CNAG2.0 Manual

I. Overviews

CNAG2.0 was developed to enable high-quality analysis of copy number alterations and allelic imbalances in cancer genomes, congenital disorders, and normal individuals using Affymetrix GeneChip platforms. In addition to the algorithms already implemented in CNAG1.0 for eliminating systemic deviations of copy number calculations, a number of new features have been added to maximize functionality and facilitate analysis of multiple samples. Though not comprehensive, this manual is intended to provide some guidelines for use of this program, explanations about how to use the program. However, before delving into the details, we briefly overview the entire sequence of analysis involved when using CNAG2.0, as this sequence has entirely changed since the previous version

1. Extracting data from .CEL files and .CHP files.

Instead of manually exporting and preparing data from each array, data required for the subsequent analysis is directly extracted and imported from multiple .CEL files and .CHP files by CNAG2.0 using “Extract Data” from the “data” menu. Through this step, each array data is assigned into one of the four categories by the user, 1) test (e.g., tumor) samples without a paired reference from the same individual, 2) reference samples without their corresponding tumor samples, 3) test samples that should be paired with a specific reference and 4) reference samples that should be paired with specific test samples. When extracting the data, “array data files” are created and stored in the directory assigned in the “setting” menu. These new files have an extension of .CFH and are ready for analysis.

2. Analysis of array data

The extracted data can be analyzed for copy number/allelic imbalances either manually or automatically in the background by using “sample manager” from the “data” menu or “background analysis” from the “automation” menu. After the analysis, “array result files” having extensions, .CFN and .CFS, are created for each array data file and stored into the directory assigned in “setting” from the data menu. Several attribute may be also set using Sample Manager, which can be used for filtering or grouping samples in the “Display Data” thereafter.

In Sample Manager, each result is displayed immediately after each analysis is done. At this point, copy number analysis has been performed using default parameters assuming that all the autosomal regions are diploid and the mean log2 ratio for single copy regions should be 0.49. Usually, however, more appropriate parameter values should be provided by users for getting more elaborating copy number analysis.

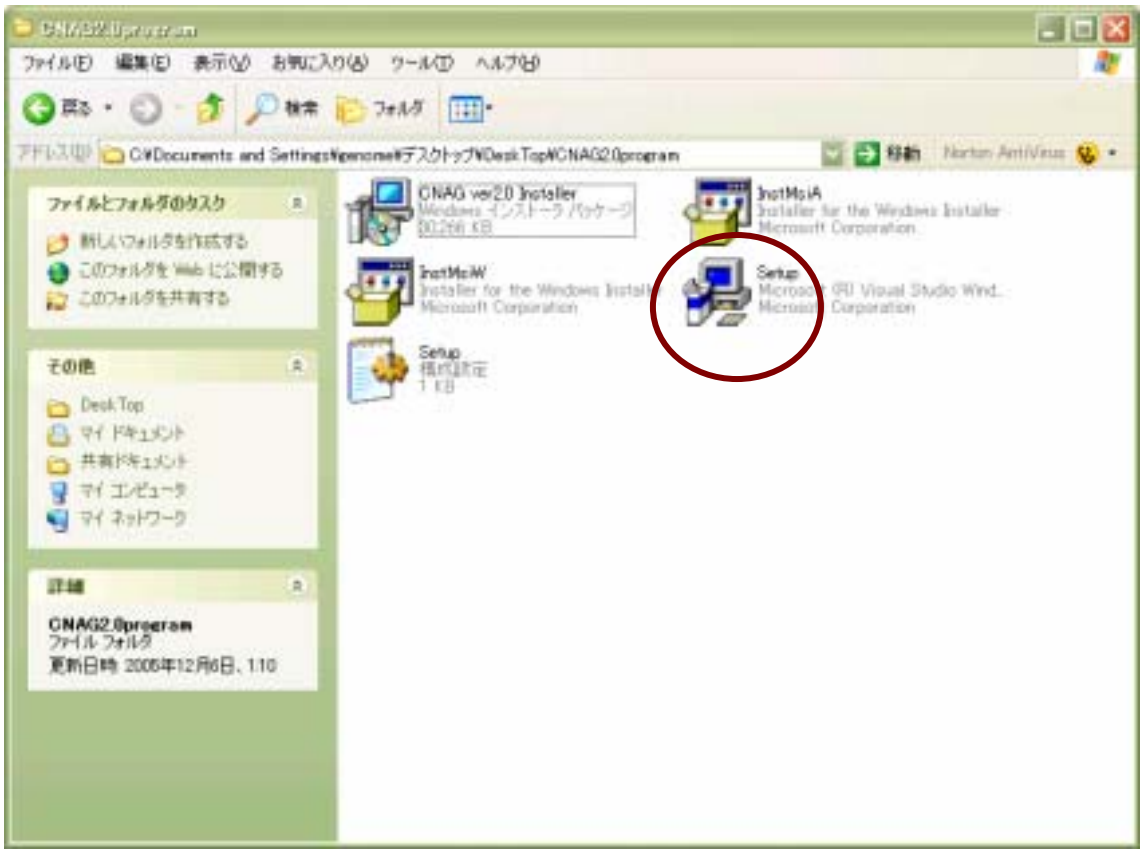
3. Viewing the results

The analyzed data is individually displayed after each sample is analyzed in Sample Manager. The “Display Sample” mode, from the “data” menu, provides greatly extended viewing functions, allowing the user to toggle between different samples. Results of multiple samples can be displayed at one time, and the user may group any samples for this purpose. Moreover, this mode enables integration of all abnormalities within a set of samples.

II. Installation and initial setting

2.1 Installation from the setup program

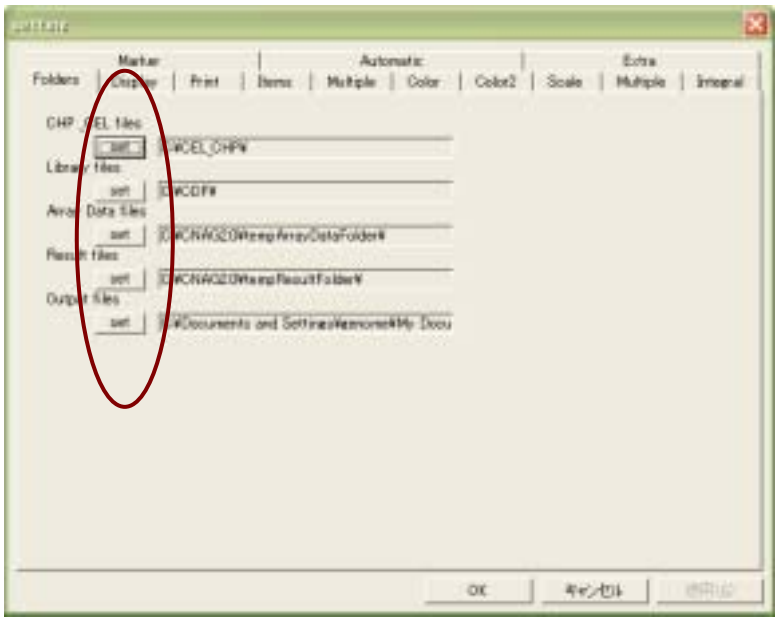
CNAG2.0 is provided with an installer program. The user can install CNAG2.0 by double clicking on the CNAG2.0 “Setup” icon. Any publication of the results obtained using CNAG2.0 are required to refer to our original papers, and this should be agreed upon before installation. Any commercial use of CNAG2.0 is strictly prohibited unless otherwise licensed.



2.2 Setting

Before using CNAG2.0, you need to configure the settings, where the location of several files and directories used by the program are specified. To do this, select “default

setting" from the data menu. The following window will appear:



2.2.1 Specifying the directories that contain .CHP/.CEL files and .CDF files.

CNAG2.0 extracts the array data from .CHP and .CEL files created by GCOS and GTYPE using array information defined by .CDF files (library files), which are provided from Affymetrix and which can be downloaded from the Affymetrix website. You need to specify the directories containing these files by clicking the "set" buttons that appear under "CHP_CEL files" and "Library files" tags. When CNAG is installed on the same PC where GCOS and GTYPE (or GDAS) are working, these files should be found in the directories accessible from the PC and thus, can be specified by toggling with the Windows file manager.

2.2.2 Creating the folders for array data files and for result files.

The extracted data is stored in .CFH files within an "Array Data files" folder which you need to create and specify in the default setting menu. Any folder can be used for this purpose but we recommend you create a new folder such as "array_data_files". Similarly, you also need to create a folder, for example, "array_result_files", and specify this folder under "Result files". This is where the analyzed result data will be stored. The result files will have .CFN or .CFS extensions, depending on the mode of the analysis (i.e., non-allele specific or allele-specific, respectively).

2.2.3 Creating the folders for data export

Likewise, you need to create a new folder, such as "data_output" and specify this under "Output files" folder. In CNAG2.0, data is manipulated in binary, but the results can be exported as text files into this folder. It is worthwhile creating a folder called "CNAG2.0"

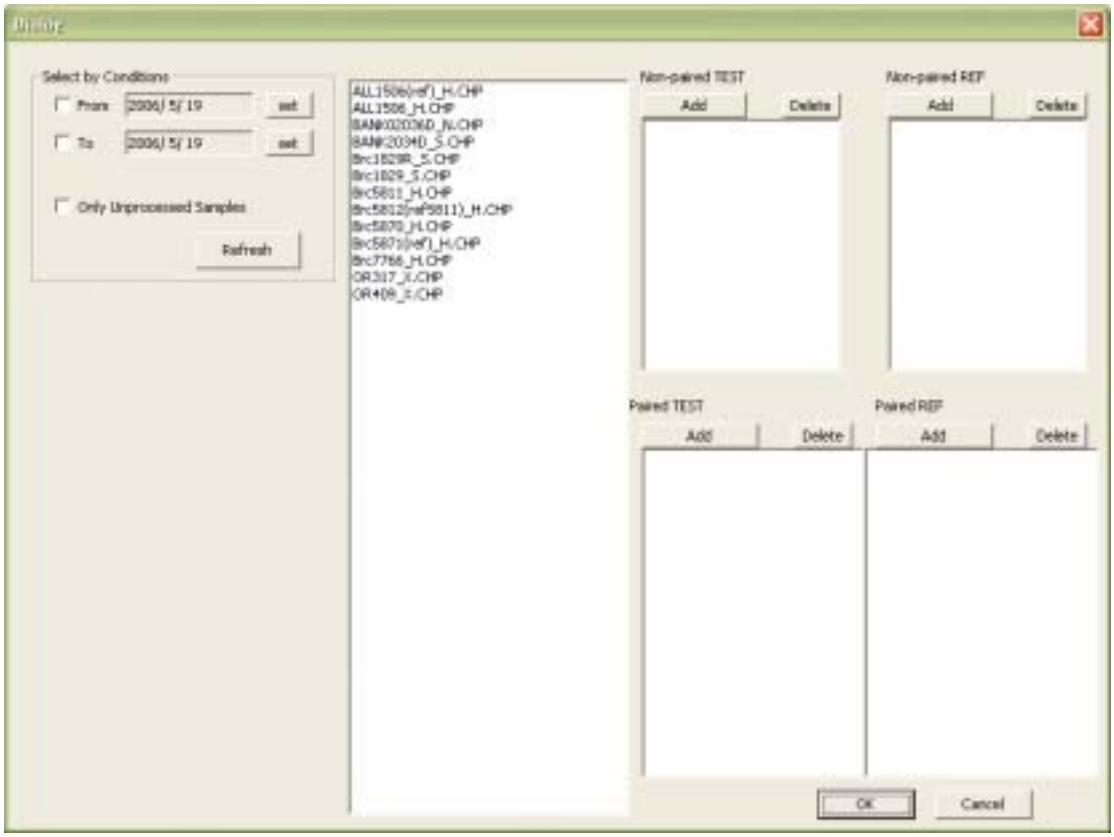
for storing the subfolders described in the above paragraphs. These include three subfolders, “array_data_files”, “array_results_files”, and “data_output”.

2.2.4 Other settings

CNAG2.0 allows users to set a variety of parameters that defines how CNAG displays its results. Users can choose items to display, including copy number plots, LOH inference bars, cytobands, hetero SNP call bars, LOH without copy number loss, copy number inferences from HMM, color-coded copy numbers/LOH, etc. Other settings include color for each item, scales for copy number plots, default magnifications. Copy numbers can be plotted at an equal interval in genetic order or according to real distance. Many of them will not require explanations. The relationship between adjacent SNPs as to the restriction fragments in which they are contained, is indicated by different colors, which is useful to interpret the copy number results. To view and adjust these different settings, click on the tabs within the “settings” window; for example, the “Items” tab allows users to choose which items are displayed.

III. Extract data from .CHP and .CEL files.

The first step to CNAG analysis is to extract signal intensity, genotype information, and gender calls from .CEL and .CHP files. Gender information is used to accurately infer copy number in the X chromosome. CNAG automatically extracts these data to create .CFH files, which are then used for copy number calculations. The user no longer needs to manually export intensity files and SNP files separately from GDAS or GTYPE and to put them into appropriate folders.



#1. Select “Extract Data” from data menu to get the window shown above.

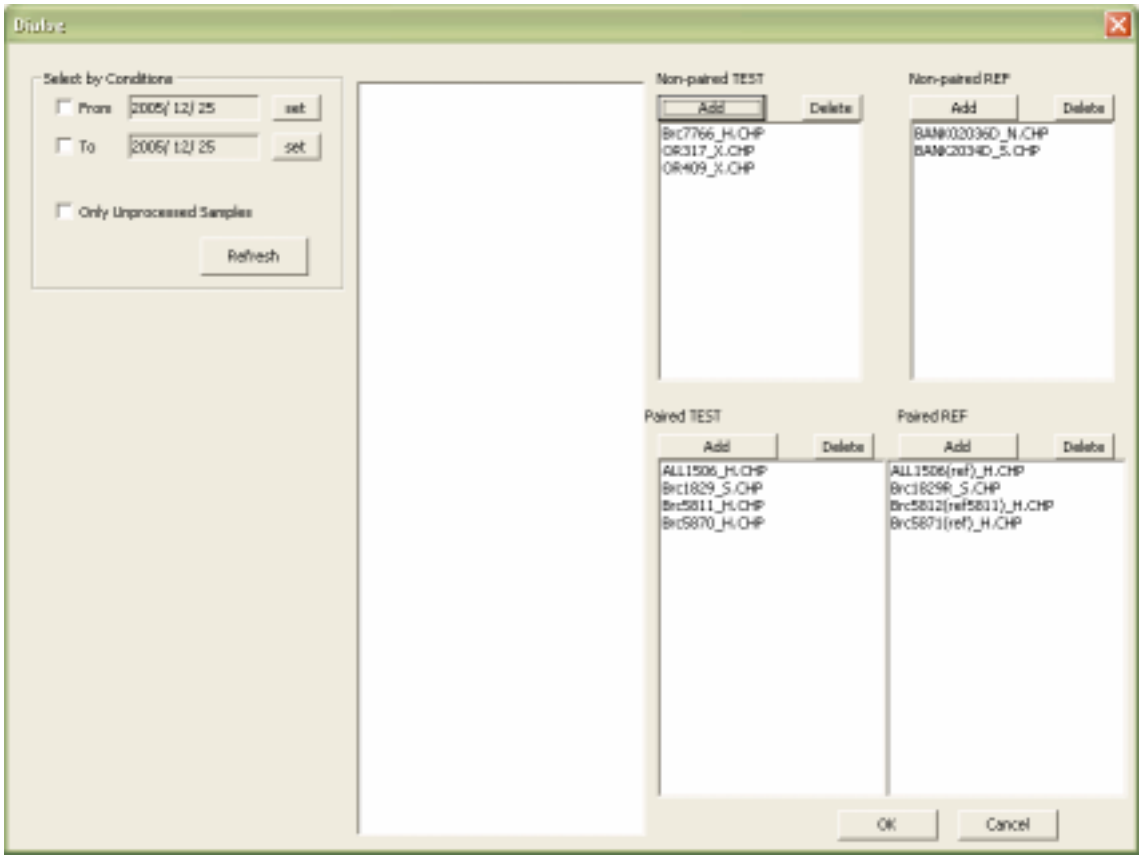
#2. CNAG displays all the samples whose .CEL/.CHP files are stored in the “CHP_CEL files” directory, as specified in “default setting”. You may apply filters according to the date of experiments. The user may also filter to display “Only Unprocessed Samples”.

#3. Assign the samples to be analyzed into four categories:

1) non-paired test sample
2) non-paired reference
3) paired test sample
4) paired reference sample

To do this, select a sample and push the “add” button for the category where it should be assigned. To remove a sample from a category, just select the sample and push “delete”. Paired samples should be placed side by side in the bottom tow windows, so that the first samples in each window will be paired together, the second samples in each bottom window will be paired together, and so on.

#4. After all samples to be analyzed are assigned to their appropriate categories, click “OK” to start extraction.

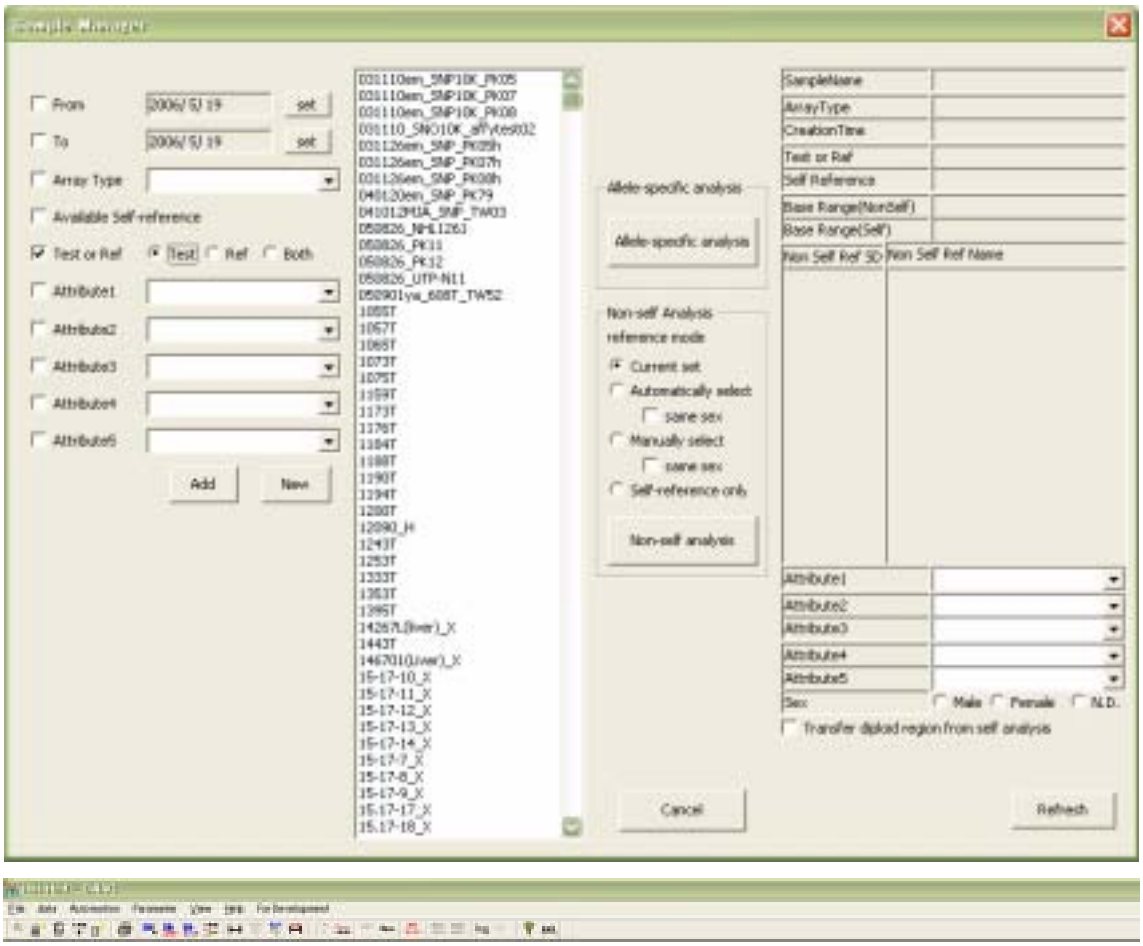


IV. Sample Manager

This mode is designed for manual data analysis, although users may use the "automated analysis" option for this purpose. The user still needs to use this mode to assign user-defined attributes to each array data file.

#1. Select "Sample Manager" from data menu.

#2. CNAG displays all the samples whose array data files stored in the "array data files" folder specified in "default setting". You may apply filters according to the date of experiments, user-defined attributes, array types, sample types (test or reference), or availability of paired reference. To do this, set a filter and click on "New". Samples filtered differently can be combined using the "Add" option. To set an attribute, select a sample, set one or more attributes, then push "Refresh". In CNAG2.0, copy numbers of reference samples can be analyzed, to do this, copy numbers are calculated using all other references besides itself.



#3. Two types of analyses, allele-specific analysis and non-specific analysis, are possible for paired samples, while only non-allele-specific analysis is currently available for

non-paired samples. In the latter case, the “Allele-specific analysis” button will not be available. For allele-specific analysis, just click on “Allele-specific analysis”. For non-allele-specific analysis, you may choose from one of four options regarding how references are selected.

The “Current set” option is available only after the sample has been previously analyzed, and it will use the same set of references used in the previous analysis. In the “Automatically select” option, CNAG performs pair-wise analysis using all of the available references and then computes the best combination of references that minimizes the standard deviation (SD) values for the region selected as “diploid”. By default, all the autosomal regions are defined as a “diploid region”.

In the “Manually select” option, CNAG also computes the SD values for all pair-wise analyses and displays the result in the window on the right. A new window and prompts users to select the references that should be used for copy number calculations. The “Self-reference only” option is available only for paired samples. With this option, CNAG calculate copy numbers using only the self-reference defined during data extraction.

#4. Adjustment of diploid region

After copy number analysis, CNAG displays the results. At this step, you have the option to define a region as diploid using the  icon. This may be useful when diploid status has been precisely determined using a cell-based analysis such as cytogenetics, FISH, or FACS analysis of DNA content. Otherwise, you may select a diploid region by visual inspection, but note that this may erroneously assign diploid region which can significantly skew copy number estimations.

#5 Copy number analysis on X chromosome

To correctly calculate copy number for the X chromosome, gender information of the sample is required. Female gender is clear when a significant number of heterozygous SNP calls appear on the X chromosome, but loss of heterozygous SNP calls can incorrectly indicate a male gender for tumor specimens in which one of the two X chromosomes is deleted. However, for most cases, gender of the specimens is known and the user can correct an erroneously called gender by checking the correct one and clicking on the “refresh” button at the lower right corner.

For copy number calculation on the X chromosome, check the “ same sex” checkbox when performing a non-self analysis. In this case, CNAG will calculate copy number by exclusively using the references of the same sex as the tumor sample.

#6. Adjustment of the parameters for copy number inference using HMM.

When the sample is contaminated with normal components, you need to set the parameters for HMM analysis in order for HMM analysis to work correctly. To do this, choose a region that is considered to have copy number 1 or 3 using the  icon.

#7 Other things

The user may select a log axis or a linear axis, by clicking on the “log” or “ln” icon. The user may also choose an equal-interval horizontal axis or a real-distance axis by choosing the “SNP” or “Base” icon, respectively.

Views may be genome-wide, or of single chromosomes. Moving between chromosomes and changing magnification is possible using the right button.

A region of interest may be measured by selecting the “M” icon, and dragging the mouse over the region of choice. You will be prompted the next action to jump to the UCSC browser. CNAG is now based on NCBI build 35.

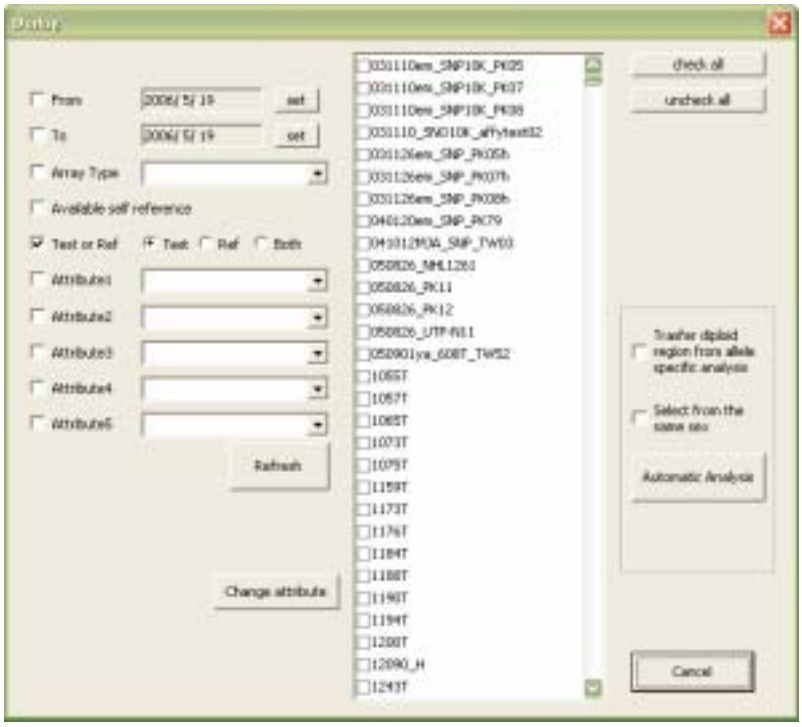
V. Automated analysis

#1. In CNAG2.0, automated analysis is possible. With this option, CNAG performed pair-wise tests for all of the references that exist within the array data folder. As the number of references increases, this takes more and more time, in which case users should avoid processing multiple files at once.

For automated analysis, select “batch analysis” from the “automation” menu, and CNAG displays a dialog box where all the samples in your “array data files” folder are shown in the window. You can apply filters according to the date of experiments, array types, or user-defined attributes. Check any number of samples that you want to analyze, and only checked samples will be analyzed during this automation. To start batch analysis, select “automatic analysis”. Note that when you have a large number of references, this will consume memory and other resources, and the performance of the foreground jobs will be decreased. We recommend that when you analyze a large number of samples, automated analysis should be done during idle time for your PC, for example over night.

During automation, both allele-specific analysis and non-allele specific analysis are performed.

.Automation window:



#2. Changing the attributes in multiple files.

The attributes can be set or changed for multiple files at one time when using this mode. To do this, first choose the samples whose attributes you want to change or set and click on the " change attributes" button. You can change the TEST/REF attribute as well as attributes 1 to 5.

V. Display mode.

The "Display Sample" mode is intended for detailed analysis of multiple results. The user may toggle between samples or display multiple results at one time. The results from multiple samples may be integrated to determine regions effected in multiple samples.

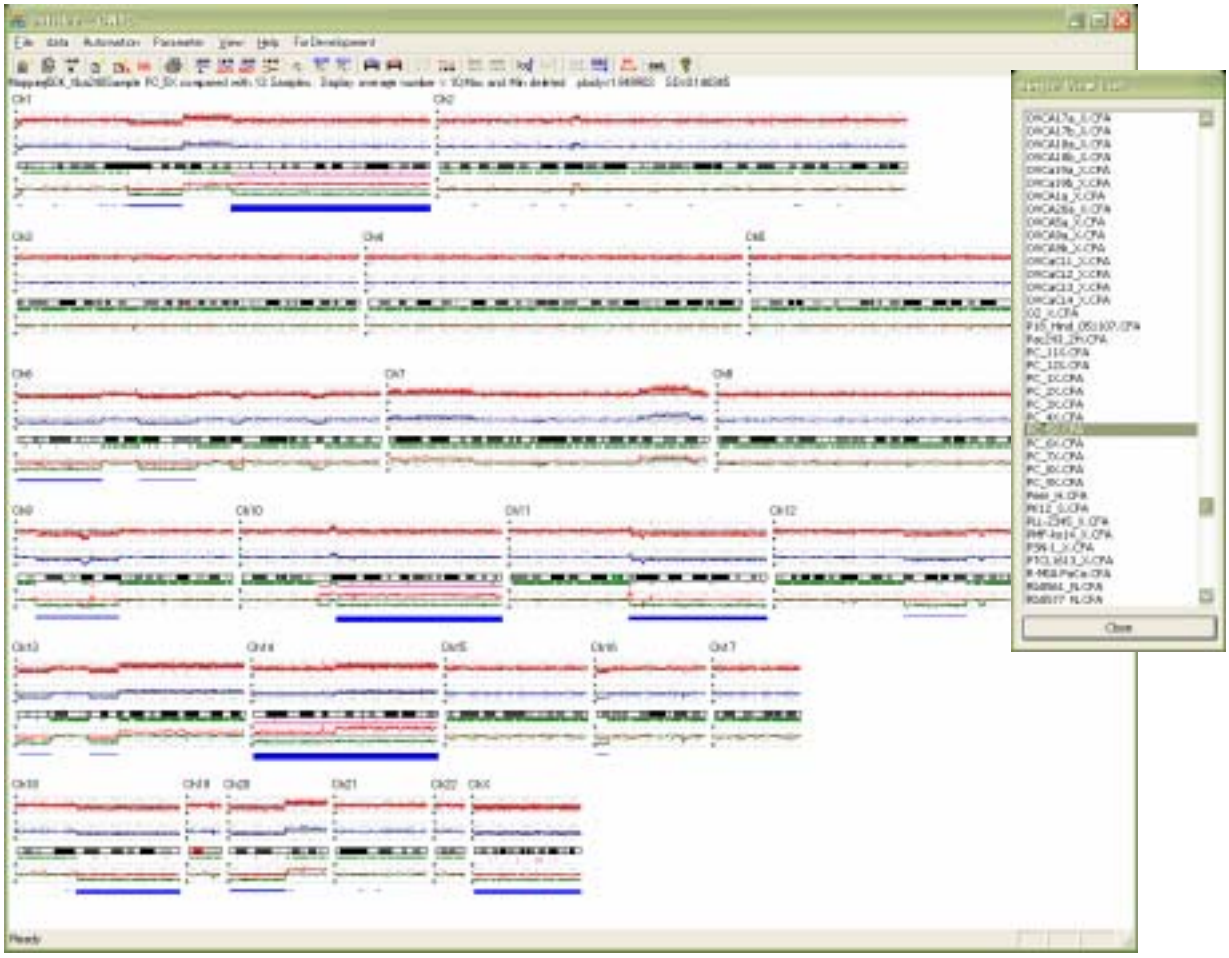
5.1.1. Entering into Display Mode

Select "data" -> "display samples". Initially CNAG displays all of the samples whose result files are stored in the "results" folder specified in the "default setting". You need to specify the type of analysis you want to display, allele-specific or non-allele-specific analysis. You may also apply filters according to the date of experiments, array types,

and other user-defined attributes (attribute 1-5). Check the attributes by which you want to choose samples and if necessary, specify the attributes, and click on “Refresh”. Samples conforming to the specified attributes will be shown in the windows. You may preserve a set of samples selected as a user-defined group, using “Set as a group” option. You can add different sets of samples to the pre-defined group. To do this, select a group in the “Select a group” field. The items of the group will be displayed. Push “Add” button, select samples to be added and push reset the group. After selecting samples, you can toggle the results of these samples by choosing the “Single display” option, or display multiple samples at a time by “Multi display”.

5.1.2 Single display

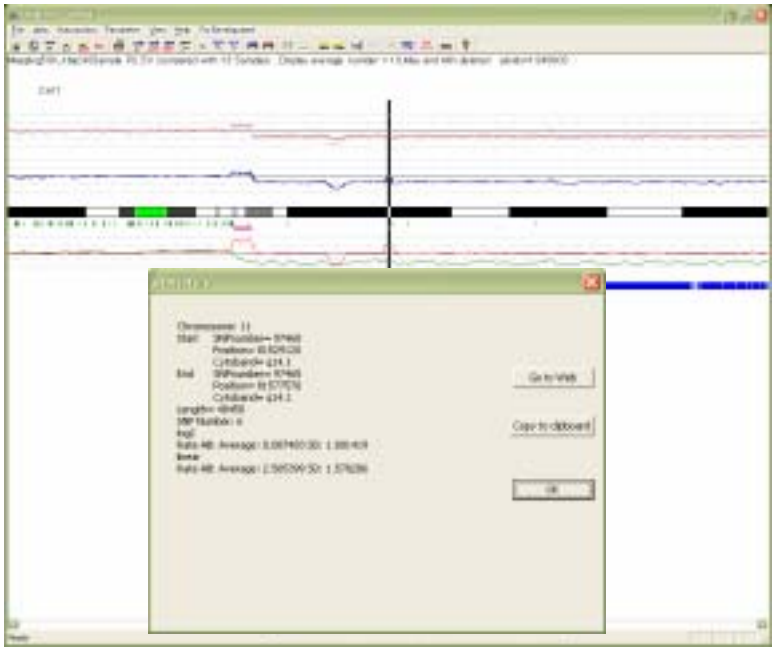
5.1.1With this option, CNAG shows the “Single View List” in a separate window. When you select a sample, the result is displayed in the main window. You can toggle between different samples by clicking on their names in the Single View List.



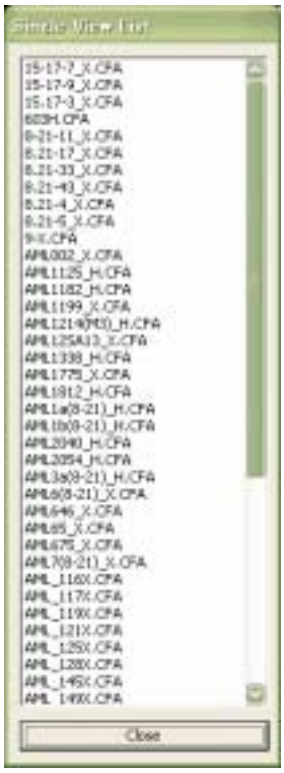
The following is a brief overview of some of the menu bar icons:

#1. (Info) Get information about the current sample
#2. (log, ln) Copy numbers are displayed either log2 ratios or raw ratios
#3. (SNP, BASE) The chromosomal axis of copy number plots is interchangeably scaled either by real distance or by the SNP order using SNP/base icon.
#4. (Edit) Editing HMM calls or marking user defined abnormalities (Edit), which can be summarized in “Integration” function
#5. (ave) The number of SNPs locally averaged and drawing mode
#6. (HMM) Seeing and changing the parameter values of HMM analysis
#7. (M) Specifying a region under interest and getting several measures from the region

The user may select a region of interest for viewing on the UCSC browser. Select this region using the M icon, then click on “Go to Web”:



The user may reset the diploid region within this view. However, note that in this mode, the parameters for quadratic regression are not changed, i.e. they retain values determined by the previous regressions. Only the base line of copy number analysis is changed. To set the parameters for regressions, you need to return to the Sample Manager. This is inconvenient and may be improved in the future version. Items to be displayed with the Single View option can be specified in the “default setting” from data menu.



5.1.3. Select items displayed in the Single view mode

A number of items can be displayed in the single view but users may select only some of them for simplicity. For example, CNAG2.0 supports allele-specific copy number inference using HMM, but the user may choose to suppress this graph by unchecking this.

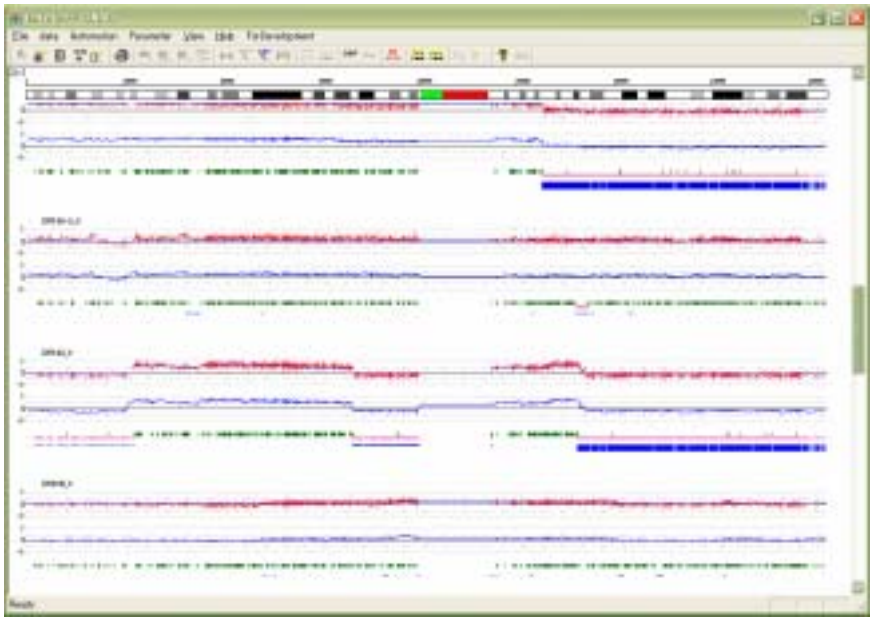
5.1.4. Combining different array data

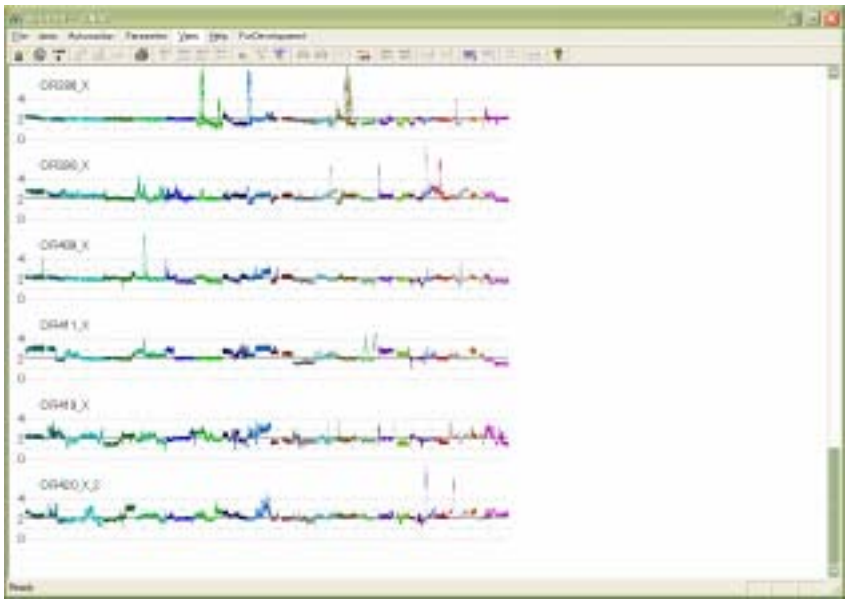
Different array data, for example Xba and Hind 50K data, can be combined in this mode. To do this, select the two array data for the same sample to be combined and click on “combine”. CNAG creates a new file named (sample name_array1)_and_(sample

name_xba), which now appears in the sample window.

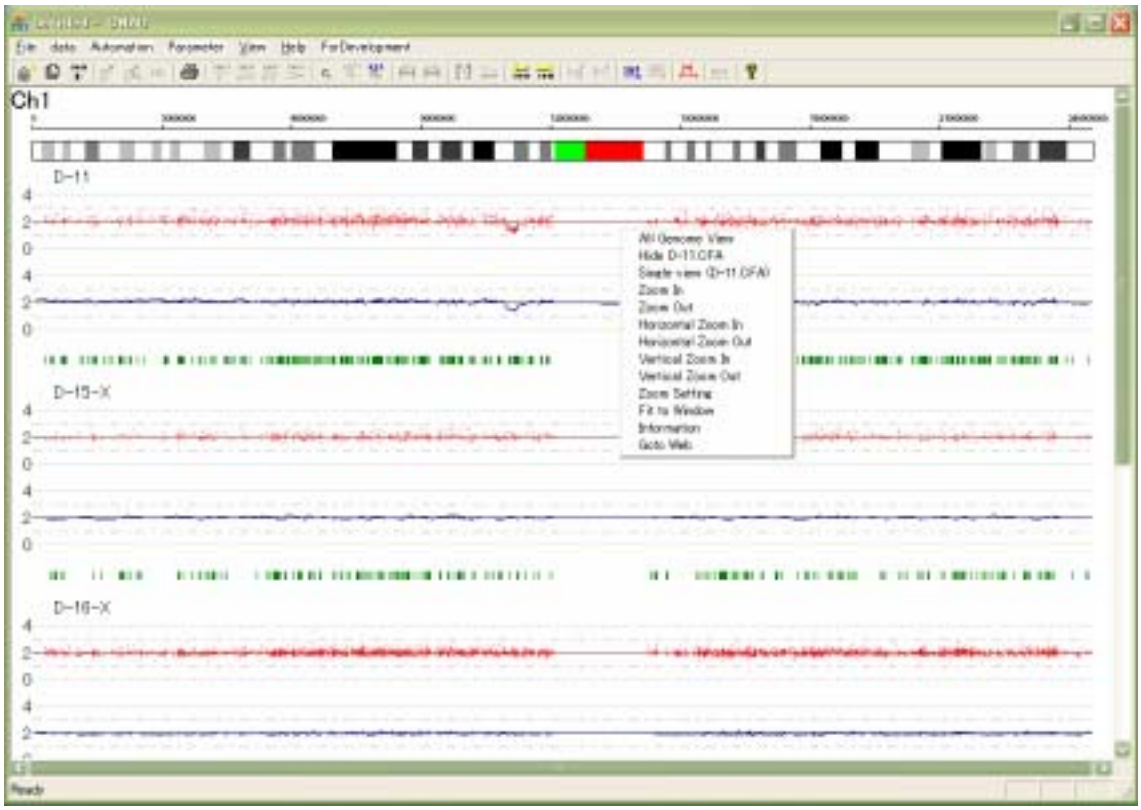
5.2 Multiple Display

5.2.1 Multiple results can be viewed within a single window. The items to be displayed in this mode are freely chosen in the default setting menu. Choices for display include the scale bar, the cytoband at the top or above each result, SNP call bars, copy number graph, copy number inference from HMM, real copy number plot, allele-specific copy number, LOH inference, and color coded copy number and LOH. These features are also viewed in both single and multiple displays.





5.2.2. Moving to Single view

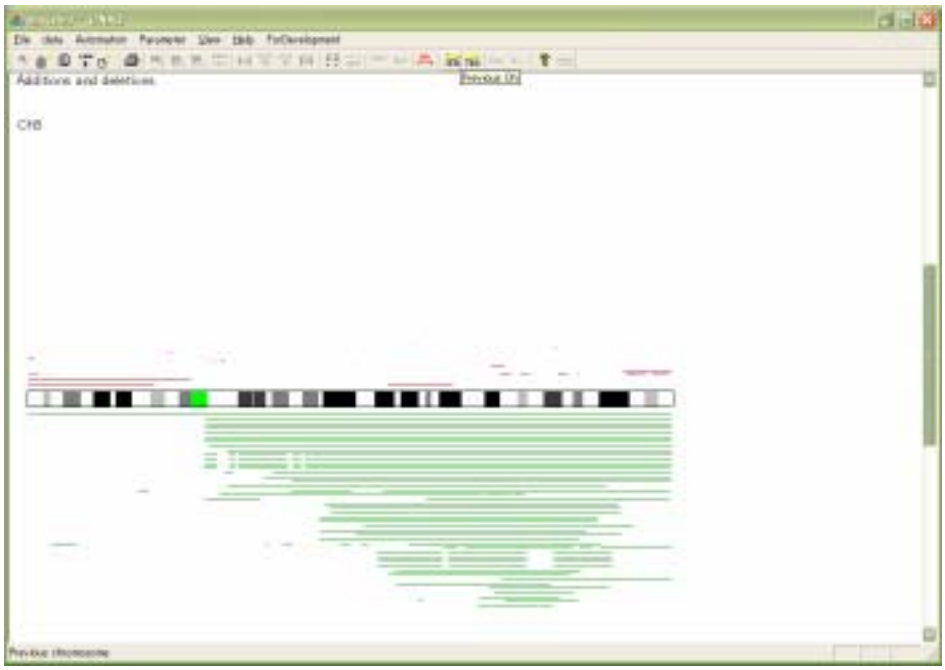


#4. The order of samples can be changed by dragging individual samples. The user can right-click on the image for options including hiding one or more samples, defining the currently selected set as a new group, zooming in or out, etc.

5.3 Integration of multiple results

CNAG can report the integrated results for specific features found in multiple samples. It summarizes and graphically reports regions showing gains and losses, homozygous deletions, high grade amplifications, LOH, allelic imbalances, and chromosomal breakpoints of unbalanced translocations.

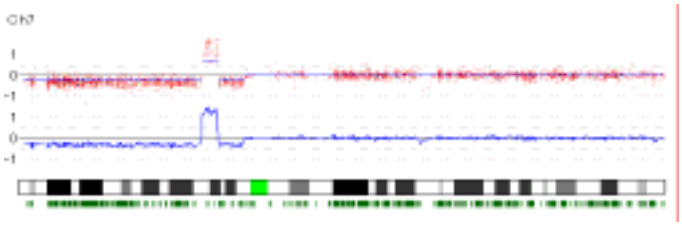
Below is an example of a summary reporting chromosomal gains and losses found in 150 MDS cases in chromosome 5.



IV Other new features

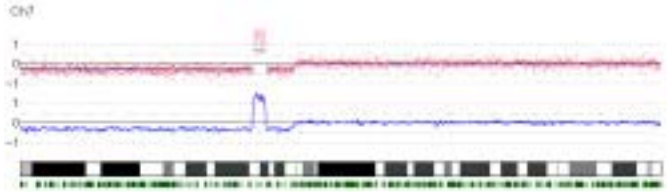
6.1 Display results in true scale versus in SNP order

The chromosomal axis of copy number plots is interchangeably scaled either by real distance or by the SNP order using SNP/base icon.



1) True scale

2) SNP order



In true scale display, you can see SNP probes are less densely distributed in some parts of chromosomes than in others.

6.2 Display relationship between adjacent restriction fragments

6.3 Color code copy number and LOH results

6.4 Export raw data out put and IGB files

Users can export the raw data to a tab limited text file. The following is the short explanation of the columns of the output file. (The columns are varied according to the analysis mode)

Call_reference, Call_tests

are the genotyping calls for ref and test 1:AA, 2:AB, 3:BB, 0:NoCall

N_AB, N_A,N_B

are the HMM calls for both alleles and allele specific analyses.

log2ratio_AB, log2ratio_A , log2ratio_B

are the raw log2ratio data.

6.5 Filter SNPs based on fragment size