (Type)
  “Deletion” or “Duplication”

Loading CNV sets

To visualize CNV within the main display view select the menu item “Load” from the menu “CNV”.

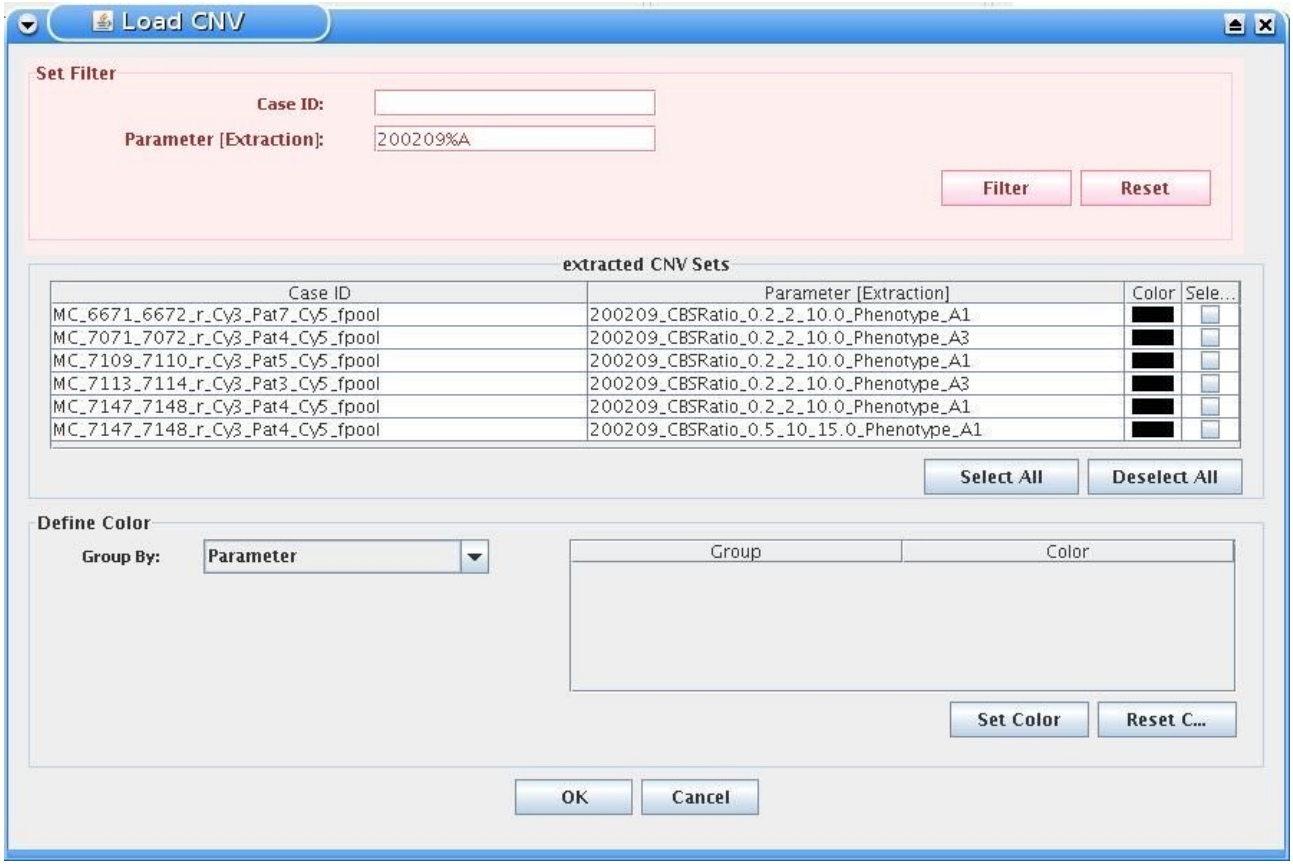


Perform the following steps to retrieve CNV sets from the database and customize their display properties:

- define a subset of CNV sets by setting some filter criteria
- assign group and colour attributes to the filtered CNV sets.
- load all selected CNV sets to be visualized.

Set filter criteria

To access the CNV sets from the the database put some search criteria into the filter fields and press the button “Filter”. All CNV sets matching the filter criteria are presented at the table “extracted CNV Sets”.



To get an overview about all available extracted CNV sets press the button “Filter” without any input at the filter fields.

The filter field “Case ID” is used to filter by the experiment identifier. A valid input would be the complete experiment identifier (to filter out one single extraction experiment) or a matching part of it. The filter field “Parameter [Extraction]” is used to filter based on the extraction parameters.

It is possible to filter the extraction parameters by two ore more search parameter. To do so join the values with the place holder character '%'

e.g. “110209%0.2%A”

where “110209” stands as a matching pattern for the Extraction Label, “0.2” for the threshold and “A” as pattern for the phenotype.

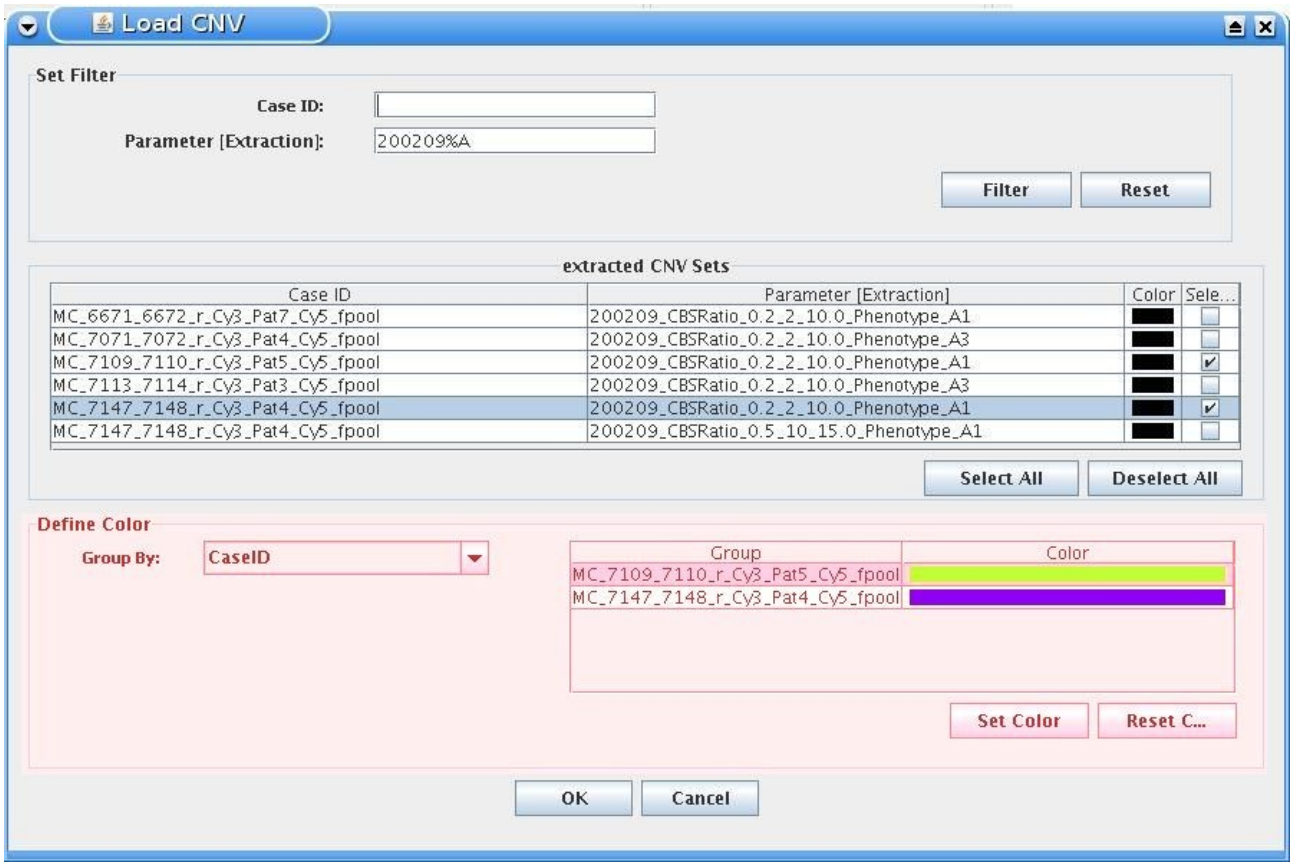
Define groups and colours

A particular colour can be assigned to each CNV set enabling their distinction in the visualization. Per default all CNV will be displayed in black unless a customized colour is chosen.

Independent CNV sets can also be grouped together and a single colour can be assigned to this group.

The colour of a CNV determines the membership of a group. All CNV sharing the same colour are considered as one group.

The feature colour setting is of particular interest when you want to compare the frequency of CNV in e.g. two distinct disorders. Select the CNV set(s) representing disorder A and B, respectively. If there is more than one CNV set for one disorder, combine them in one group. Assign one colour to each disorder. As outlined below, CNV derived from both disorders can be displayed simultaneously and easily distinguished based on their colour code.



To assign groups and colours perform the following steps:

Check the group type:

Choose one of the following group types from the box *"Group by":*

CaseID:

All CNV extracted within one experiment will get the same (unique) colour.

Parameter:

All CNV sets sharing identical extraction parameter will get the same colour and will be assigned to one group.

None:

All selected CNV sets will get the same colour and will be considered as one group

Select CNV sets

To visualize CNV from one or more particular extraction experiment set the hook at the corresponding boxes in the column "Selected". To display all extraction sets use "Select all".

According to the chosen group type and the selected CNV sets all available groups are listed at the group table.

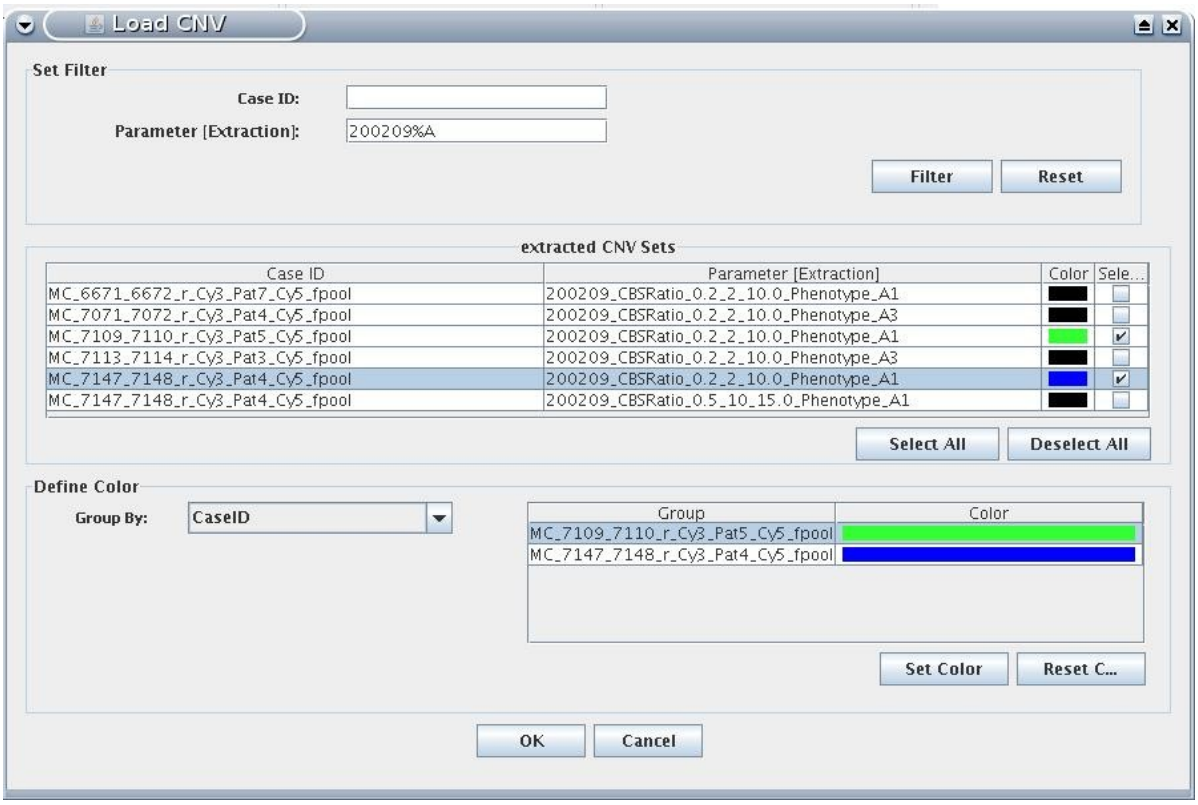
If the filter criteria or the group type is changed again, all listed CNV sets are reset as "unselected" and must be selected again.

Set the group colour

For each group a unique colour (taken from a set of 10 different colours) is proposed.
To modify the pre-set colour click the field "Color" at the group table. The colour editor is opened and an a colour can be chosen.

If more than 10 different groups are selected the colour for the 11th group has to be modified by the user, due to the fact that all CNV sharing the same colour are considered as one group.

Press “Set Color” to assign the group dependent colours to all selected CNV sets.



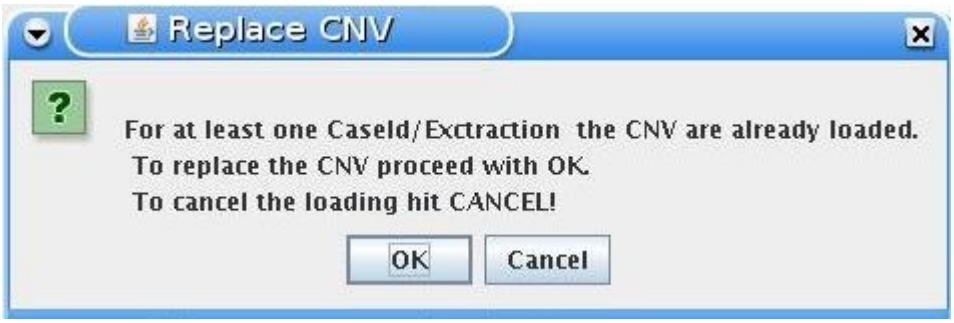
It is possible to modify the assigned colour for a certain CNV set by clicking the field “Color” at the table “extracted CNV sets”. Please remark that changing the colour will also change the the group membership.

Display CNV sets

Press “OK”.to load all CNV of the selected CNV sets. The CNV are added to the already visualized CNV.

If some of the newly loaded CNV have the same colour attribute as any of the already visualized CNV they will get part of the same group.

If some of the newly loaded CNV sets already have been visualized a warning dialogue appears:

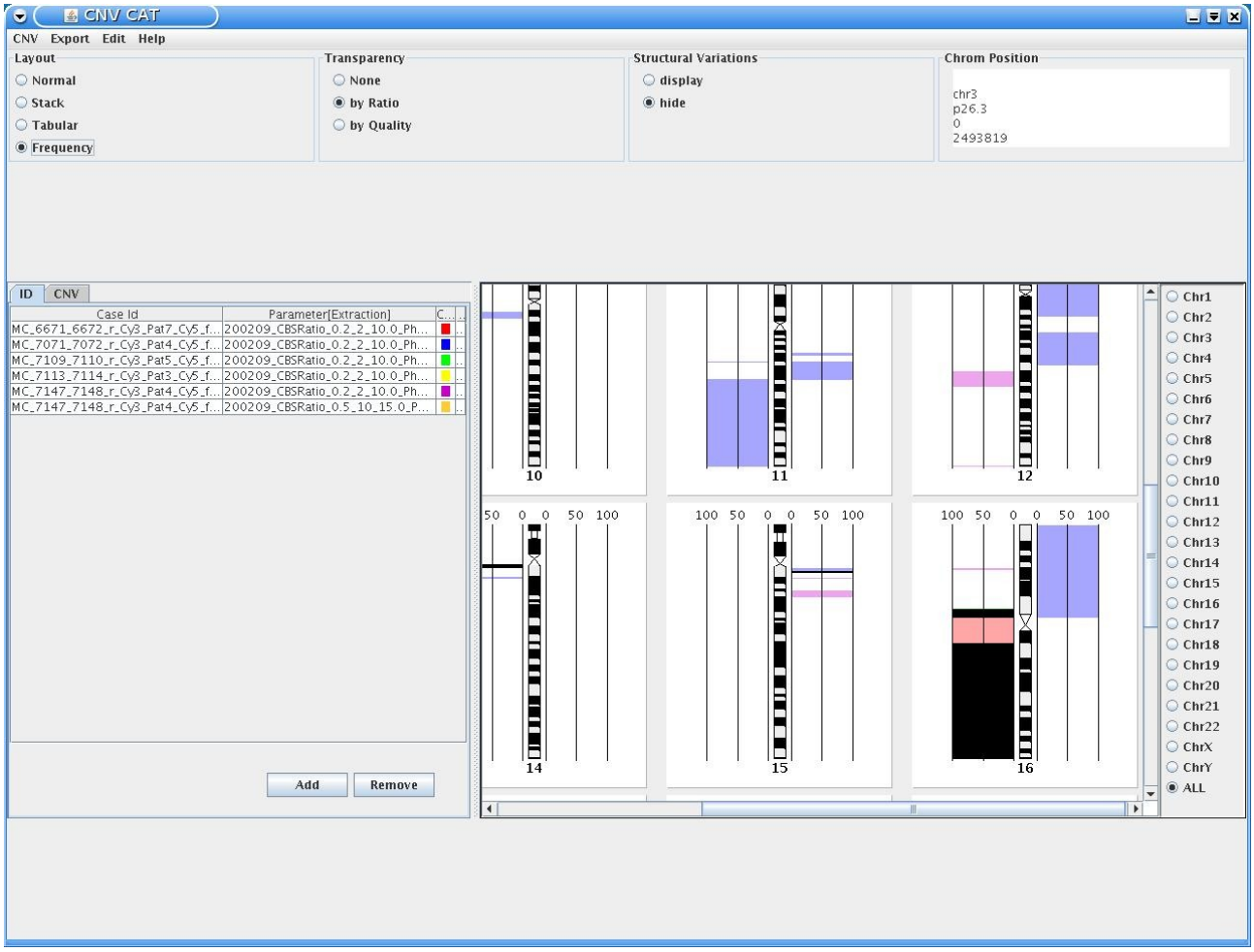


Pressing “OK” will replace the formerly loaded CNV and their attributes, “Cancel” will stop the loading and the formerly loaded CNV will remain unaffected.

Explore CNV

Main display view

The main window depicts all loaded CNV sets. They can be displayed for each chromosome separately (by selecting the chromosome at the right-hand tab pane) or as genomic view, where all chromosome are shown at once (tab “ALL”). Use the right mouse bottom to zoom in or out the chromosome view.



The left hand table at the main screen named “ID” shows case identifier, extraction parameter and the assigned colour for each loaded CNV set.

The table named “CNV” provides detailed information about CNV, like genomic position , ratio, number of spots, first and last spot identifier etc. How to get this data see “Show Details on CNV” for more information.

The assigned colour can be changed for each CNV set by clicking the colour field at the table "ID". The colour editor is opened and and a suitable colour can be chosen. Please remark that changing the colour of the CNV will also change its group membership.

The size of the table view and the chromosome view can be modified by dragging the central split bar with the mouse.
Each field itself is resizeable by mouse, its position can be modified by drag and drop.
To sort the list click the header of one of the fields.

The button "Add" opens a dialogue box for loading additional CNV sets. Pressing the button "Remove" will clear all data tables and views.

Layout modes for the chromosome and genomic view

CNVCat features the following layouts:

Normal:

The CNV are displayed from top to bottom along their genomic position. Overlapping CNV from the same set are placed at the same column, if the overlapping CNV belongs to a different CNV set it is moved to the next column.

Stack:

Displays CNV in the most condensed form. Overlapping CNV are moved to the next column, but the first non overlapping CNV is placed in the first column again.

Tabular:

All CNV identified in a single CNV set are placed and displayed at their own column.

Frequency:

This view depicts the relative frequencies of CNV in groups respective the genomic position.

The frequency is calculated for each distinct genomic region. The size of these region is given by mapping each screen pixel to a certain genomic location. The resolution of the frequency calculation therefore depends on the zoom factor.

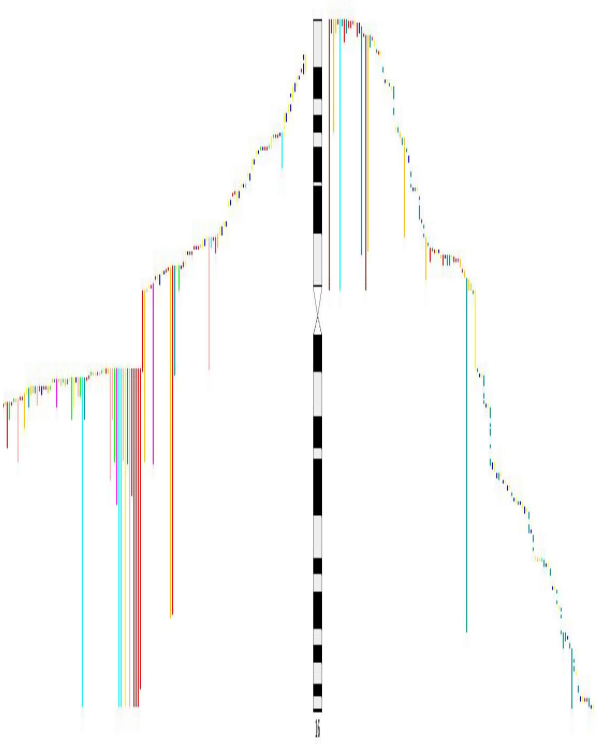
$$f(ij) = \frac{n_{ij}}{N_j}$$

n_{ij} *number of CNV sets of group j with CNV at genomic position i;*
N_j *total number of CNV sets of group j;*

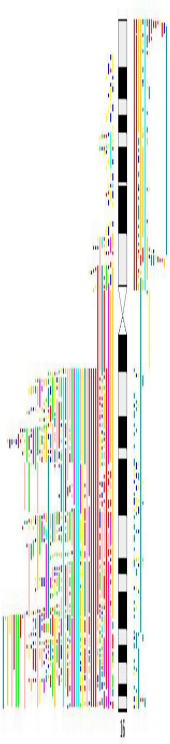
If CNV from two (ore more) groups overlap, the minimal shared relative frequency is indicated by a separate colour (default setting is grey).

The colour indicating overlapping frequencies is customizable; see "Customize database connection and application properties".

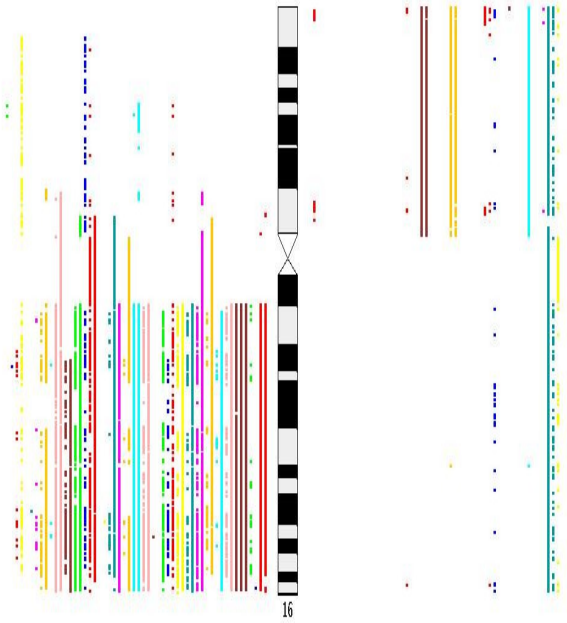
Normal



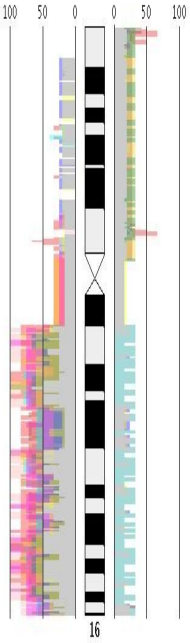
Stack



Tabular



Frequency

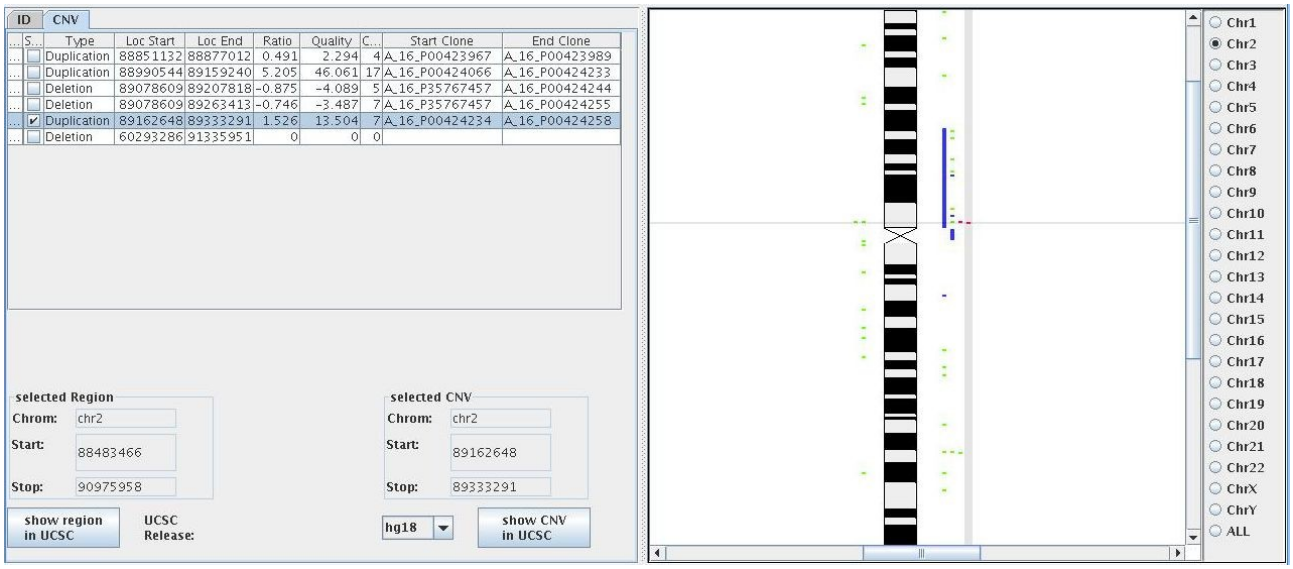


Setting Transparency

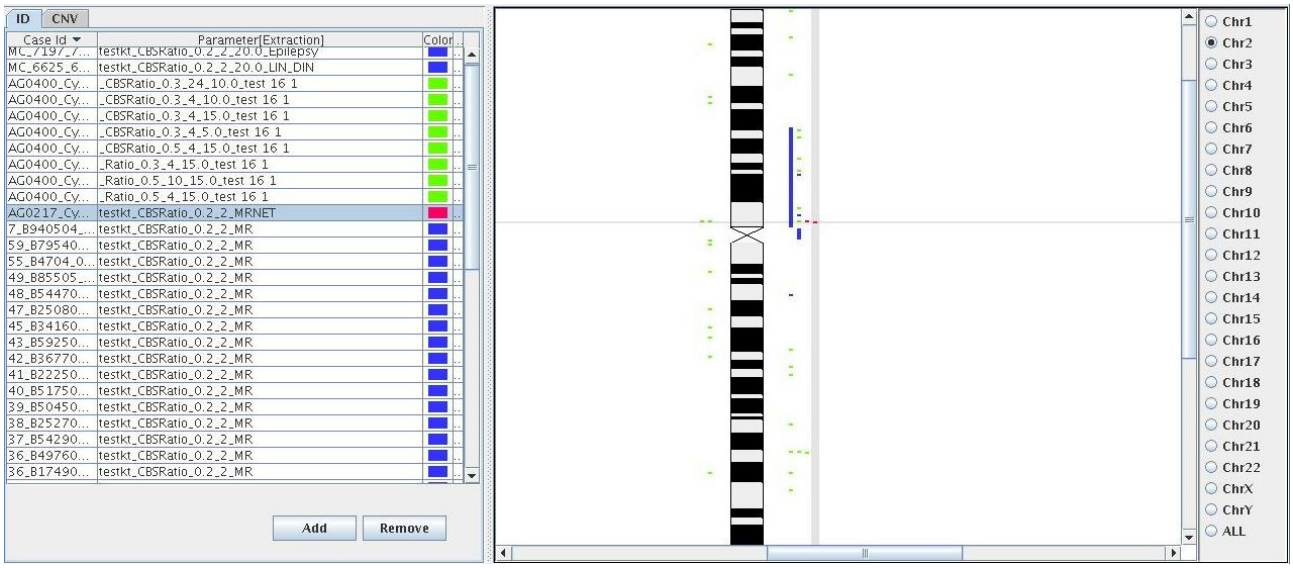
The bars representing CNV at the right hand view pane are displayed with a certain degree of transparency. The base for this transparency can be chosen from the panel „Transparency“. If “none” is selected, the CNV are displayed full coloured. Otherwise, transparency is set according to the quality score or the extend of the ratio shift.

Show Details on CNV

To show detail information for a certain genomic region or a CNV of interest click the mouse within the chromosome view. The CNV exactly beneath the mouse pointer will be highlighted by a horizontal and a vertical bar. The table „CNV“ will be refreshed to show all CNV located in the selected region, the item for the selected CNV will be marked with a hook.



To show the data from the associated CNV set like experiment ID, extraction parameter first select a certain CNV by clicking the mouse at the respective item at the table “CNV”. Second switch to the table “ID” by clicking the tab. The selected single CNV item and the associated item for the set data are indicated with a highlighted background.



It is also possible to select a CNV directly be replacing the hook within the table „CNV“. The newly selected CNV will be highlight at the chromosome view.

The colour of the highlighting bars can be customized; see “Customize database connection and application properties”.

The panel “selected Region” gives the chromosome coordinates for the selected genomic region, while the panel “selected CNV” gives the chromosome coordinates for the selected CNV.

With the buttons *“show ... in UCSC”* these intervals are displayed inside the “UCSC Genome Browser” [3]. The internet browser is opened as an external window. The appropriate UCSC's release must be selected beforehand.

Note that the CNV data are not automatically converted to the different releases!

Show Structural Variations

The database CGHPRO maintains data from the “Database of Genomic Variants” [4].

To display these variations choose the associated button at the panel “Structural Variations”.

The structural variations are superimposed as horizontal bars onto the chromosome view (placed over the cytoband). Note that we have not collected these dataset ourselves. Thus, if you are referring to any data from the Database of Genomic Variants, please cite the appropriate reference [4].

Customize database connection and application properties

To customize the application properties or to define the database connection select the menu item "Properties" from the menu "Edit".

All user defined properties within this dialogue are saved to an external xml-file (named "coreproperties.xml") and are therefore available after restarting the application.

Database connection

If CNVCat is started for the first time the parameters to connect to your local CGHPRO database must be specified.

If CNVCat runs inside the CGHPRO application these database connection information are already set. In this case there is no need to change the database parameter here.

Setting this connection parameter is essential to run CNVCat as standalone application (note remarks above), otherwise you will get the following error message:

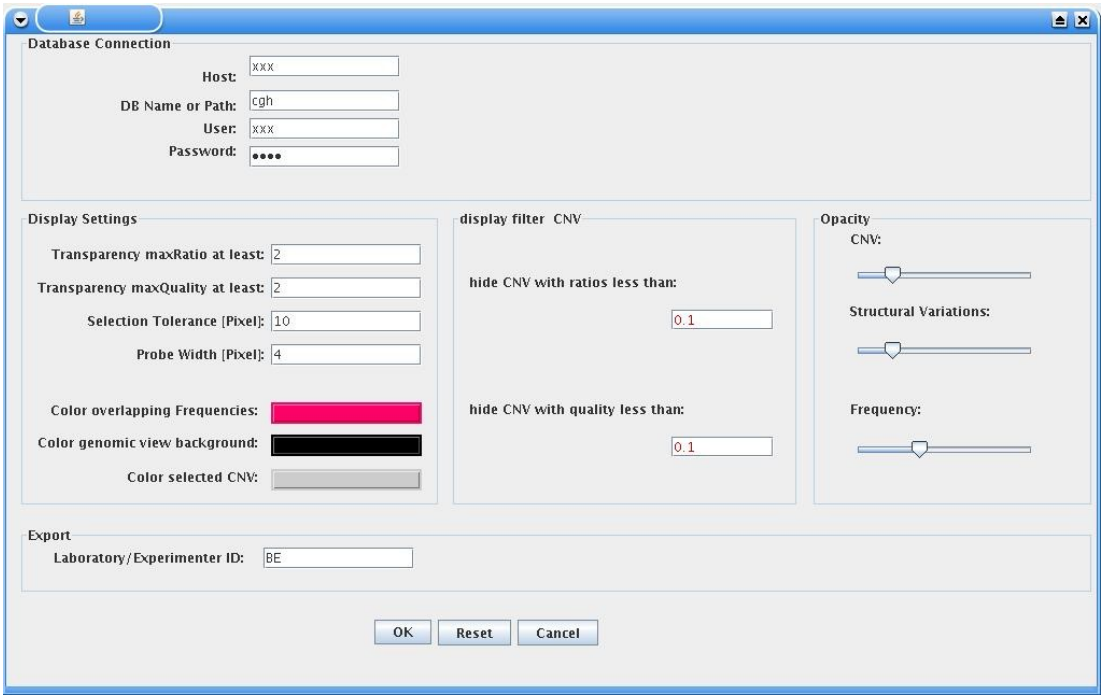
To set up the database connection you have to enter:

Host: computer name hosting the database

Database name: name of the database

User: user with access rights to the database

Password: password for the given user



CNVCat offers a range of options to customize the standard application behaviour, mostly regarding the display properties.

Display settings:

Transparency maxRatio at least/Transparency maxQuality at least:

The CNV are displayed with a certain degree of transparency.

The transparency reflects the intensities of the ratios , CNV with high ratios are displayed with less transparency (rich coloured) and vice versa.

In order to scale the transparencies the highest ratio value of the currently displayed CNV is taken as the upper bound. These upper bound will be replaced by the value set here, if it is higher than the calculated one.

If for example the ratios of the currently displayed CNV span only values between 0.02 and 0.09 the transparency is scaled within this minimal range. The CNV with the ratio value of 0.09 is displayed rich coloured.
This is not always the desired behaviour, because those low ratios are unremarkable and should be displayed less coloured. To avoid this it is possible to define a minimum value for the upper bound.
If this value is set to 2.0 all CNV with ratios beneath will be displayed with a degree of transparency, all ratios above will be displayed full coloured.

Selection Tolerance (Pixel):

By clicking the mouse within the chromosome view the selected CNV is highlighted and information about all CNV within this region are shown at the table "CNV".

The selected region is located by mapping the pixel position (i.e. the y coordinate) of the mouse cursor to a genomic location.
To help on hitting a specific CNV by mouse click the covered area can be extended by setting the parameter "Selection Tolerance".

If for example the selection tolerance is set to 2, the area spanning 2 pixels above and beneath the current mouse position is added to define the selected genomic region.

Probe Width (Pixel):

This value defines the display width for the bar representing one single CNV.

Colour overlapping frequencies:

Defines the colour for the overlapping regions at the view mode "Frequency" (see "Layout modes for the chromosome and genomic view" for more information about the frequency view mode).

Colour genomic view background:

The genomic view displays all chromosome within one picture. The background between the single chromosome panels can be changed here.

Colour selected CNV:

By clicking the mouse within the chromosome the selected region is highlighted with horizontal and vertical bars. The colour for this bars can customized here.

Display filters:

hide CNV with ratio less than:

CNV with absolute (negative/non negative) ratios less the given threshold are removed from the main CNV view pane.

hide CNV with quality less than:

CNV with absolute (negative/non negative) quality scores less the given threshold are removed from the main CNV view pane.

Opacity:

To optimize the CNV display properties you can customize the transparency for CNV by moving those sliders.

Less opacity means more transparency and vice versa.

CNV are displayed with a certain degree of transparency, which depends on their ratio or quality scores. Changing the settings here affects the range of this transparency.

Export:

Laboratory/Experimenter ID:

This identifier is useful for the export of CNV (see "Export CSV File").

The text will be added prior to the experiment identifier. This may be helpful to index the result data by laboratory or experimenter information.

If there is no need for such index this field should be left empty.

Additional Features

Export CSV File

A list of CNV sets can be exported as a csv-file.

The export is started via the top menu bar. A dialogue window opens where you can specify the cases you want to export (see “Filter CNV set”).

Next you are asked to choose the directory and to name the csv file. The file extension (“.csv”) is added automatically. The successful export is reported by the following message:

Export Images

The pictures of the chromosomal or genomic view pane can be exported as image files.

Selecting the menu item “Image” from the menu “Export” will open a dialogue box in which you can name the image and define the folder where the current view pane should be stored.

The image format must be defined manually by adding the extension “.jpg” or “.png”, respectively. Currently, no other formats are available.

Delete CNV set

This submenu offers you the chance to remove CNV from obsolete extraction sets. After selecting one or more CNV sets from the filter dialogue (see “Filter CNV set”) the chosen CNV will be deleted.

The successful deletion is reported by the following message:

Filter CNV set

This dialogue pops up prior to one of the functions described above and helps to define a subset of CNV sets.

To select on ore more CNV sets place the hook at the field “Selected” at the respective item or use the button “Select all”.

If you type in some text into the filter field the database is queried and the list of CNV sets is updated immediately by typing.

So the input at the filter field can be delayed if the list contains a great amount of data.

Bibliography

1: Chen, F. Erdogan, H. H. Ropers, S. Lenzner, and R. Ullmann, Cghpro – a comprehensive data analysis tool for array cgh, 2005

2: Olshen AB, et al, Circular binary segmentation for the analysis of array-based DNA copy number data, 2004

3: Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D, The human genome browser at UCSC, 2002

4: Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW, Lee C, Detection of large-scale variation in the human genome, 2004