Agilent G4175AA
CGH Analytics 3.4

User Guide

Agilent Technologies

Notices

Edition

First Edition, July 2006

Printed in USA

Agilent Technologies, Inc.
3501 Stevens Creek Blvd.
Santa Clara, CA 95051 USA

Software Revision

This guide is valid for the Agilent G4175AA CGH Analytics 3.4 software.

Warranty

The material contained in this document is provided "as is," and is subject to being changed, without notice, in future editions. Further, to the maximum extent permitted by applicable law, Agilent disclaims all warranties, either express or implied, with regard to this manual and any information contained herein, including but not limited to the implied warranties of merchantability and fitness for a particular purpose. Agilent shall not be liable for errors or for incidental or consequential damages in connection with the furnishing, use, or performance of this document or of any information contained herein. Should Agilent and the user have a separate written agreement with warranty terms covering the material in this document that conflict with these terms, the warranty terms in the separate agreement shall control.

Technology Licenses

The hardware and/or software described in this document are furnished under a license and may be used or copied only in accordance with the terms of such license.

Restricted Rights Legend

U.S. Government Restricted Rights. Software and technical data rights granted to the federal government include only those rights customarily provided to end user customers. Agilent provides this customary commercial license in Software and technical data pursuant to FAR 12.211 (Technical Data) and 12.212 (Computer Software) and, for the Department of Defense, DFARS 252.227-7015 (Technical Data - Commercial Items) and DFARS 227.7202-3 (Rights in Commercial Computer Software or Computer Software Documentation).

Safety Notices

CAUTION

A **CAUTION** notice denotes a hazard. It calls attention to an operating procedure, practice, or the like that, if not correctly performed or adhered to, could result in damage to the product or loss of important data. Do not proceed beyond a **CAUTION** notice until the indicated conditions are fully understood and met.

WARNING

A WARNING notice denotes a hazard. It calls attention to an operating procedure, practice, or the like that, if not correctly performed or adhered to, could result in personal injury or death. Do not proceed beyond a WARNING notice until the indicated conditions are fully understood and met.

In This Guide...

This guide contains information to use the CGH Analytics 3.4 program.

1 How-To

This chapter provides brief instructions on how to accomplish necessary tasks and procedures for using CGH Analytics and its features quickly and efficiently.

2 Main Windows Reference

In this chapter, each window and pane is described with information on when, why and how to use it.

3 Dialog Box Reference

In this chapter each dialog box is described with information on why and how to use it.

4 Statistical Algorithms

This chapter discusses several algorithms that can be used to facilitate the statistical analysis and calibration of aberrant regions.

5 Reference

This chapter provides additional information on several topics that will assist you in using Agilent CGH Analytics.

What's New in CGH Analytics 3.4 Software?

- New circular binary segmentation (CBS) copy number change algorithm
- New fuzzy zero correction algorithm
- Agilent multi-array format support
- Common aberration visualization
- Spike-in support for reference DNA
- Visual recommendation of QC thresholds
- Virtual design creation functionality via array set fusing
- Universal data file (UDF) import

Contents

1 How-To

This chapter provides brief procedural instructions on how to accomplish tasks common to everyday use of CGH Analytics 3.4. For more detailed information, please refer to the appropriate sections in Chapter 2, “Main Windows Reference”, and Chapter 3, “Dialog Box Reference”.

Importing Data

CGH Analytics supports Agilent CGH XML based array design file, Feature Extraction (FE) file, and vendor-neutral Universal Data File (UDF) import in addition to various other types of sample or genome annotation files. Imported data files are presented for analysis options within the **Navigator view** for experiment selection and attribute modification.

To import data files

1. In the Navigator, right-click the **Data folder**. The **Import** submenu displays three types of files that you can import.
   - **Design Files**
   - **FE Files**
   - **UDF Files**
2. Click the type of file that you want to import to display a corresponding Import *Type* Files dialog box.
3. Complete the dialog box, and click **OK** to access the folder containing the data files.
4. Select the file or files to import, and click **OK**.

See "File Menu" on page 60 for other ways to import array files and design files.

To import design files

Design files are Gene Expression Markup Language (GEML) compliant metadata files. Array specific information such as probe identification, descriptors, sequence, and genomic location is loaded into CGH Analytics when a design file is imported.

1 In the **Navigator**, right-click the **Data** folder, then click **Import > Design File**. The Import Design Files dialog box appears. See Figure 3-1 on page 93.

2 Complete the dialog box, and click **Open** to access the folder containing the data files.

3 Click **Open**.

Alternatively,

1 In the **Menu Bar**, click **File > Import > Design File**. The Import Design Files dialog box appears. See Figure 3-1 on page 93.

2 Complete the dialog box, and click **Open** to open the folder containing the data files.

3 Click **Open**.

You may also download design file updates directly from Agilent eArray:

1 In the **Menu Bar**, click **Tools > User Preferences.** The Preferences dialog box opens. Select the Miscellaneous tab. See "Preferences dialog box displaying Miscellaneous tab options" on page 169.

2 In the eArray User Details area, enter your account information.

3 Click **Apply** to save the changes.

4 Click **OK** to proceed.

5 In the **Navigator**, within the **Data node**, right-click on any design node. A shortcut menu opens.

6 Select **Update From eArray**.

To import experiments

CGH Analytics experiments can be exported from one computer and imported into another as special .zip files.

1 In the **Menu Bar**, click **File > Import > Experiments**. The Import Experiments dialog box appears. See Figure 3-2 on page 94.

2 Complete the dialog box, and click **Open** to open the folder containing the experiment .zip files.

3 Click **Open**.

To import Feature Extraction files

Agilent CGH and gene expression arrays processed with the Feature Extraction (FE) software can be imported directly into CGH Analytics for analysis. FE files contain the processed intensities from Cy3 and Cy5 fluorochrome channels as well as array-specific metadata.

1 In the **Navigator**, right-click the **Data** folder, then click **Import > FE File**. The Import FE Files dialog box appears. See Figure 3-3 on page 95.

2 Complete the dialog box, and click **Open** to access the folder containing the data files.

3 Click **Open**.

Alternatively,

1 In the **Menu Bar**, click **File > Import > Array Files > FE Files**. The Import FE Files dialog box appears. See Figure 3-3 on page 95.

2 Complete the dialog box, and click **Open** to access the folder containing the data files.

3 Click **Open**.

To import filters

Aberration, array level or feature level filters created in CGH Analytics can be exported from one computer and imported into another as XML files. The imported filter(s) can be found in the corresponding filter option in the **Menu Bar** under **Tools > *filter type* Filters > Apply**.

1 In the **Menu Bar**, click **File > Import > Filters**. The Import Experiments dialog box appears. See Figure 3-4 on page 96.

2 Complete the dialog box, and click **Open** to access the folder containing the data files.

3 Click **Open**.

To import CGH Analytics 3.1 files

CGH Analytics is backwards compatible with Experiments created in CGH Analytics version 3.1.

1 In the **Menu Bar**, click **File > Import > CGHAnalytics 3.1 Experiment(s)**. The Import dialog box appears. See Figure 3-5 on page 97.

2 Complete the dialog box, and click **Open** to access the folder containing the data files.

3 Click **Open**.

To import genome builds

In general, the genome build specified in the array design file will be used and that genome build will be protected from changes. If a genome build is not available in the design file, you may import a genome build.

NOTE If an experiment consists of multiple array designs that use different genome builds, you must choose one genome build for the entire experiment. See "To resolve genome builds" on page 27.

1. In the **Menu Bar**, click **File** > **Import** > **Genome Build**. The Import Genome Build dialog box appears. See Figure 3-6 on page 98.
2. Complete the dialog box.
3. Click **OK**.

To import microarray attributes

1. In the **Menu Bar**, click **File** > **Import** > **ArrayAttributes**. The Import microarray attributes dialog box appears. See Figure 3-7 on page 99.

NOTE The microarray attributes file format must follow the specifications outlined in "Array Files" on page 61.

2. Complete the dialog box.
3. Click **OK**.

To import universal data files (UDFs)

You can import tab delimited aCGH data as a new custom-formatted array data file. The order of your tabular data must follow the column headers in the **UniversalData Importer - Map column headers** dialog box. See Figure 3-12 on page 104.

1. In the **Menu** bar, click **Files > Array Files > UDF Files**. The UDF Files dialog box appears. See Figure 3-8 on page 100.
2. Select the file name of the UDF file. **The Files of This Type** filter displays files containing .txt as a suffix.
3. Click **Open**. The Select Data Type for Experiments dialog box opens and presents four options for the data type (See Figure 3-9 on page 101):
   - **ratio**
   - **log2 ratio**
   - **log10 ratio**
   - **ln ratio**
4. Select **log2 ratio** for Agilent CGH arrays or the appropriate ratio for the data to be imported.
5. Select the design type.
6. Click **Continue**. The Universal Data Importer - Map column headers dialog box appears. See Figure 3-12 on page 104.
7. In the **Species Information** area**,** select the correct **Genome Build** and **Species** for your experiment.
8. In the **Mapping Info** area, choose **Select Mapping** to label this mapping as **Custom** or a choose **Save Mapping As** to label the mapping specification with a specific name.
9. Go to the table and **Select** a header from those provided for each column. If the data is from Agilent CGH array(s), the correct header entry will be displayed on the top row of the table.

NOTE

These headers are mandatory and must be labeled in specific order from left to right with the six specific headers provided:

- The first column (left most column) header must be: ProbeName
- The second column header must be: ChrName (chromosome)
- The third column header must be: Start
- The fourth column header must be: Stop
- The fifth column header must be: Description
- The sixth column (right most column) header must be: LogRatio

One or more columns following the LogRatio column may contain signal values for each of the specific arrays.

10 Click **Import** to import the UDF files. A dialog box opens:

- Click **Yes** if you want to do a bulk or batch import of additional UDF files. Repeat steps 1-9 for each additional UDF file.
- Click **No** to proceed with UDF file import.

11 The UDF Import Summary dialog box appears. See Figure 3-11 on page 103.

12 Click **OK**. A new **Data** node appears in the Navigator for the UDF file(s). You can now use this data file for experiments.

To include Agilent multi-pack array

Agilent provides 2-pack, 4-pack, and 8-pack multi-pack array formats. You can include Agilent multi-pack arrays in your experiment in the same way as single-pack arrays (see "To import data files" on page 12). Each Agilent array within a multi-pack generates an individual FE file. Within the Microarray Properties dialog box, new microarray attributes are available for these samples.

1. In the **Navigator**, click on an Experiment to expand it.
2. Click on a Design to show available array(s).
3. Right click on an array.
4. Select **Show Properties** in the submenu. The Microarray Properties dialog box appears. See Figure 3-15 on page 108.
5. Check the **Attribute** list for:
   - Is Multi-pack = TRUE
   - Grid Position = the Bar Code followed by _x_y. (Where the 'x' designates the array row in the grid and the 'y' designates the array column.).
6. Click **Close**.

NOTE

Another indicator that the experiment has a multi-pack array set attached is the small grid icon next to the '**CGH**' in the experiment name. See Figure 2-2 on page 50.

Experiment Creation and Modification

Analysis within the CGH Analytics program is based on the concept of an Experiment. An Experiment is accessible within the **Navigator view** as a node which allows collection, organization, and persistence of multiple instances of analysis information and attributes for efficient data mining and exploration.

To activate an experiment

The CGH Analysis program proceeds from the data and attributes specified in one Experiment at a time. To load the Experiment into the application's memory for analysis, the Experiment node in the **Navigator** must be active.

1. Go to the **Navigator View** of the **Main Window**.
2. In the Navigator, double-click an experiment to display a query asking you to confirm or cancel your decision to activate that experiment.
3. Alternatively, you can right-click the experiment, then click **Select Experiment** to elicit the same query.
4. Click **Yes**. The Experiment node name text will turn blue upon activation. See Figure 2-2 on page 50.

To combine replicates

Replicates within an array and also in between multiple arrays can be combined by to increase the statistical power of your experiments.

1 In the Toolbar's Combine Replicates tool box, select whether to combine replicates as **Intra Array** or **Inter Array** replicates.

NOTE For expression arrays, the Intra Array replicate methodology uses the genes to combine replicates. For CGH arrays, the probe name is used to combine replicates.

2 From the **Group By** list box, select the attribute to use for grouping the arrays to perform Inter Array combining of replicates.

3 Click **Go**. See "Combine Replicates" on page 85

To create and modify filters

You can specify aberration, array, and feature conditions within a given experiment. For example, you can filter using attributes from the selected experiment, or filter by features using data from the FE files.

1 In the **Menu Bar,** click **Tools.** The submenu displays three types of files that you can create:

- **Aberration Filters**
- **Array Level Filters**
- **Feature Level Filters**

2 Select ***filter type*** **> Edit Filter.** The Edit *filter type* Filters dialog box appears. See Figure 3-26 on page 121.

3 Click **New**. An Input text box appears.

4 Type a name for your new filter and click **Close**.

5 You are returned to the Edit *filter type* Filters dialog box with the name of your new filter displayed along with default values for the four parameters.

6 Accept the displayed default values or change the parameters appropriately.

7 If you change any values, click **Update** to apply the changes.

NOTE If you want to return to the previous parameters after updating, you must re-type the values manually.

8 If you want to return your changed values to the unchanged values before you update, click **Reset**.

9 Click **Close**. A Confirm filter save dialog box appears, and you will be prompted **Save changes to filter?**

10 Click **Yes** and the new filter will be listed as one of your available filters.

11 See also, "To import filters" on page 15.

To create experiments

Before analyzing CGH data, log ratio values across array(s) must be associated with array attributes into a working unit called an Experiment. The Experiment allows each instance of analysis to be stored in the CGH Analysis application.

1 In the Navigator, right-click the Experiments folder and select **New Experiment** to open the Create Experiment dialog box. See Figure 3-14 on page 107.

2 Type in a **Name** and **Description** for the experiment.

3 *Do not click* **OK** *at this point***.** The experiment does not contain data, and you will have to populate this experiment before you can use it.

4 Click the **Properties** button to display the Experiment Properties dialog box. See Figure 3-34 on page 131.

5 From the list of **Arrays**, select and move arrays you want in your experiment to the **Selected Arrays** list using the arrows.

   You can import additional CGH 3.1 and FE files by clicking the corresponding buttons at the bottom.

6 Click **OK**.

NOTE You can add arrays to an experiment at any time by dragging and dropping an array from the data node to a selected experiment. See "To populate experiments" on page 27.

To define attributes

Attributes are basic items of array-specific information that you would record about particular arrays, such as hybridization time (hyb time), hybridization date, or sample type.

1 In the **Menu** bar, click **Tools > Attribute** to display the Attributes dialog box.

2 The characteristics of a selected attribute are displayed where you can

   a change values

   b delete value

   c create new attributes

3 Click **Close**.

To explore or update design files

1 In the Data node, right-click a design file. A shortcut menu appears with three options.

- **QC Metrics.** Click to display a QC Metrics table dialog box. See Figure 3-49 on page 150.
- **Update from eArray.** Click to update the design from the eArray site.
- **Delete.** Click to delete the design.

2 Select an option and proceed.

To fuse arrays

You can combine, or fuse, two different design files to create a new virtual design file that enables you to analyze data in two individual arrays or in the entire genome CGH array.

When fusing arrays, the CGH Analytics program has the following restrictions:

- Arrays from the same design cannot be fused.
- CGH Analytics arrays and expression design arrays cannot be fused.
- If the design is the result of a fused array, then it cannot be fused again.
- If the designs to be fused have common probes then they will appear as replicates in the fused array. You can combine these probes by selecting **Intra Array** in the combine section of the **Toolbar**.

You can overwrite a fused design manually by fusing the same set of designs again. This function is useful if the designs are changed or upgraded.

NOTE If you wish to fuse large quantities of files, you can import an array attributes file. See "To import microarray attributes" on page 16.

1 In the **Navigator**, create a **New Experiment** node. See "To create experiments" on page 23.

2 Select at least two array files each with a different design and drag these files to the newly created experiment node.

NOTE The samples to be fused must be aliquots from the same prep. Preferably, they should be labeled and hybridized together under the same conditions.

3 Select the newly created Experiment node (it turns blue onscreen).

4 Right click on one array from the array set. A shortcut menu opens.

5 Select **Show Properties**. The Microarray Properties dialog box appears. See Figure 3-15 on page 108.

6 In the **Attribute** list, go to **ArraySet** attribute. Type in an identifier for the selected array.

7 Click **Close**.

8 Select the second array to be fused in the **Fuse Experiment** node.

9 Right click on this array to open a shortcut menu.

10 Select **Show Properties**. The Microarray Properties dialog box appears. See Figure 3-15 on page 108.

11 In the **Attribute** list, go to **ArraySet** attribute and again type in the same identifier for the selected array. By typing in the same identifier, the two arrays are partnered for fusing in the CGH Analytics program.

12 Click **Close**.

13 Activate the newly created Experiment node. See "To activate an experiment" on page 20. The array data for this Experiment will be combined into one file.

14 Go to the **Combine** section of the **Toolbar**. See Figure 2-20 on page 85.

15 Click on the **Fuse** button. The Array Set: Fuse Experiment dialog box appears. See Figure 3-13 on page 106,

NOTE Check to ensure that each ArraySet attribute has the same identification number for all the array data. This ID number joins the arrays in the fuse process.

16 Click **Continue**. A data table appears in the Tab View.

17 A new **Design Node** is created under the newly created fused **Experiment Node** when the array fuse process has completed.

18 In the **Toolbar > Combine area > Replicates area**, select Inter Array to combine the common probes between two designs.

To modify array files

1. In the **Navigator**, right-click an array file under a parent design in the **Data Node**.
2. A shortcut menu appears with four options.
   - **Show Properties.** Click to display the Microarray Properties dialog box with tabs for **Attribute**, **FE Headers**, and **FE Features**. See Figure 3-16 on page 109.
   - **QC Metrics.** Click to display a QC Metrics table dialog box. See Figure 3-49 on page 150.
   - **Rename.** Click to display an Input dialog box for renaming an array file.
   - **Delete.** Click to delete the array.
3. Select an option and proceed.

To populate experiments

If you did not populate your new experiment when you created it, you can easily add samples to the experiment node. You may add or remove samples from the experiment at any time.

1. Go to the **Navigator** and click the **Data** folder to expand it and display the designs it contains.
2. Click a design that contains microarrays to expand it.
3. Select one or more arrays, and drag and drop the files onto the folder of the new experiment you created.

To resolve genome builds

If an experiment consists of multiple array designs that utilize different genome builds, you must select one appropriate genome build to represent the entire experiment.

1. In the **Navigator**, right click on the active Experiment Node which contains multiple array designs. A dropdown menu appears.
2. Select **Change Genome Build.** A dialog box appears.
3. Select the genome build that is correct for your experiment from those available in the dropdown list.
4. Click **OK**.

To select arrays

The first array is selected for analysis within an experiment by default. You can specify which array(s) to include in an experiment analysis.

1. Click an experiment in the **Navigator > Experiment** node list.
2. Click the array design folder to expand the list of available arrays.
3. Click the array you want to include in the analysis. To choose more than one array, press **Cntrl+click**.
4. Right-click on the selected array(s).
5. A pop-up menu appears. Choose **Select**.
6. Subsequent analysis steps will derive from the selected array(s).

To select experiments

See "To activate an experiment" on page 20.

Analysis and Visualization

The visual display of information in CGH Analytics is a multi pane window which allows simultaneous course and fine grain data exploration.

To apply centralization

The centralization algorithm re-centers the log ratio values by finding a constant value to add to or subtract from all log ratio measurements, ensuring that the zero-point reflects the most-common-ploidy state. Centralization is turned on by default in CGH Analytics.

1 In the **Menu Bar,** select **Tools > User Preferences**. The preferences dialog box opens.

2 Select the **Miscellaneous** tab. See "Preferences dialog box displaying Miscellaneous tab options" on page 169.

3 In the Centralization area, toggle application of the function by selecting **Apply**. Change Threshold and Bin Size parameters if necessary. See "Centralization Algorithm" on page 193 for information about the parameters available in the centralization algorithm.

4 Click **OK** to accept the changes and apply or remove centralization.

You can quickly check the status of centralization by looking at the **Status Bar**. If the fifth column displays a bold-face '**C**', centralization is applied. The 'C' is gray if centralization is not currently applied. See "Status View" on page 90.

NOTE

Depending on the density of the array, centralization processing may take up to 1 minute per array. This can lead to extensive processing time and should be planned for or controlled accordingly. For example, if you have an experiment which has 50 arrays with centralization turned on, the first time you run the application it may take 50 minute.

To apply fuzzy zero correction

ADM-1 and ADM-2 scores may identify extended aberrant segments with low absolute mean ratios. Often such aberrations represent noise, and are detected because of high number of probes in the region. If long, low aberrations are detected in your analysis, you can apply the fuzzy zero algorithm to correct for the reliance on segment probe number. Fuzzy Zero is turned off by default in CGH Analytics. "Fuzzy Zero" on page 209 for more information.

1. In the **Menu Bar,** select **Tools > User Preferences**. The preferences dialog box opens.
2. Select the **Miscellaneous** tab. See "Preferences dialog box displaying Miscellaneous tab options" on page 169.
3. In the Fuzzy Zero area, toggle application of the function by selecting **Turn On**.
4. Click **OK** to accept the changes and apply or remove Fuzzy Zero correction.

You can quickly check the status of Fuzzy Zero by looking at the **Status Bar**. If the fifth column displays a bold-face '**F**', Fuzzy Zero is applied. The 'F' is gray if Fuzzy Zero is not currently applied. See "Status View" on page 90.

To assess sample quality via spike-in references

CGH Analytics allows inspection of the performance of an external DNA reference standard present in your sample(s) to sample quality through the CGH Analytics program.

You can visualize spike-in reference DNA performance through expected vs. observed value charts and fold change vs. spike-in ratios by the following procedure:

1. In the **Navigator**, activate an **Experiment** that has samples with an external DNA reference. See "To activate an experiment" on page 20.
2. Right-click an the activated experiment file or on any sample(s) in the activated experiment with an external DNA reference. A shortcut menu opens.
3. Select **Show Spike-in.** The spike-in view opens in the **Main Window.**

Alternatively:

4 In the **Menu Bar**, select **View > Spike In**.

5 Select the graph option (Fold Change vs. Spike-in Ratios or Expected vs. Observed) that plots the expected outcome of the DNA reference and the outcome of the array(s) you selected in a new dialog box. See Figure 3-56 on page 158.

6 To change the parameters of the graph for visualization, right-click on an individual graph to open the Chart Properties dialog box.

You can visualize median $\log_2$ ratio value charts sorted by array by the following procedure:

1 Go to the Navigator's **Experiment** node, right-click on an experiment that has an external DNA reference added.

2 Select the **Show Spike in ratios** option. The Spike In Plots dialog box opens. See Figure 3-58 on page 160.

3 If you right click on an individual graph, a Properties menu appears that provides for **Save As**, **Print, Zoom In** and **Zoom Out**. The **Auto Range** function allows you to change the parameters of the graph for viewing.

The Spike In Ratios graph for each reference spike ratio is presented in a series of graphs with the median log2 value (y axis) given for each array (x axis) in the experiment with an external DNA reference. Ideally, the **Show Spike-in ratios...** graph(s) will appear as a horizontal straight line. However, experimental systems will vary based on the array data.

NOTE

Expected vs. Observed Spike-In plot correlation, Intercept, and slope data are available as QC metrics and are a useful way to incorporate Spike-In reference data to your analysis workflow. See "QC Metrics Table" on page 150 for more information.

To change display orientation

By default, the Main screen is displayed with a vertical orientation with Navigator, Genome, Chromosome, and Gene views displayed side-by-side.

To change the display to a horizontal orientation with three of the views (Gene, Chromosome, and Genome) placed below one another:

1. In the **Menu Bar**, click **View > Orientation.**
2. Mark the **Horizontal** check box.
3. To return to the vertical orientation, repeat the procedure and mark the **Vertical** check box.

To change visualization layout

You can modify the data displayed and the way it is displayed by changing the information in the Preferences dialog box.

1. Select an experiment.
2. In the Menu bar, click **Tools > User Preferences** to display the Preferences dialog box.
3. Click the **View** tab to verify or change
   - Orientation
   - Rendering Style
   - Data Visibility
   - Rendering Patterns.
4. Click the **Gene Symbols** tab to verify or change
   - Show Gene Symbols
   - Font
   - Font Style
   - Font Size
   - Font Orientation.
5. Click the **Miscellaneous** tab to verify or change
   - eArray User Details
   - Error Model

- Application of Fuzzy Zero
- Data Location
- Centralization
- Scale.

6 Click the **License** tab to verify or change
- Server Location
- Text License

7 Click **Apply > OK**.

To create common aberration summaries

You can test for statistically significant aberrations shared in common genomic intervals between two or more samples. Once created, common aberration summaries are added to the Experiment data node under the parent design in the **Navigator**. The resultant common aberration summary node is used as the basis for common aberration report summaries.

NOTE Common aberration analysis can be applied to two or more arrays. You must first analyze each of the arrays to be used in creating a common aberration summary using any of the available aberration algorithms. See Figure 2-19 on page 84.

1. Select the arrays to be used in the common aberration analysis. See "To select arrays" on page 28.
2. In the **Menu**, click **Reports > Common Aberration** to display the Common Aberration dialog box. See Figure 3-21 on page 114.
3. From the list of available algorithms, select the common aberration calling method. See "Common Aberration Analysis" on page 201 for more information on available common aberration algorithms.
4. Change the input parameters if necessary.
5. Select the scope of the common aberration analysis. **Genome Scope** is selected by default and will include all chromosomes.
6. Enter a reference name for the new common aberration node to be created in the **Navigator** and press **OK** to proceed.
7. The Navigator will update the node list with **Common Aberrations**. See Figure 2-2 on page 50.

To create gene lists

1. Select an area of interest in the **Chromosome View**.
2. Right-click anywhere in the **Gene View** area.
3. From the shortcut menu, click **Create Gene List** to display the Create Genelist dialog box.
4. Type a **Name** and **Description** for the new gene list.
5. Mark the **User Defined** radio button.
6. Click **OK**.

To create graphical common aberration filters

Graphical common aberration summaries can be filtered by genomic interval to focus on or exclude chromosomal regions.

1. From the Graphical Common Aberration Summary pane (see "To visualize common aberrations" on page 39), click **Create Filter**. The Interval Filter dialog box appears. See Figure 3-41 on page 141.
2. Click **New**. An Input text box appears.
3. Type a name for your new filter and click **OK**.
4. Complete the Interval Filter dialog box by assigning attributes, operators, values, and logical expressions to match intervals. Click **New Condition** to add another set of criteria.
5. Application of an interval filter reflects intervals which pass a boolean conditional across all criteria. Select to include or exclude such intervals.
6. Click **Update** to apply the conditions to the named filter.
7. Click **Close** to return to the Graphical Common Aberration Summary pane.
8. Check **Apply Filter** to visualize results.

To generate gene lists for common aberrations

You can generate a list of genes contained within a genomic interval which share a common aberration across samples.

1. In the Graphical Common Aberration Summary dialog box tab pane, choose the genomic interval to display the Common Aberrations for a given chromosome. See Figure 3-38 on page 136.
2. Click **Create Gene List** to display the Common Aberration Create Gene List dialog box.
3. Enter a name and description for the new gene list.
4. Change the gene list node color code if necessary to distinguish from other gene lists.
5. Click **OK** to proceed. A new gene list node is created in the active Experiment folder in the Navigator.

To isolate and visualize data

You can isolate and enlarge an individual view to help you better visualize the data it displays. Regardless of orientation, there is a toggle button at the top and center of each view—**Navigator**, **Genome**, **Chromosome**, **Gene**, and **Tab** views. The toggle button resembles the lower edge of a “hidden” button.

1. Click the toggle button at the top of the view you want to enlarge. That view is displayed as an isolated window of the same size.
2. Click the **Maximize** button in the center of the three buttons in the upper right corner of the isolated window.

   The size of the window is maximized to a full screen with a corresponding enlargement of the displayed data. To return to the standard view, click the **Close** button (X) in the upper right corner of the maximized view.

To view microarrays

You can view individual columns of array data or multiple columns of data simultaneously.

1. In **Tab** view, click in the column heading to select a single array for viewing See Figure 2-21 on page 87.
2. To view multiple columns of array data, press the **Ctrl** key while clicking on each desired column. (It may take a few seconds before you see the results of your selection.)
3. To cancel the selection of any column, click on it a second time.

Scatter Plots have limited utility when visualizing multiple arrays. A better metric for visualization is to use the **Moving Average** display.

NOTE

On Apple platforms, you must hold down the **Shift** key, rather than the **Control** key while clicking column headers. **Control** + click is used on Macintosh platforms for launching context (pop-up) menus.

Performance Tip: You can compute all statistics for a set of parameters at one time by right-clicking the data column header, and clicking **Select All Arrays**. You can change the time necessary for the operation by selecting or deselecting the arrays.

To view QC metrics

Color-coding the QC Metrics thresholds provides a quick assessment of the quality of your experimental results. The QC Metrics thresholds in data table Appendix B of this User's Manual are grouped by color and are rated as Yellow=Excellent, Turquoise=Good, and Pink=Poor.

1 In the **Menu bar**, click **Tools > QC Metric** to display the QC Metrics table dialog box (see "QC Metrics Graph" on page 147). This dialog box displays quality control outcomes for your experimental results.

Alternatively,

1 In the **Navigator**, in the **Data** folder, right-click on a design node or a sample node. You may also right-click on an active **Experiment** node, or any array(s) listed in the active experiment. A shortcut menu opens.

2 Select **QC Metrics**.

3 You can display the data based on selected attribute that appears in the **Group by** list box in the lower left corner of the QC Metrics table dialog box.

4 Mark one or more array's check boxes (**DLR Spread**, **SignalTo**, **SignalInt**, etc.). If your array(s) contain external spike-in reference DNA, you can apply measures of the performance of these spike-ins to your QC analysis. See "QC Metrics Table" on page 150.

5 Click the appropriate button to view the data's QC Metrics: **Frequency Distribution** or **Plot** (see "QC Frequency Distribution" on page 146).

6 Click **Close** when finished viewing.

For more information on metrics, see "QC Metrics Thresholds - Recommendations (Plot and Table)" on page 152.

To visualize aberrations

You can visualize statistically significant aberrations across any samples in an Experiment detected on a given chromosome.

NOTE If you want to average log ratios on aberrant probes, you should first group the samples by attribute. See "To combine replicates" on page 21.

1. In the **Navigator,** select the array(s) to be used in the aberration analysis.
2. Select the chromosome from which to visualize aberrations from the **Genome View**.
3. In the **Menu** bar, click **Reports> Graphical Aberration Summary** to display the Graphical Aberration Summary.

To visualize common aberrations

You can visualize statistically significant aberrations shared in common genomic intervals between two or more samples. Differential subsets of the samples can be created for comparative analysis, and common aberrations can be filtered by attribute(s).

1. In the **Menu** bar, click **Reports> Graphical Common Aberration...** to display the Common Aberration dialog box. See Figure 3-37 on page 134.
2. In the **Navigator,** select the arrays to be used in the common aberration analysis and drag the icons to the main window in the Common Aberration dialog box.
3. From the list of available algorithms, select the common aberration calling method. See "Common Aberration" on page 114 for more information on available common aberration algorithms.
4. Select the scope of common aberration analysis. Genome scope is selected by default and will include all chromosomes.
5. Change the input parameters if necessary and click **Calculate Aberration** to continue.
6. Select the chromosome from the **Graphical Aberration Summary** window by pressing the < or > button to decrement or increment the chromosome number under review See Figure 3-39 on page 138.

7 Choose the attribute to group the samples by in the graphical common aberration summary by selecting from the **Select Attributes** list.

8 Select the samples for inclusion in up to two differential common aberration analysis sets.

9 Select the samples for exclusion (ignore) from any common aberration analysis set.

10 Once a chromosome of interest is chosen, click **Common Aberration...** to generate a graphical representation of shared significant aberrations. The Graphical Aberration Summary dialog box appears. See Figure 3-38 on page 136.

11 Optionally create and apply an interval filter to the graphical representation of common aberrations or sort the samples by aberration score. For more information, see "To create graphical common aberration filters" on page 35.

12 Select the genomic interval and chromosome from the Graphical Aberration Summary dialog box tab pane to refocus the graphical representation of common aberrations.

Report Creation and Export

CGH Analytics reports are tabulated files suitable for subsequent analysis in spreadsheet applications.

To create and export aberration summaries

You can export tabulated aberration reports for any number of arrays.

1. In the **Navigator**, select array(s) for aberration summaries.
2. In the **Menu** bar, click **Reports > Text Aberration Summary** to display the Text Aberration Summary Setup dialog box. See Figure 3-63 on page 165.
3. Select from the list of available **Report Types:**
   - **Probe Based** – Reports by each probe on specified array(s).
   - **Interval Based** – Reports by genomic intervals.
   - **Probe & Interval Based** – Reports by probe and interval.
4. Select the scope of the aberration report. **Complete Genome** is selected by default and will include all chromosomes.
5. Enter a file name for the aberration report and click **Browse** to specify the new file path location.
6. Click **Save**. An aberration report is generated and saved to disk. See "Common Aberration..." on page 78 for a description of the output format.

NOTE Interval based aberration summary reports may be nested and thus differ from what is seen in the visualization panel.

To create and export common aberration text summaries

You can export tabulated reports by interval for shared aberrations in common genomic regions.

NOTE You must first create a common aberration node within the experiment. See "To create common aberration summaries" on page 34.

1. In the **Navigator**, right-click a **Common Aberration** node.
2. Choose Generate Text Summary.
3. Enter a file name for the common aberration output file and click **Save**. An aberration report is generated and saved to disk. See "Reports Menu" on page 76 for a description of the output format.

To create and export penetrance experiments

Penetrance plots display the percentage of selected arrays that have an aberration at each probe position on the array for the selected chromosomes. You can create tab-delimited report files of these summaries.

1. Click **Reports > Text Penetrance Summary**.
2. Navigate to the folder where you want to save the gene list and type a file name.
3. Click **Save**.

To export experiments

CGH Analytics experiments can be exported as special .zip files.

1 In the **Menu Bar**, click **File > Export > Experiments** to open the Export Experiments dialog box. See "Export dialog box" on page 161.

Alternatively,

2 In the **Navigator**, right-click on the Experiments folder or on any experiment subfolder and select **Export**.

3 Select the experiment(s) you would like to export as special .zip files.

4 Click **OK**. The Export dialog box opens. Choose the disk location to save the file(s). See "Export dialog box" on page 161.

5 Click **Export.**

To export filters

Aberration, array level or feature level filters created in CGH Analytics can be exported as XML files.

1 In the **Menu Bar**, click **File > Export > Filters** to open the Export Filters dialog box. See "Export Filters dialog box" on page 163.

2 Select the filter(s) you would like to export as .xml files.

3 Click **OK**. The Export dialog box opens. See "Export dialog box" on page 161. Choose the disk location to save the file(s).

4 Click **Export**.

To export gene lists

1 In the **Navigator**, go to the **Gene List** folder and select a gene list. To create a gene list, see "To create gene lists" on page 35.

2 Right-click the **Gene List** and select **Save As** from the drop-down menu. The Save As dialog box appears.

3 Navigate to the folder where you want to save the **Gene List** and type a **File name**.

4 Click **Save**.

1 How-To

To export gene lists

2
Main Windows Reference

CGH Analytics allows you to visually explore genome-wide patterns in the data as well as zero in on specific chromosomes or specific features of microarrays. The Main Window Components section provides a brief overview of CGH Analytics main components. To get a basic understanding of how to operate the software, start by reading this overview.

Window Components

The main display for CGH Analytics is divided into several interconnected panels that display data at various levels of detail.

The philosophy behind CGH Analytics is to show simultaneous levels of expanded views so that you can view overall trends at a glance in the main window and still be able to scrutinize details at the various zoom levels with minimal manipulation of the user interface.

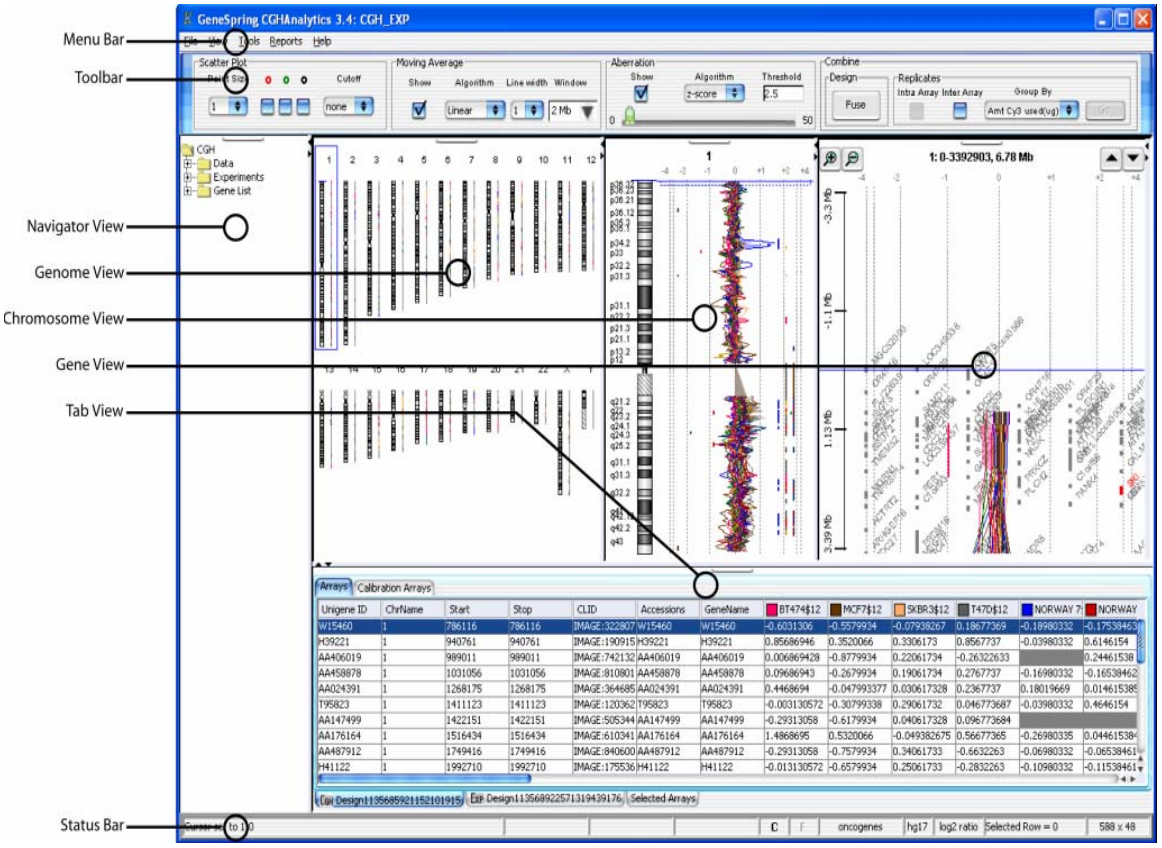


Figure 2-1 The main display with the multiple viewing areas labeled

The main display consists of eight basic components.

Navigator View

The Navigator area is used to organize microarrays in experiments and run analyses on these experiments. The area displays the individual experiments, the design files, and the raw data files (microarrays). See "Navigator" on page 49. Right-clicking some Navigator elements enable shortcut menus that are described in more detail under "Shortcut Menus" on page 53.

Genome View

The Genome view display the selected data sets plotted in the context of a full genome. This allows you to see an overview of all chromosomes in the genome from which you can select specific chromosomes to view in more detail.

Click a chromosome in the overview to select it and set the view to the location clicked within the chromosome.

NOTE

The Genome view is disabled and grayed by default. You can click on any chromosome to select it, even if the Genome View is disabled. You can enable the Genome View by right-clicking somewhere in the pane, and clicking the **Show Data** command option.

Chromosome View

The Chromosome view displays a single, user-selected chromosome. To select an area on the chromosome, right click anywhere in the Chromosome View and drag the mouse to another location. A rectangle will form to enclose the selected area, and the cursor will appear as a blue line across the middle of the area.

Gene View

The Gene view is linked to the Chromosome view in that it shows an even greater magnification of the selected location to the level of an individual transcripts. Click in the Gene view to select that position and center the cursor at that position in the view pane. Since the cursor location is always in the center of the pane, you can perform a pseudo-scrolling operation by repeatedly clicking at the top or bottom of the view.

Menu Bar View

The Menu bar contains five main menu choices from which, in turn, you can access one or more submenus. These menus control a variety of options ranging from simple computer functions, such as printing and copying, to more application-related functions, such as creating experiments, importing existing files, and generating specific reports. For more information, go to "Menu Bar" on page 60.

Toolbar View

The Toolbar enables you to change important display settings. These settings remain on display at all times, providing you with information on the current settings and which other settings are available. For more information, go to "Toolbar" on page 80.

Tab View

The table across the lower portion of the screen provides detailed information on the individual microarray ratios along with positional information and a few columns of annotation (usually associated with gene or transcript data). Clicking in a table row will select the chromosome and position that is associated with the row. All other views are then synchronized to that position. See "Tab View" on page 87.

It is important to note that all data views are linked and clicking in one view or a portion of one view synchronously navigates all other views to the same location (but at different levels of detail).

Status Bar View

The Status bar at the bottom of the screen contains nine subpanels which show current applied filters, algorithms, and genomic display focus. See "Status View" on page 90.

Navigator

The Navigator area provides a place where you can view all of the experiments, designs, microarrays, and various gene lists arranged in a simple treelike structure.

You can also drag and drop microarrays from Data Nodes to Experiment Nodes. (Only arrays of the same species can be added to an experiment.) If a Design node corresponding to an array does not exist in an experiment, a new node is created for that design and the array is added to that Design node.

The Navigator area is located in the left pane of the main window. Folder categories listed in collapsed format can be expanded by clicking the plus (+) symbol to the left of the folder name. Below is a representative view (in partially expanded format). The root-level nodes—Data, Experiments, and Gene List—are displayed in a treelike structure by default.

Important sample attributes are visible in the Navigator area. Data files that are annotated as dye-flip (a reversal of the Cy3 and Cy5 fluorochrome labelled samples) are indicated by a box enclosing an italic letter *f*. Data nodes created from multi-pack arrays are indicated by a four small boxes adjacent to the parent design. The selected Experiment for analysis is indicated by blue text. When applicable, active common aberration nodes are also indicated by blue text within the selected Experiment. Arrays selected for analysis within a selected Experiment are indicated by colored nodes.

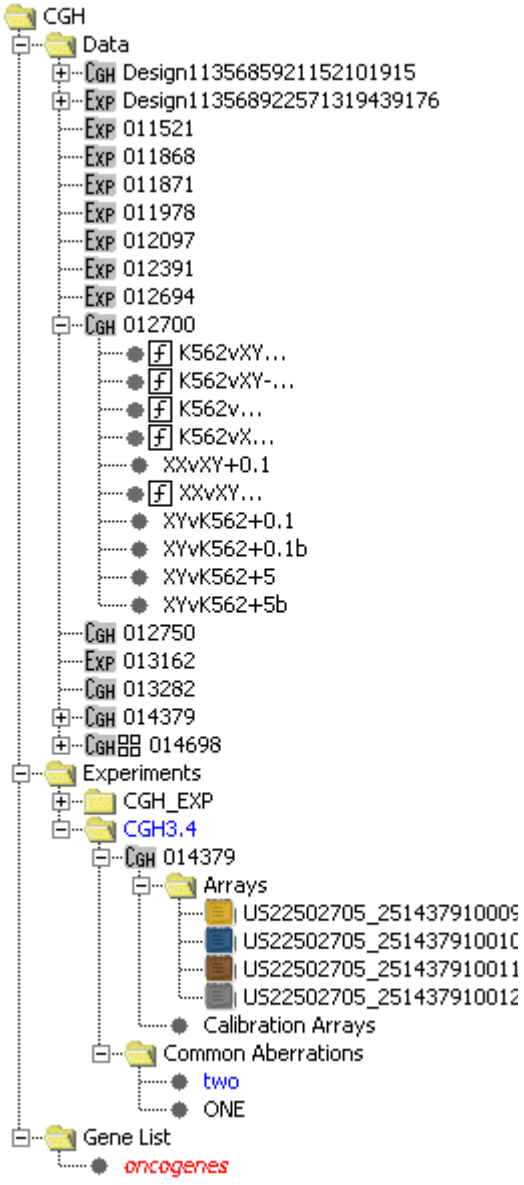


Figure 2-2 Navigator area displaying the file tree

Data Node

Imported microarrays are displayed in the Data node folder, grouped by their design number.

- All microarrays in a design are listed below the parent design number.
- CGH designs were developed to analyze comparative genome hybridization (CGH) data. Designs labeled EXP were developed to analyze **Gene Expression** data.
- Data files that are annotated as dye-flip (a reversal of the Cy3 and Cy5 fluorochrome labelled samples) are indicated by a box enclosing an italic letter *f*.
- Data nodes created from multi-pack arrays are indicated by a four small boxes adjacent to the parent design.
- Arrays imported in version 3.1 format are displayed with a random design identifier generated by the software.
- Arrays imported in UDF format are displayed with the name of the file as the design name.

Experiments Node

All experiments are listed in folders below the Experiments node with the selected microarrays listed below each experiment. They are categorized by their parent designs.

Designs are listed in folders below an Experiment node where

- Microarrays are listed below the design under either Arrays or Calibrated Arrays.
- Inter array replicated arrays are listed as merged arrays. The arrays are combined under one node following the naming convention **<Attr_name>=<value>**, where **Attr_name** is the name of the attribute used for grouping combined inter array replicates.

The Navigator view supports a drag-and-drop feature that you can use to move microarrays from the Data node to the Experiments node—a simple way to add arrays to an experiment.

Gene List Node

The Gene List node lists all of the gene lists that are available at the application level. See also “Sample Data” on page 230 and “Shortcut Menus” on page 53.

Shortcut Menus

When you right-click on the nodes in the Navigator's tree structure—Data, Experiment, or Gene List—various shortcut menus appear depending on the node. These are categorized here as

- Data node shortcut menus
- Experiments node shortcut menus
- Gene List shortcut menus

Data Node Shortcut Menus

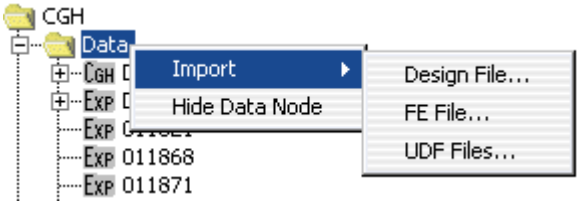


Figure 2-3 Data node shortcut menus

Import

Selecting Import provides options for importing three types of files:

Design File Imports Agilent design files. See "To import design files" on page 13.

FE File Imports microarray files created in the Feature Extraction application. See "To import Feature Extraction files" on page 14.

UDF Files Imports Universal Data Files (UDFs), tab-delimited text files created by another, non-Agilent application. See "To import universal data files (UDFs)" on page 17.

Hide Data Node

Selecting this option hides the Data node from view. It can be used to reduce desktop clutter when no additional transfers of microarrays are needed. To redisplay the Data node, right-click anywhere in the Navigator area and click **Unhide Data Node**.

Modifying Design Files

Design files are imported into the Data folder, and data files are imported into the Design folder. Right-clicking the design file name displays a shortcut menu with three options. See also "To explore or update design files" on page 24.

QC Metrics Displays a QC Metrics table dialog box listing the metrics for the array within that design. See Figure 3-49 on page 150. For more information on metrics, see "QC Metrics" on page 56 under the Tools menu.

Update from eArray Updates the annotations for your Agilent microarray design from the eArray Web site. Every three months, Agilent microarray annotations are updated at the eArray site to incorporate new genome annotations as they become available from public sources.

Delete Displays a Confirm array deletion dialog box to affirm or cancel a decision to permanently delete the selected array.

Modifying Array Files

Right-clicking a data file name displays a shortcut menu with four options. See also "To modify array files" on page 26.

Show Properties Displays the Microarray Properties dialog box giving you access to a tabbed list of the array attributes. See also.

- **Attribute.** At the array level, **Attribute** lists the parameters of the experimental conditions for that array. This is an important way to classify arrays into categories. For example, you can use attributes as filter elements in configuring Array Level filters. When combining by inter array, attributes can be used in deciding which arrays to combine. See Figure 3-16 on page 109.
- **FE Headers.** Displays information on each array with the Feature Extraction statistics and parameter table information. See Figure 3-17 on page 110.
- **FE Features.** Displays all of the imported data for each feature of a microarray from a Feature Extraction file. See Figure 3-18 on page 111.

QC Metrics Displays a QC Metrics Table dialog box listing the metrics for the selected array. See Figure 3-49 on page 150. For more information on metrics, see "QC Metrics" on page 56 under the Tools menu.

Rename Displays a **Input** dialog box where you can rename the selected array.

Delete Displays a Confirmation dialog box where you can affirm or cancel a decision to permanently delete the selected array.

Experiments Node Shortcut Menus

Experiment creation / export

Right-click the **Experiments** folder in the **Navigator** pane and a shortcut menu with two options appears.

New Experiment The **New Experiment** dialog box opens where you can name the new experiment, describe it, and specify its properties in terms of the microarrays it contains.

Export The **Export Experiments** dialog box opens where you select the experiment to export as a .zip file and the disk location to save the file.

Experiment selection

Right-click on an available deselected experiment, and a shortcut menu with 11 options appears.

Select Experiment Displays the Experiment Selection dialog box prompting you to confirm or cancel your selection. If you confirm, the selected experiment is restored. You can also select an experiment by double-clicking the experiment's name.

Rename Displays and Input dialog box where you can rename the selected experiment.

Delete Removes the selected experiment.

Show Properties Displays the Experiment Properties dialog box where you can edit the experiment's association with specific microarrays and their descriptions. See Figure 3-34 on page 131. This field is grayed for experiments with non-Agilent arrays.

Export The **Export Experiments** dialog box opens where you select the experiment to export as a .zip file and the disk location to save the file.

QC Metrics Displays a QC Metrics table dialog box listing the metrics for the selected experiment. See Figure 3-49 on page 150. For more information on metrics, see "QC Metrics" on page 56 under the Tools menu. This field is grayed for experiments with non-Agilent arrays.

Change Genome Build If an experiment consists of multiple array designs that utilize different genome builds, you can select one appropriate genome build to represent the entire experiment.

Show Spike-in Displays the Show spike-in .dialog box, which shows a graph of the calculated DNA fold change (x axis) and the spike-in ratios (y axis) in the experiment that uses a spike-in reference DNA. See Figure 3-56 on page 158.

Show Spike-in Ratios Displays the Show spike-in ratios dialog box, which shows a graph for each ratio (all arrays of the experiment) in separate panels with the number of arrays (x axis) and the median of log2 value for each arra;y (y axis). Figure 3-58 on page 160.

Edit Array Color Displays the Edit Array Color dialog box where you can change the color used to display one or more microarrays in your experiment. See Figure 3-27 on page 122.

Edit Array Order Displays the Edit Array Order dialog box where you can change the order of the microarrays in a specific design. See Figure 3-29 on page 125.

Design selection

Right-click on any design in an experiment, and a shortcut menu containing four options appears.

Set for Calibration Assigns all microarrays from that design for calibration as part of the calibration step, and moves those arrays to the Calibration Array folder under the same design node of that experiment. This option is not recommended for most analysis situations.

QC Metrics Displays a QC Metrics table dialog box listing the metrics for the arrays of the selected design. For more information on metrics, see "QC Metrics" on page 57 under the Tools menu. The QC Metrics option is not enabled (grayed) for non-Agilent data.

Save as Text file Displays a Save Design dialog box in which you can designate the folder and file name where you can save those design files in CGH Analytics v. 3.1 tab-delimited format.

Delete Deletes the highlighted design and its associated microarrays from the experiment.

Array selection

Right-click on any array listed under a design in an experiment, and a shortcut menu with eight options appears.

Select or Deselect (toggle) Selects a microarray for an experiment. In the Tab view, the microarray is marked with a colored box of the same color as the data line plots and aberration calls for that microarray in the display windows. If an array is already selected, clicking this option deselects the selected array

Select or Deselect for Calibration Includes the selected microarray in the list of microarrays used in calibration. If the microarray is already selected, clicking this option removes it from the calibration list.

Edit Array Color Displays the Select Color dialog box. See Figure 3-53 on page 155. A tabbed color selector allows you to specify the color of the gene list based on color samples displayed as swatches, based on an HSB schema (Hue, Saturation, and Brightness or Value), or based on an RGB schema (Red-Green-Blue). Only one gene list can be displayed at any one time.

Show Spike-In Displays the spike-in summary pane.

Show Properties Displays a submenu with a single option.

- Attribute... This option displays the Microarray Properties dialog box with access to a list of the array's attributes. See Figure 3-16 on page 109.

  At the array level, **Attribute** lists the parameters of the experimental conditions for that array. This is an important way to classify arrays into categories. For example, you can use attributes as filter elements in configuring **Array Level** filters. When combining interarrays, attributes can be used in deciding which arrays to combine.

QC Metrics Displays a QC Metrics table dialog box listing the metrics for that particular microarray. For more information on metrics, see "QC Metric" on page 71. This field will be grayed here for experiments with non-Agilent arrays.

Rename Provides a way to rename the microarray.

Delete Deletes the microarray from an experiment.

Common Aberration selection

Right-click on any common aberration node listed under a common aberration folder in an experiment, and a shortcut menu with five options appears.

View details Displays an analysis summary about the common aberration. See "Common Aberration details dialog box" on page 115. The information displayed includes:

- Name of common aberration node
- The aberration calling algorithm used to create the common aberration
- The threshold applied to the aberration calling algorithm
- The genome build of the samples used in analyzing the common aberration
- The common aberration algorithm used
- The p-value threshold applied to the common aberration algorithm
- The overlap threshold applied to the common aberration algorithm
- The scope of the common aberration analysis
- The sample names used in the common aberration analysis

Show in UI/Hide Toggles display of the common aberrations shared between samples in the genome, chromosome, and gene views.

Generate text summary Creates an tabular interval based aberration report and saves the file to disk. See "Reports Menu" on page 76 for a description of the output format.

Rename Provides a way to rename the common aberration node.

Delete Deletes the common aberration node from an experiment.

Gene List Shortcut Menus

Import Gene List

Right-click the **Gene List** node to display an Import dialog box in which you can specify the folder and file name of a gene list you want to import. You can import **All Files** in the specified folder or only those that have a *.txt file extension.

Once a gene list is imported or if a gene list is already available, you can right-click on the gene list to display a menu with seven options.

View in Table Allows you to select the gene list that applies to the selected experiment if a gene list exists. See Figure 3-62 on page 164. The genes are listed in a Gene List dialog box by number and name.

Clicking the **Color** button displays a tabbed color selector where you can specify the color of the genes in the list. See Figure 3-53 on page 155. Only one gene list can be displayed at any time.

Rename Displays an Input dialog box where you can enter a new name for the gene list.

Delete Displays a Confirmation dialog box asking you to affirm or cancel a decision to delete the gene list.

Save As Displays a Save As dialog box in which you can designate the folder and file name in where you want to save the gene list as a file *.txt file extension

Add to Gene list Selecting this option allows you to add the selected gene list to another gene list that you specify.

The following options will vary depending on user selections:

Show only Applies any changes you made to the gene list and applies the gene list to the selected experiment.

Show all Allows you to remove the gene list restriction from display consideration. Right-clicking the gene list file name a second time provides a submenu where the **Save As** option is replaced by **Highlight**. Clicking **Highlight** will re-apply the selected gene list.

Menu Bar

CGH Analytics includes a main menu bar containing five menu choices, which in turn lead to related submenus. The application also features additional menus accessed from column headers the **Tab** view and menus for Web searching. Each menu type provides access to one or more submenus.

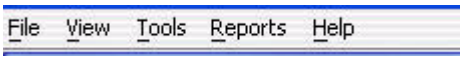


Figure 2-4 Main Menu Bar

File Menu

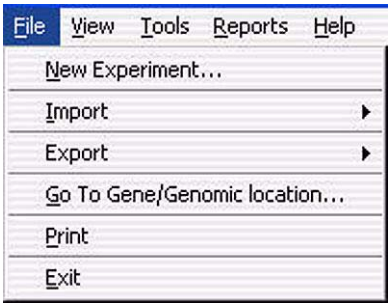


Figure 2-5 File Menu

New Experiment...

This option opens the Create Experiment dialog box where you can name and describe a new experiment. See Figure 3-14 on page 107.

Import

The Import option leads to seven submenus.

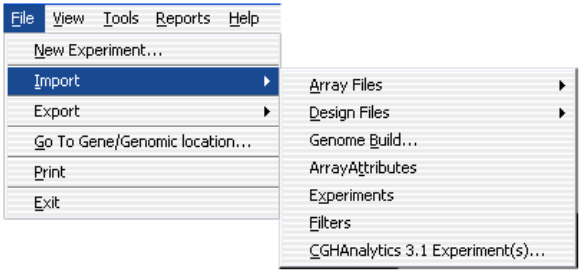


Figure 2-6 File Import submenu

Array Files Displays two additional submenus.

FE File Displays the Import FE Files dialog box where you can select one or more **Feature Extraction** files in microarray format. The files are identified, validated, and imported. See Figure 3-3 on page 95.

UDF File Displays a CGH 3.1 Files dialog box where you can select a file in the CGH Analytics 3.1 format. See Figure 3-5 on page 97. The tab-delimited file is used to import microarray data from other technologies.

Design Files Displays a submenu.

Design File Opens an Import Design File dialog box from where you can select an Agilent microarray design file in GEML file format. See Figure 3-1 on page 93.

Genome Build Information for human, mouse, and rat genomes is shipped with the software. Clicking this submenu option opens an Import Genome Build dialog box giving you access to updated genome builds provided by Agilent. See Figure 3-6 on page 98.

Array Attributes Opens an Import microarray attributes dialog box. See Figure 3-7 on page 99. From this dialog box you have access to any folder on your system that contains **Microarray Attribute** information files in tab-delimited .txt file format.

The file format constructed as follows:

- The first row contains header information that corresponds to the attribute the values of which you wish to update.
- The header of the first column must be named Barcode.

- The rows below it list individual microarrays to be updated. Each row starts with the microarray's barcode followed by columns of the information to be updated.

The following illustrates an update file. It lists the barcodes of two microarrays, the name of the person who hybridized them, and the temperature used in the K562 study.

Barcode	Hyb'd By	Hyb Temp
251270010402	Sam Spade	65
251270010411	Sam Spade	65

Experiments Opens an Import Experiments dialog box. See Figure 3-2 on page 94. From here you can access any folder on your system that contains experiments previously exported from CGH Analytics in ZIP file format.

Filters Opens an Import file dialog box. From here you can access any folder on your system that contains filters previously exported from CGH Analytics in XML file format.

CGH Analytics 3.1 Experiment(s) Opens an Import dialog box giving you access to any folder on your system that contains data files in CGH 3.1 format. See Figure 3-5 on page 97. By default, this folder is in: **C:/Program Files/CGHAnalytics3.X.X/data** folder.

Export

Selecting this option displays two submenus.

Experiments Opens the Export Experiments dialog box that lists the experiments on your system. See Figure 3-59 on page 161.

To export one or more experiments in ZIP format, select the check boxes of experiments you choose to export, then click **OK**. An Export dialog box appears where you can specify the destination folder for receiving the exported experiments.

Filters Displays an Export Filters dialog box that lists the filters on your system by name and type. See Figure 3-61 on page 163.

To export one or more experiments in XML format, select the check boxes of filters you choose to export, then click **OK**. An Export dialog box appears where you can specify the destination folder for receiving the exported filters.

Go To Gene/Genomic location...

Displays a dialog box that allows you navigate directly to a gene name selected from a scroll-down list displayed in the **Gene** view or to a specific genomic location based on coordinates. See Figure 3-35 on page 132.

Print

Opens a standard Print dialog box from where you can send the document to a designated printer based on the printer and print settings selected.

Exit

Closes the CGH Analytics application.

View Menu

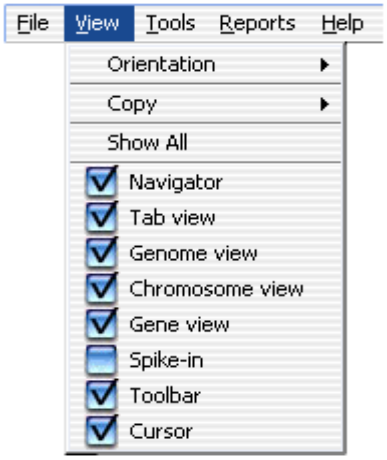


Figure 2-7 View menu

Orientation

This option allows you to modify the display of three of the main screen elements from a vertical presentation (chromosomes displayed top to bottom of the screen) to a horizontal presentation (chromosomes displayed left to right on the screen). To return to the previous orientation, again click **View > Orientation** and select the desired view. See "Main view with horizontal orientation" on page 65.

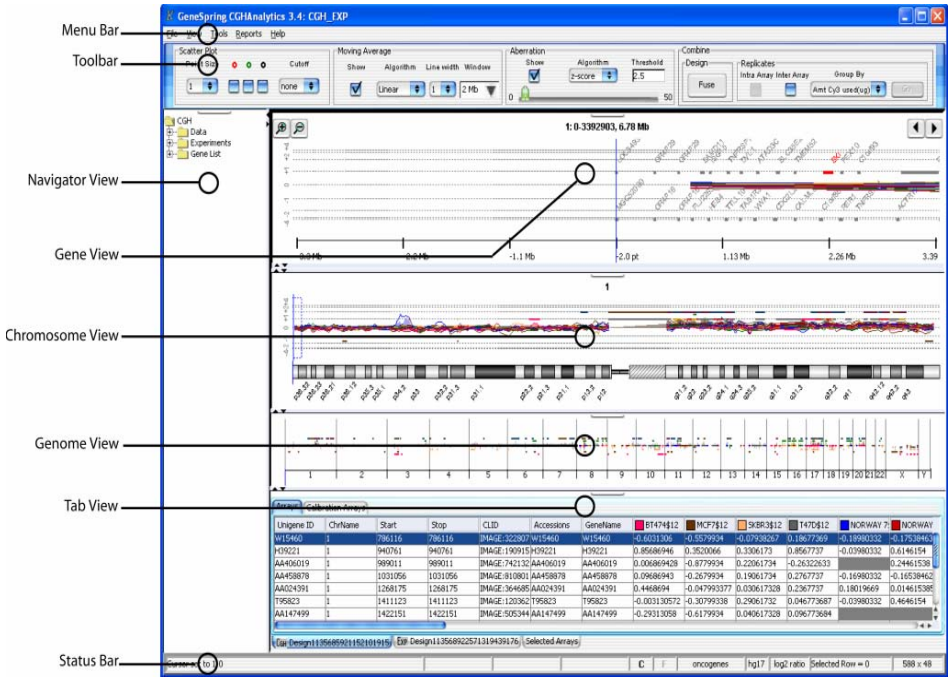


Figure 2-8 Main view with horizontal orientation

Copy

Copy functions make it easier to move image data from the display into an application. Selected components of the display are copied to the system clipboard from where you can simply paste the component into other applications.

Click **Copy** to display a submenu with seven options.

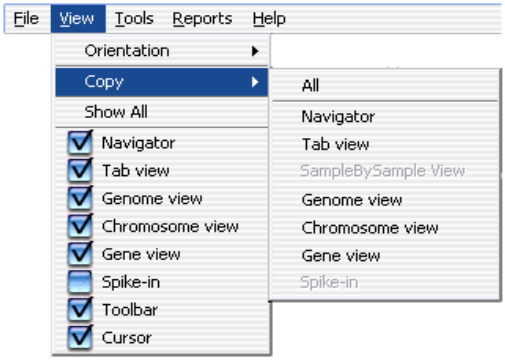


Figure 2-9 View menu showing Copy submenu

All Copies all of the main panes of the display (excluding the toolbar). The other options copy specific portions of the main display to the clipboard.

Navigator Isolates and copies the **Navigator** area to the clipboard.

Tab view Isolates and copies the **Tab** view to the clipboard.

Sample By Sample view This option is disabled (grayed) by default. When enabled, allows you to isolate and copy the **Sample By Sample** view to the clipboard.

Genome view Isolates and copies the **Genome** view to the clipboard.

Chromosome view Isolates and copies the **Chromosome** view.

Gene view Isolates and copies the **Gene** view to the clipboard.

Spike-in This option is disabled (grayed) by default. When enabled, allows you to isolate and copy the **Spike-in** view to the clipboard.

Show All

The following **View** menu options control the display of the featured areas of the **Main Window**. Selections that are checked are visible and are displayed on the desktop. Selections not checked are hidden.

The **Show All** menu is disabled (grayed) by default if all views are selected. If any view selection is not checked (hidden), S**how All** is enabled and can be clicked to enable the display of all views on the desktop.

The **Show All** menu controls the display of eight views.

- **Navigator** view
- **Tab** view
- **Genome** view
- **Chromosome** view
- **Gene** view
- **Spike-in** view
- **Toolbar**
- **Cursor**

Tools Menu

The **Tools** menu provides a place for filtering results, changing array attributes, executing Agilent-supplied or user-written plug-ins, examining **QC Metrics**, and setting user preferences.

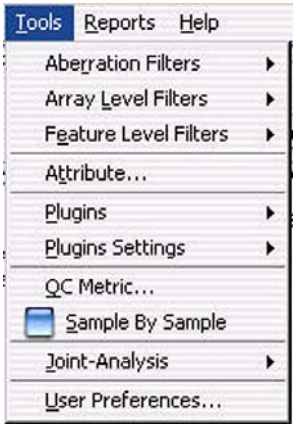


Figure 2-10 Tools menu

Aberration Filters

Aberration filters reduce the number or extent of aberrations displayed following user-set parameters.

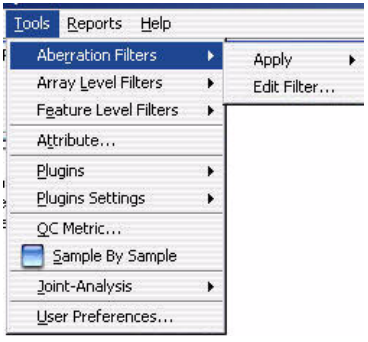


Figure 2-11 Tools menu showing Aberration Filters submenu

Apply Displays a list of the available filters. The current filter is checked. If filters have not been imported yet, **No Filter** is checked. You can select another filter by checking it, or you can remove the applied filter by checking **No Filter**.

NOTE All four aberration level filters are applied sequentially from top to bottom of the list. To make only one filter effective at a given time, you must put "safe values" in the other filters. For example, to use an Aberration Level Filter on absolute average log ratio, you must set the other three filters to safe values such as: Minimum number of probes in a region = 1; Maximum number of aberrant regions = 2000; and Percent penetrance per feature = 1.

Edit Filter Displays the Edit Aberration Filters dialog box that allows you to create, modify, or delete filters. See Figure 3-26 on page 121.

Array Level Filters

Array Level filters set attributes or values that you can use to identify similar arrays. You can filter an array or group of arrays in or out using a specific attribute from the selected experiment. You can also create a new filter by combining more than one attribute from the list using logical AND/OR operators.

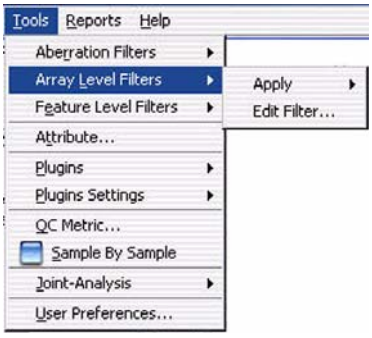


Figure 2-12 Tools menu showing Array Level Filters submenu

Apply and Edit Filter Procedures for Aberration filters are applicable to Array Level filters, but filter parameters differ as shown in Figure 3-28 on page 123.

Feature Level Filters

Various columns from Feature Extraction output files are loaded in this software. You can apply filtering at the feature level using data in these columns. The default Feature Level filter removes features that have a non uniformity flag or a saturated flag set in either channel, i.e. either glsFeatNonUnifOL, or rlsFeatNonUnifOL, or glsSaturated, or rlsSaturated set. Currently, you cannot create compound filters.

Apply and Edit Filter

Procedures for Aberration Filters are applicable to Feature Level filters as well, but the parameters differ as shown in Figure 3-31 on page 127.

Attribute

This option displays the Attributes dialog box for creating, modifying, or deleting attributes. See Figure 3-30 on page 126.

Attributes are basic items of information that you would record about particular arrays, such as the hybridization time (hyb time), hybridization date, or sample type. Attributes are array specific. An attribute's values for specific microarrays can be changed by right-clicking the microarray in the Navigator area, and selecting **Show Properties**. Alternatively, you can select a tab-delimited text file that contains the attributes of the microarrays by clicking **File > Import > Microarray Attributes**.

Plugins

The plug-ins used by CGH Analytics are displayed in a submenu.

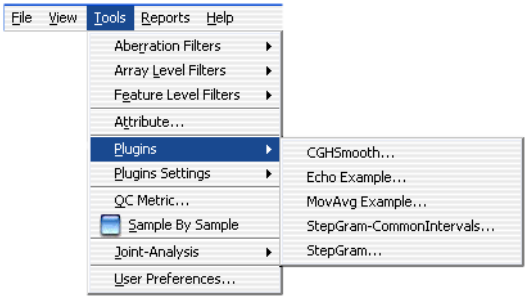


Figure 2-13 Tools menu showing Plugins submenu

Five plug-ins are supplied with the standard installation.

CGHSmooth Does various data-smoothing operations. For more information, see "CGHSmooth Plug-in" on page 229.

Echo Example Echoes STDIN back to STDOUT. For more information, see "Echo Example.c" on page 228. This is a very small C program.

MovAvg Example Computes a 10-point moving average of each column of array data. For more information, see "MovAvg Example.pl" on page 228 and Figure 3-44 on page 145. This is a short Perl program.

Plugins Settings...

This option allows you to set the parameters for two types of plugins.

CGHSmoothing See "CGHSmooth Plug-in" on page 229

MovAve See "MovAvg Example.pl" on page 228 and Figure 3-44 on page 145 for an example of the plug-in's parameters.

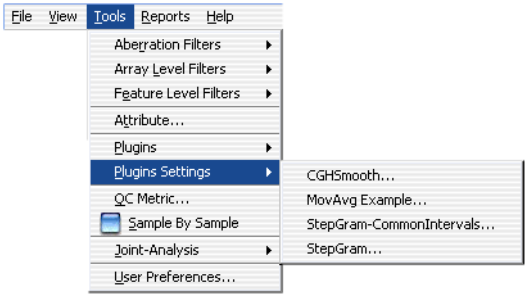


Figure 2-14 Tools menus showing Plugins Settings submenu

QC Metric

The QC Metrics table dialog box displays the various metrics for each array in a selected experiment grouped by a specified attribute. See Figure 3-49 on page 150. In this table and the associated plot, you can evaluate the quality of your aCGH microarray results and assign a QC status to each microarray based on this evaluation.

NOTE The QC Metrics table also can be displayed in two other ways: By clicking **Experiments > Selected Experiment** then right-clicking the experiment to display its microarrays or by clicking **Data > A Selected Design > Selected Array** then right-clicking the array to display microarrays in the design. For more information on QC Metrics, see Microarray QC Metrics.

SampleBySample

This option allows you to analyze one sample at a time and create or edit any research notes on that sample. Two text fields appear with array-level attributes, **Karyotype** and **Research Notes,** displayed above the Tab view pane. The attributes are pre-named but you can edit them, but you cannot swap them for new attributes.

When you toggle to this view, only the first array in the experiment is displayed, but you can select another array by double-clicking its name in the Navigator view.

Joint Analysis

This option provides two ways that to compare CGH Analytics data with Gene Expression data—**Correlation Analysis** and **Enrichment Analysis**.

Correlation Analysis

This type of analysis determines if correlations exist between selected CGH and Gene Expression (GE) arrays in an experiment.

You can choose

- **Setup** – Click to set up a new analysis. See Figure 3-24 on page 118.
- **Perform** – If enabled, click to perform an analysis you set up previously.

Gene Selection You can choose to analyze data from the

- **Entire Genome** – All arrays in the genome.
- **Active Gene List** – Only regions of the genome included in the active gene list.
- **Genomic Location** – Only analyzes genomic regions that you select in the following dialog box. Click to display the Select Chromosome Interval dialog box where you name the chromosome or chromosomes, and the starting and stopping locations of each interval. See Figure 3-52 on page 154.

Apply to

- **All** – Apply the correlation analysis to all arrays in the experiment.

- **Selected** – Apply the correlation analysis only to the arrays currently selected in the experiment.

Threshold Display the user-defined threshold value that separates samples that show correlation (above the threshold and colored red) from samples that show insignificant or no correlation (below the threshold and colored grey). The default value is 3.5.

Match Sample By Indicate the attribute to use for matching the samples arrays. The default is Sample Name.

Perform Analysis Click to display the Matched Sample dialog box. See Figure 3-42 on page 143.

Statistical Algorithms Three statistical approaches are used to analyze correlation data:

- Student-t-test (amplified vs. non-amplified)
- Student-t-test (deleted vs. non-deleted)
- Pearson p-value

Student-t-test approaches (both types) Let P be the set of samples for which a region spanning g is called amplified. Let Q be the set of samples for which a region spanning g is called deleted. Let U denote the entire set of samples.

- Let
  - μ_P be the average expression value, of g, for samples in P,
  - μ_Q the average expression value, of g, for samples in Q,
  - μ_{U-P} the average expression value, of g, for samples in U-P (the complement of P), and
  - μ_{U-Q} the average expression value, of g, for samples in U-Q (the complement of Q).
- Student-t statistics and the corresponding p-values are computed for one-sided alternative hypotheses:
  - $\mu_P > \mu_{U-P}$ (g's expression levels in amplified samples are larger than those in non-amplified samples)
  - $\mu_Q > \mu_{U-Q}$ (g's expression levels in amplified samples are larger than those in non-amplified samples)

Pearson p-value Correlation of C and E and a corresponding p-value.

In the tab view of the "Correlation Results dialog box" on page 119, all three scores are reported for every gene *g*. The largest value of the three scores is reported as the maximum value the fourth column. Output formats include graphical output as well as a text report reflecting the bottom panel table.

Enrichment Analysis

CGH-Expression Enrichment Analysis can be computed for any sample in which both CGH and expression were measured. The output is a list genomic intervals for each sample with a statistical score showing the enrichment of over- and under-expressed genes in that interval.

You can choose

- **Setup** – Click to set up a new analysis. See Figure 3-33 on page 130.
- **Perform** – If enabled, click to perform an analysis that you set up previously.

Input scores

You can select from three types of enrichment analyses:

- **Enrichment of correlation score** – This selection is not enabled until after a correlation analysis has been completed and you have correlation scores available.
- **Enrichment of expression values** – These are expression values you already have.
- **Enrichment of external score** – This allows you to detect enrichment in non-expression data in the detected aberrations of a sample. You are prompted to type in the path to the folder holding the external scores. A **Browse** button is provided. The file format should be tab-delimited text.

  The first line is a header line with columns titled Chr, Start, Stop, and Score <array name>. Subsequent rows contain data on the chromosome number, start and stop position of the value being reported and the score (a decimal number or name). For example, the first two rows of a file to analyze one of the CGH_EXP arrays (BT$&$) could look like:

```
Chr          Start          Stop         Score-BT4
1           779448       808099      -0.6
```

Interval Selection

You can choose to analyze the entire genome or only the aberrant intervals.

Input Parameters

This are user-designated parameters where you can type in the `Maximum Interval Length` and `mHG Threshold`. The defaults are 10 and 0.01, respectively.

Perform Analysis Click to display the Match Sample dialog box. See Figure 3-42 on page 143.

User Preferences

The Preferences dialog box allows you to select how you would like to display the applications properties that you will use for rendering views, calculations, etc. These properties remain unchanged across sessions, but they can be edited. Four tabbed options are available:

View Tab Controls the general characteristics of how data are displayed on your monitor. See Figure 3-67 on page 171.

Gene Symbols Tab Controls typeface, style, size, and text orientation of gene symbols. See Figure 3-64 on page 166. Change these settings to improve the legibility and clarity of the gene symbols in the Gene view.

Miscellaneous Tab Controls the settings for several aspects setting up CGH Analytics. See Figure 3-66 on page 169.

License Tab The License tab options displays information on the license you provided when you initially installed CGH Analytics 3.3. See Figure 3-65 on page 167.

Reports Menu

The Reports Menu provides access to six types of reports.

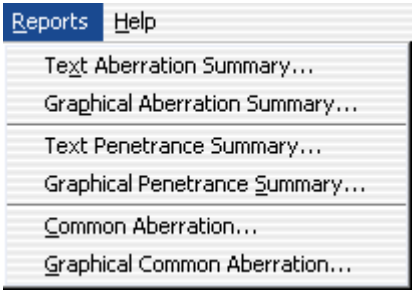


Figure 2-15 Reports menu

Text Aberration Summary...

This selection displays a Select Report File dialog box. See Figure 3-63 on page 165. When completed, a tab-delimited, Excel-compatible spreadsheet is generated containing those microarray probes that have high scores according to the algorithm used. All scores for these cases are included in the generated spreadsheet. After the report is generated on Win32 platforms, you can view the report in Excel.

The report displays three columns for each array: **Log Ratio**, **Amplification Score**, and **Deletion Score**, in that order.

NOTE This summary will be only for the chromosome aberration detection algorithm currently selected for the microarrays.

Graphical Aberration Summary...

This selection generates an alternative, graphical view of the aberrations. See Figure 3-36 on page 133. Each microarray is shown as a heat map aligned with the ideogram of the selected chromosome.

NOTE The summary is constructed only for the selected microarrays in a selected chromosome. However, the display is actively linked to the main display. Changes to the microarray selections and/or Z-scoring parameters are reflected immediately in the summary display.

On Win32 platforms, the Menu bar displays an Edit menu containing a single function: **Copy**. You can use **Copy** to copy the current Aberration Summary to the system clipboard from where you can copy and paste it into a report, publication, or presentation.


Due to technical issues, this **Copy** functionality is not supported on Apple and Linux platforms.

Tip: You can create separate summaries for different chromosomes by selecting a chromosome, generating a summary, and then selecting another chromosome, and generating a second summary for that chromosome, etc. This can be used for chromosome-to-chromosome comparisons that are not possible with the main four-panel display.

Text Penetrance Summary...

This selection displays a Select Report File dialog box identical to the one displayed for the Text Aberration Summary report. The completed dialog box generated a Text Penetrance Summary report.

Graphical Penetrance Summary...

This report generates a graphical view of a penetrance plot and displays it in a new window. See Figure 3-40 on page 140. The penetrance plot displays the percentage of selected arrays that have an aberration at each probe position on the array for the selected chromosomes.

Common Aberration...

Common aberration analysis is used with multiple samples to identify genomic intervals that have common aberrations that are statistically significant. For more information, see "Common Aberration Analysis" on page 201.

A common aberration analysis can create either an interval based report or a probe based report. An interval based report output format includes the following columnar headers:

- Aberration number - The number of the aberration listed (in order, beginning with chromosome 1).
- Cytoband - The cytoband position.
- Chromosome - The chromosome number of the aberration.
- Start - The first base included in the aberration.
- Stop - The last base included in the aberration.
- p-value - Statistical p-value for this aberration (or z-score threshold).
- Amplification/Detection - Indicates if this aberration is an amplification or deletion.
- Aberration Type - Indicates whether this aberration is a single sample aberration or a common aberration between samples.

A probe based report output format includes the following columnar headers:

- Aberration number - The number of the aberration listed (in order, beginning with chromosome 1).
- Cytoband - The cytoband position.
- Chromosome - The chromosome number of the aberration.
- Start - The first probe base included in the aberration.
- Stop - The last probe base included in the aberration.
- Gene Name - The gene name that the probe interrogates.
- Log Ratio - Measured log ratio of this probe.
- p-value - Statistical p-value for this aberration (or z-score threshold).
- Amplification/Detection - Indicates if this aberration is an amplification or deletion.
- Aberration Type - Indicates whether this aberration is a single sample aberration or a common aberration between samples.

Graphical Common Aberration...

- This report generates a graphical view of a common aberration summary and displays it in a new window. See Figure 3-38 on page 136. The **Graphical Common Aberration** colors significant Z-scores in heat-map fashion (red for putative amplifications and green for putative deletions). The scoring is the same as it is in the main display.

Help Menu

Help

Displays the CGH Analytics User's Guide.

Quick Start

Displays the CGH Analytics Quick Start Guide. The Quick Start Guide is brief introduction and an easy reference companion to common analysis tasks in CGH Analytics.

About CGH Analytics

Displays the ubiquitous "About" box information on the software's version, copyrights, etc.

Toolbar

The Toolbar is displayed across the top of the screen below the Menu Bar. It contain four groups of tools.

- Scatter Plot
- Moving Average
- Aberration
- Combine Replicates

Scatter Plot

Figure 2-16 Scatter Plot dialog box

A scatter plot simply plots the data aligned with the corresponding chromosome.

Performance tip: Rendering scatter plots for more then 10 high density arrays in the chromosome view may take significant time. It is advisable to not select more than 10 arrays if the scatter plot is turned on for chromosome view. Selecting ellipses as the rendering style for CGH scatter plots can also decrease performance. Please change the rendering style for CGH data from ellipse to the plus (+) or cross hair sign. You can make these changes from the User Preference dialog box (see Figure 3-67 on page 171).

NOTE The scatter plot is has been turned off in the genome view to enhance performance. The scatter plot will display in the chromosome and gene views.

Point Size Specify the size of the spot used to plot each data point. Sizes available range from 1 to 9 points.

Colored Circles Mark or clear the check boxes to show or hide three classes of ratios to plot:

- Red corresponds to large positive ratios (increased copy number.
- Black corresponds to log ratios near zero that are considered insignificant according to the set "Cutoff" value.
- Green corresponds to large negative log ratios (decreased copy number)

This color-coding system was adopted from standard gene-expression displays where red indicates positive ratios and green indicates negative ratios.

Cutoff Cutoff points specify the fold ratios that are to be considered significant. Any ratio between cutoff points is considered relatively insignificant and is color-coded black. The values available are 0.125 to 5.00. If a value is less than plus or minus (±) this ratio, the value is considered unchanged and is colored black.

Tip: By eliminating unchanged data, you can concentrate on only the copy numbers that have increased or decreased. While not statistically rigorous, this is a convenient way of filtering the amount of data being viewed. Setting a relatively high Cutoff value and a relatively large Point-Size value will make the very high values stand out. This can be used effectively with moving averages and aberration detection.

Spot Identification Spot identification is displayed by holding the curser over a data point. The following information is presented:

- Sample name (Array name).
- Description (if Feature Extraction files are used in the experiment, this information will come from the associated Design file).
- Probe name.
- Log ratio of probe in array.

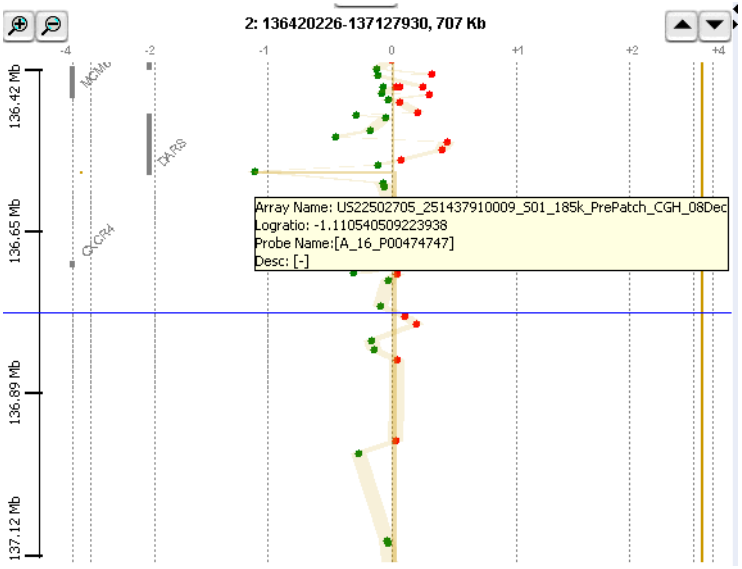


Figure 2-17 Scatter Plot spot identification on hover

CAUTION

Since the scatter plot does not have a facility for automatic differentiation of multiple microarrays beyond spot identification, this is not an optimal mode to use when trying to visualize multiple microarrays at the same time. See Viewing Multiple Microarrays for more information.

Moving Average

Figure 2-18 Moving Average dialog box

Show Select or clear to turn the Moving Average Line Plot **On** (Show) or **Off** (Hide).

Algorithm

- **Linear –** The linear algorithm is the standard method used by the Moving Average approach to smoothing.
- **Triangular –** The triangular algorithm was developed to overcome some potential problems with the linear approach. For information on this algorithm, see "Sample Data" on page 230.

Line Width Specify how thick of a line to plot (in pixels). The range available is 1 to 5 pixels.

Window Set the size of the moving average window. Moving averages can be computed with windows with sizes based on either a specific length of base-pairs (5 Kb to 50 Kb or .1 Mb to 50 Mb) or a fixed number of data points (1 pt. to 60 pt.). A moving average is computed for each point in the data set using a window size centered on that point.

NOTE While both Moving Averages and Z-scores use the window size as input, you can plot them individually or together. To reduce screen clutter, it is often useful to start with a multiple microarray view that shows just the Z-scores, in order to identify the microarrays and chromosomes that are of potential interest. See also Z-Scoring for Aberrant Regions.

Aberration

Figure 2-19 Aberration dialog box

Show Select or clear the check box to apply the chosen aberration algorithm and to turn the Aberration plot **On** (Show) or **Off** (Hide).

Algorithm Select the algorithm to use to detect aberrations. Four algorithms are available:

- Z-score
- ADM-1 (Aberration Detection Module-1)
- ADM-2
- CBS

For details on these algorithms, see Chapter 4, “Statistical Algorithms”.

Threshold / Slider Bar Select the threshold value from the slider bar or type a desired threshold value into the list box. Threshold values from 0.0 to 50.0 are listed.

The optimal threshold values for finding the correct aberration using the Z-score algorithm, ADM-1, and ADM-2 are different, even for the same array. The meaning of threshold also differs for the algorithms.

Combine Replicates

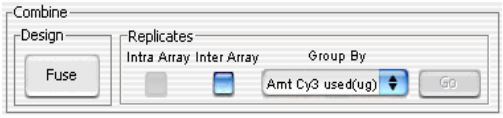


Figure 2-20 Combine Replicates dialog box

Intra Array Select to combine the log ratios for replicated probes in each array using the error model described in Error Model and Combining Replicates. For expression arrays, Intra Array uses the genes to combine replicates. For CGH arrays the probe name is used to combine replicates.

Inter Array Select to use the **Group By** selection field to define replicate probes to combine. The error model is used to combine probes.

Group By Select the attribute for grouping inter array replicate combinations. Twenty-four selections are available:

- Amt Cy3 used (µg)
- Amt Cy5 used (µg)
- Array Fab date
- Array type
- Array Set
- Chip Barcode
- Comments
- Cy3 sample
- Cy5 sample
- Hyb Date
- Hyb temp
- Hyb time
- Hyb'd By
- Labeling Method
- Model System
- Polarity

- Purpose
- QC Metric Status
- Sample
- Sample Name
- Sample Type
- Template
- Wash Conditions
- isMultiPack

Go Click Go to apply the selected options to the selected arrays. Arrays are selected in Tab view by clicking the column header. For more information, see "Tab View" on page 87.

The time required to complete the operation will vary depending on the number of arrays you select.

Tab View

The data for each array in a selected experiment is displayed across the lower portion of the screen in tabular form. This component can be hidden.

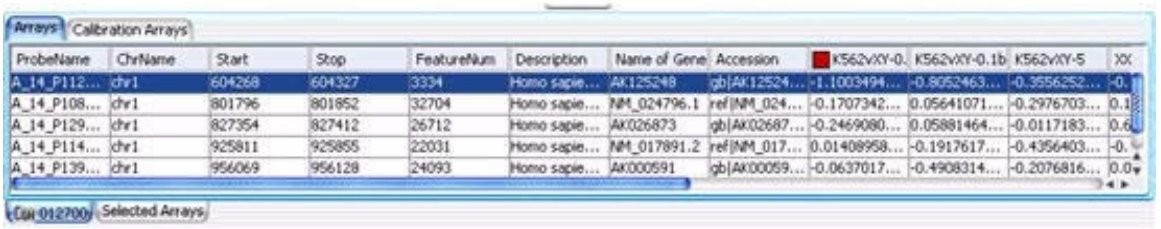


Figure 2-21 Tab view

When you select an experiment, its data is displayed in this tabular view. The column headers for Agilent data include:

- ProbeName
- Chromosome Name
- Starting point for the gene expression
- Stopping point for the gene expression
- Feature Number
- Description of source
- Gene Name
- Accession number
- One or more columns following the Accession column contain information on the specific microarrays in the experiment. The microarrays preceded by a colored box are Selected Arrays.

Column 5 and higher may display different headers with non-Agilent data.

Tabs Above the main columnar table there are two tabs:

- **Arrays** – Displays all arrays in the table.
- **Calibration Arrays** – Displays only the arrays used for calibration.

Below the main table, there are two or more tabs:

- **Numbered tabs** – Represents experiments and when clicked will display the microarrays in that experiment.
- **Selected Arrays** – Displays only the microarrays that show a colored square indicating they have been selected.

Table Header Shortcut Menus

Table headers display four types of shortcut menus in Tab View. Right-clicking any column header (columns 1–8) will display a **Scroll to Column** command button. If you click the command, the Scroll to Column dialog box appears to facilitate moving directly to any selected column based on its column header. See Figure 3-51 on page 153.

If you right-click an column header in the microarray portion of the table (columns 9+), a drop-down menu gives you several choices for manipulating the microarrays.

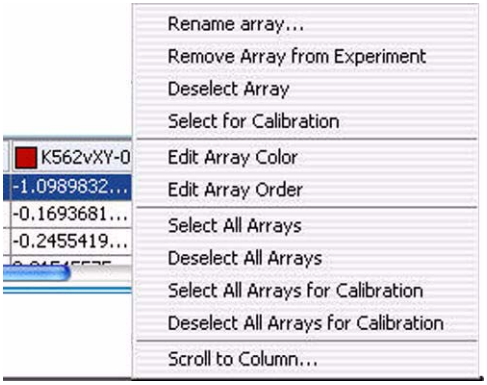


Figure 2-22 Tab view column drop-down menu

NOTE This menu can change depending on the status of the selected item. For example, the menu for an individual selected array will display Deselect Array, but if the array was not a selected array, the menu will display Select Array. Similarly, the menu changes for individual arrays that are Selected/Deselected for Calibration.

Rename array	Renames the highlighted microarray to a convenient name *in that experiment*. The original name of the microarray remains unchanged.
Remove Array from Experiment	Delete the selected microarray from the experiment and remove that column from the table.
Select / Deselect Array (toggle)	Select the microarray and mark it with a colored box. The color is same as that of the data line plots and aberration calls for that microarray in the display windows. If an array is already selected, this option cancels its selection.
Select / Deselect for Calibration (toggle)	Include the selected microarray in the list of microarrays used in calibration. If the microarray is already selected, this option removes it from the calibration list.
Edit Array Color	Display the Select Color dialog box. A tabbed color selector is provided to use for specifying the colors representing the microarrays. See Figure 3-53 on page 155.
Edit Array Order	Display the Edit Array Order dialog box where you can change the order of the microarrays in a specific design. See Figure 3-29 on page 125.
Select All Arrays	Select and mark all microarrays with appropriate colored boxes.
Deselect All Arrays	Cancel the selection of the microarrays and removes the colored boxes.
Select All Arrays for Calibration	Highlight and move all arrays from the **Experiments** folder to the **Calibration** folder.
Deselect All Arrays for Calibration	Cancel the selection of all microarrays previously selected for calibration and return them to the tabular view.
Scroll to Column	Select a column and scroll directly to it. This is useful when working with a large set of microarrays, since it may be cumbersome to scroll manually.
Web Search	If you right-click a tabular cell below any column header, a Web Search dialog box displays where you can search for more information on the microarray. See "Web Searching" on page 235.

Status View

The analysis status for the CGH Analytics application and the active experiment is displayed across the bottom of the screen for quick reference.

Figure 2-23 Status view

The columns displayed from left to right:

- Selected chromosome and cursor position.
- Feature level filter (if applied).
- Array level filter (if applied).
- Aberration filters (if applied).
- Status of centralization algorithm (bold ‘C’ indicates algorithm is applied).
- Status of fuzzy zero correction (bold ‘F’ indicates correction is applied).
- Current active gene list.
- Current selected genome build.
- Current selected source data type.
- Selected Row in the table (the table is in the tab view and is a delimited listing of candidate aberration intervals sorted by sample).
- Number of rows and columns in the table.

NOTE Status bar items can be identified at any time by holding your mouse over the column to be queried. A pop-up box will display the contents of the status column.

3 Dialog Box Reference

CGH Analytics allows a variety of file imports, analysis and visualization of data, and summary report data file generation. This section provides illustrations of the major dialog boxes that you will encounter when using CGH Analytics 3.4. The sections are arranged by workflow and correlate to the order of procedural instructions found in Chapter 1, “How-To”. Within each section, the dialog box references are arranged alphabetically.

Importing Data

This section provides illustrations of the major dialog boxes that you will encounter when using CGH Analytics 3.4 to import data. The dialog boxes are arranged alphabetically.

Import Design Files

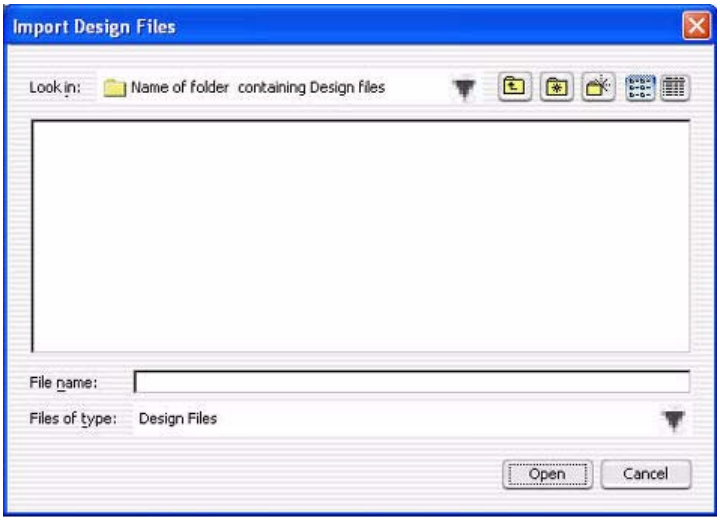


Figure 3-1 Import Design Files dialog box

Look in Displays the name of a selected folder containing Design files. Click the down arrow to select a different folder from your system's directory.

File Navigation icons Click to move to a higher level folder, return to the desktop, create a new folder, or change the way folder and file names are displayed in the in the list box.

List Box Displays the names of folders and files of the indicated type contained in the selected folder. Click to select a folder or file.

File name Displays the name of the selected file.

Files of type Displays the type of files displayed in the list box. In this case, the default type is **Design Files** format. Click the down arrow to select **All Files**.

Open Open the selected file.

Cancel Cancel the selection and close the dialog box.

Import Experiments

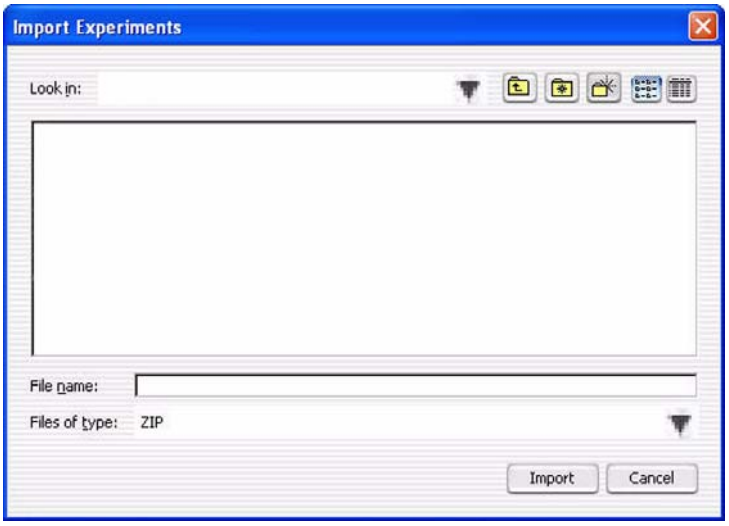


Figure 3-2 Import Experiments dialog box

Look in Displays the name of a selected folder containing Experiment files. Click the down arrow to select a different folder from your system's directory.

File Navigation icons Click to move to a higher level folder, return to the desktop, create a new folder, or change the way folder and file names are displayed in the in the list box.

List Box Displays the names of folders and files of the indicated type contained in the selected folder. Click to select a folder or file.

File name Displays the name of the selected file.

Files of type Displays the type of files displayed in the list box. In this case, the default type is **ZIP** file format. Click the down arrow to select **All Files**.

Import Import the selected file

Cancel Cancel the selection and close the dialog box.

Import FE Files

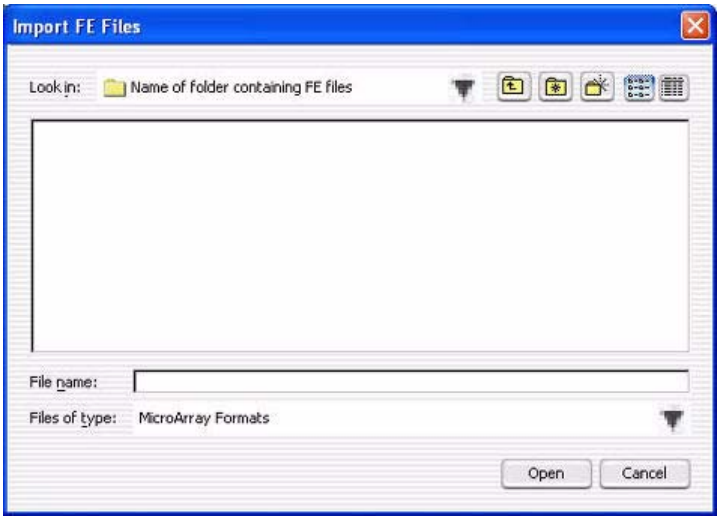


Figure 3-3 Import Feature Extraction Files dialog box

Look in Displays the name of a selected folder containing Feature Extraction files. Click the down arrow to select a different folder from your system's directory.

File Navigation icons Click to move to a higher level folder, return to the desktop, create a new folder, or change the way folder and file names are displayed in the in the list box.

List Box Displays the names of folders and files of the indicated type contained in the selected folder. Click to select a folder or file.

File name Displays the name of the selected file.

Files of type Displays the type of files displayed in the list box. In this case, the default type is **Microarray Format** files. Click the down arrow to select **All Files**.

Open Open the selected file.

Cancel Cancel the selection and close the dialog box.

Import Filters

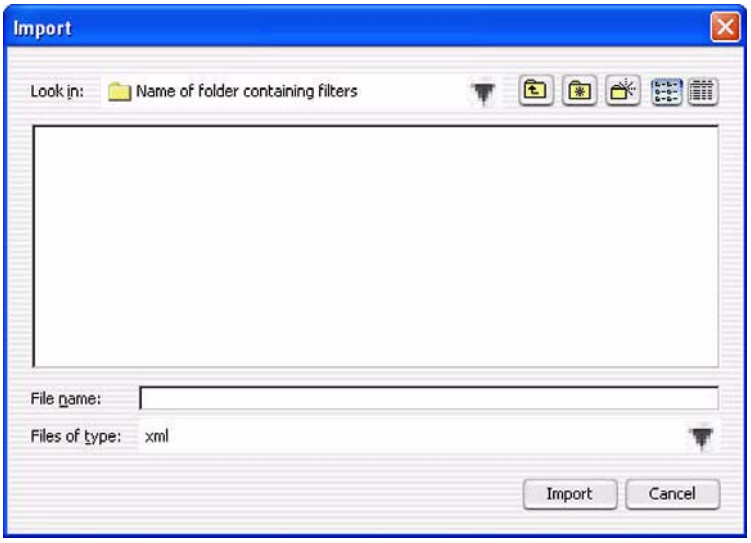


Figure 3-4 Import filters dialog box

Look in Displays the name of a selected folder containing filter files. Click the down arrow to select a different folder from your system's directory.

File Navigation icons Click to move to a higher level folder, return to the desktop, create a new folder, or change the way folder and file names are displayed in the in the list box.

List Box Displays the names of folders and files of the indicated type contained in the selected folder. Click to select a folder or file.

File name Displays the name of the selected file.

Files of type Displays the type of files displayed in the list box. In this case, the default type is **.xml** file format. Click the down arrow to select **All Files**.

Import Import the selected file.

Cancel Cancel the selection and close the dialog box.

Import CGH Analytics 3.1 Experiments

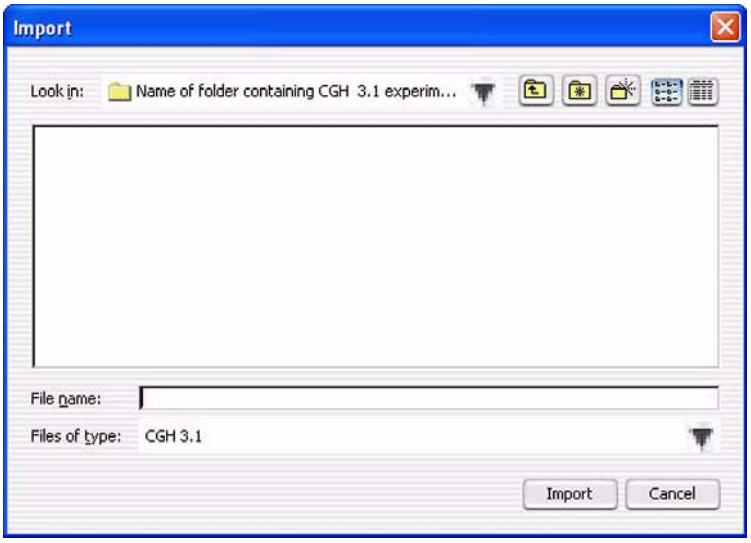


Figure 3-5 Import CGH 3.1 experiments dialog box

Look in Displays the name of a folder containing CGH 3.1 experiment files. Click the down arrow to select a different folder from your system's directory.

File Navigation icons Click to move to a higher level folder, return to the desktop, create a new folder, or change the way folder and file names are displayed in the in the list box.

List Box Displays the names of folders and files of the indicated type contained in the selected folder. Click to select a folder or file.

File name Displays the name of the selected file.

Files of type Displays the type of files displayed in the list box. In this case, the default type is **CGH 3.1** file format. Click the down arrow to select **All Files**.

Import Import the selected file.

Cancel Cancel the selection and close the dialog box.

Import Genome Build

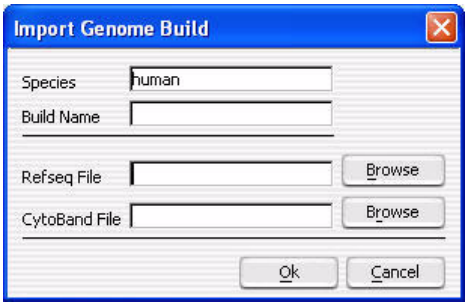


Figure 3-6 Import Genome Build dialog box

Species Specify the genome's species of origin. Human, rat, or mouse genomes are provided with this application.

Build Name The name of the build being imported.

Refseq File The Reference Sequence file accession number from the NCBI gene list.

CytoBand File The UCSC file designation.

OK Accept the selection and import the build.

Cancel Cancel the selection and close the dialog box.

CAUTION Currently, Agilent provides genome-build files for use with CGH Analytics. Do not import any file not provided by Agilent.

Import Microarray Attributes

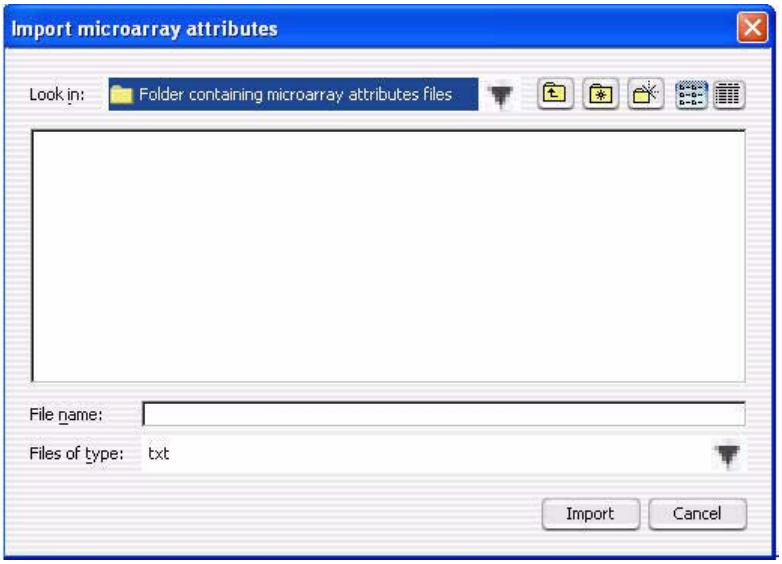


Figure 3-7 Import microarray attributes dialog box

Look in Displays the name of a selected folder containing Microarray Attributes files. Click the down arrow to select a different folder from your system's directory.

File Navigation icons Click to move to a higher level folder, return to the desktop, create a new folder, or change the way folder and file names are displayed in the in the list box.

List Box Displays the names of folders and files of the indicated type contained in the selected folder. Click to select a folder or file.

File name Displays the name of the selected file.

Files of type Displays the type of files displayed in the list box. In this case, the default type is **txt** file format. Click the down arrow to select **All Files**.

Import Accept the selection and import the attributes.

Cancel Cancel the selection and close the dialog box.

Import UDF Files

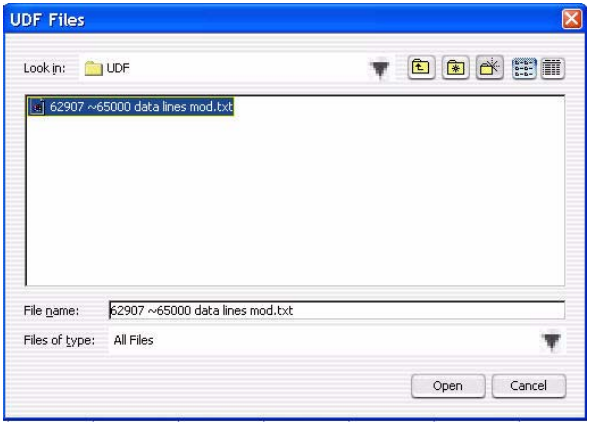


Figure 3-8 UDF Files dialog box

Look in Displays the name of a selected folder containing UDF files. Click the down arrow to select a different folder from your system's directory.

File Navigation icons Click to move to a higher level folder, return to the desktop, create a new folder, or change the way folder and file names are displayed in the in the list box.

List Box Displays the names of folders and files of the indicated type contained in the selected folder. UDF files are .txt files. Click to select a folder or file.

File name Displays the name of the selected file.

Files of type Displays the type of files displayed in the list box. In this case, the default type is **all** file types.

Open Accept the selection and import the UDF file.

Cancel Cancel the selection and close the dialog box.

Select Data Type for Experiments

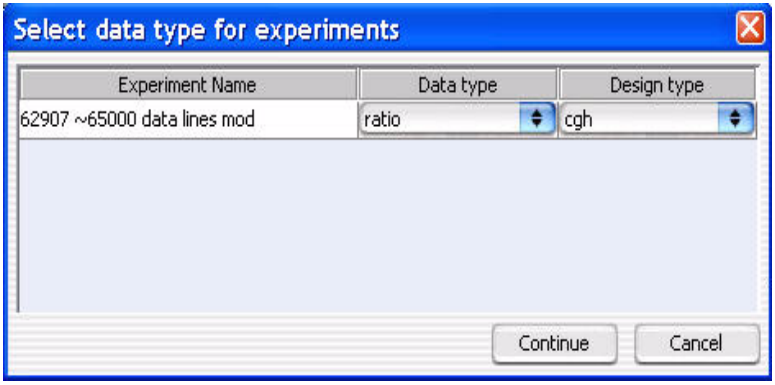


Figure 3-9 Select Data Type for Experiments dialog box

Data Type Allows specification of the ratio type for the data file import. There are four choices:

- **Ratio**
- **Log2 ratio**
- **Log10 ratio**
- **ln ratio**

Design Type Allows specification of the ratio type for the data file import. There are two choices:

- **CGH**
- **Expression**

Continue Accept the selection and import the data file with type specifications.

Cancel Cancel all selections and close the dialog box.

Select Report File

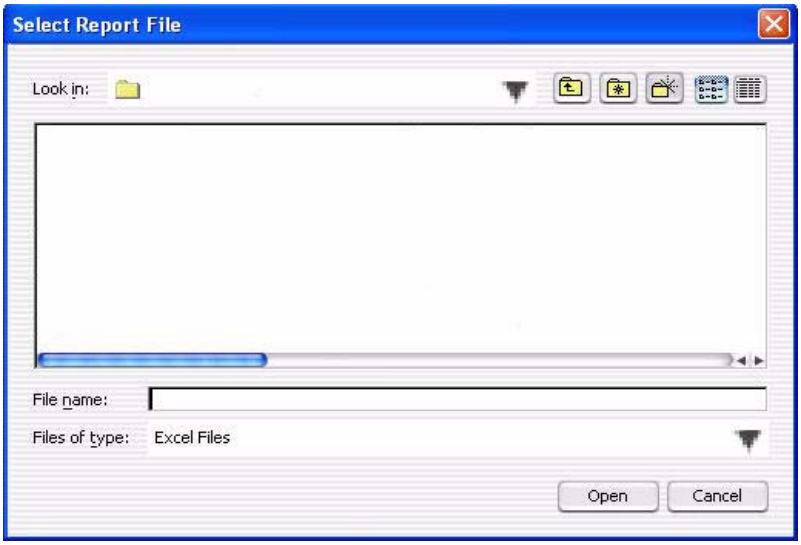


Figure 3-10 Select Report File dialog box

Look in Displays the name of a selected folder containing Report Files. Click the down arrow to select a different folder from your system's directory.

File Navigation icons Click to move to a higher level folder, return to the desktop, create a new folder, or change the way folder and file names are displayed in the in the list box.

List Box Displays the names of folders and files of the indicated type contained in the selected folder. Click to select a folder or file.

File name Displays the name of the selected file.

Files of type Displays the type of files displayed in the list box. In this case, the default type is **Excel File** format. Click the down arrow to select **All Files**.

Open Open the selected file.

Cancel Cancels your selections and close the dialog box.

UDF Import Summary

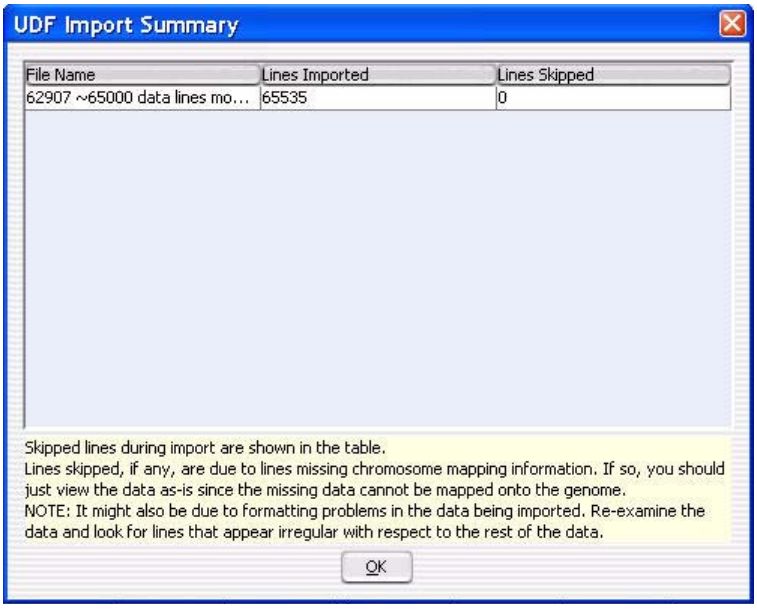


Figure 3-11 UDF Import Summary dialog box

List Box Displays the file name of the UDF file imported, the number of lines successfully imported, and which lines, if any, were skipped during import. Skipped lines may be due to lines missing chromosome mapping information, or may be the result of improper formatting of the UDF file.

OK Click **OK** to accept the summary report

Universal Data Importer - Map Column Headers

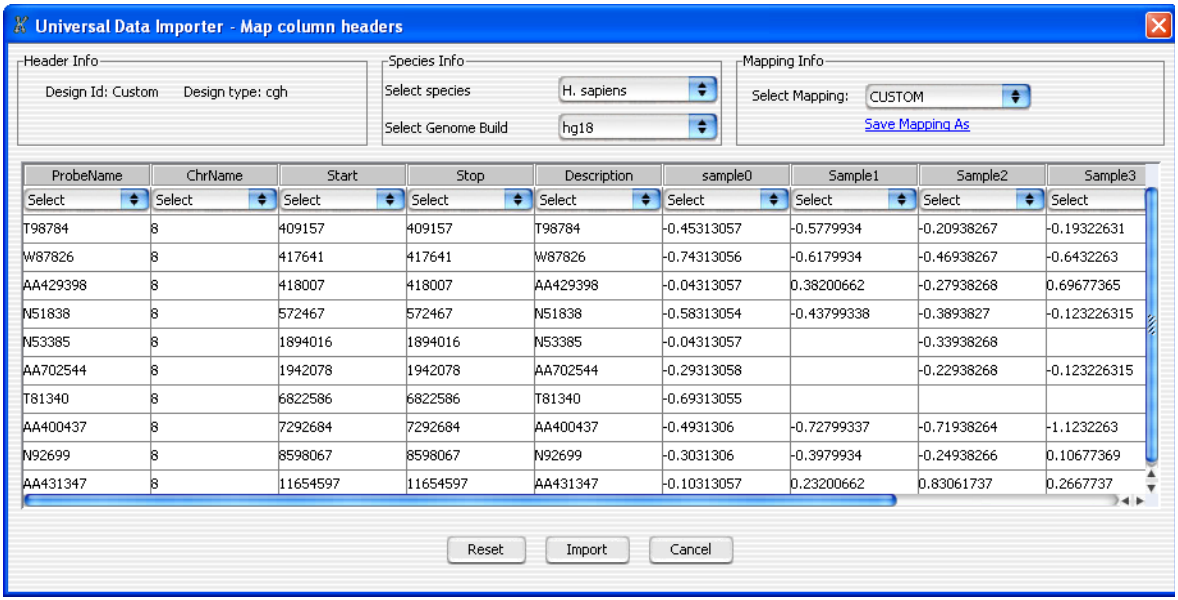


Figure 3-12 Universal Data Importer - Map Column Headers dialog box

Species Info Contains two parameters:

Select Species Set the array species. There are three choices:

- **M. musculus**
- **H. sapiens**
- **R. norvegicus**

Select Genome Build Specifies the species specific build to use.

Mapping Info Allows using a stored mapping already created or saving the mapping settings as a new mapping.

Select Mapping Set the mapping properties to those of a previously created mapping.

Save Mapping As Save the settings as a new mapping. Opens a dialog box to accept a new mapping name.

Table Allows selection of headers from those provided for each column. If the data is from Agilent CGH array(s), the correct header entry will be displayed on the top row of the table.

These headers are mandatory and must be labeled in specific order from left to right with the six specific headers provided:

- The first column (left most column) header must be: ProbeName
- The second column header must be: ChrName (chromosome)
- The third column header must be: Start
- The fourth column header must be: Stop
- The fifth column header must be: Description
- The sixth column (right most column) header must be: LogRatio

NOTE One or more columns following the LogRatio column may contain signal values for each of the specific arrays.

Reset Resets the properties in this dialog box.

Import Imports the UDF file with specified parameters.

Cancel Cancels all selections and closes the dialog box.

Experiment Creation and Modification

This section provides illustrations of the major dialog boxes that you will encounter when creating Experiments and modifying associated Attributes. The dialog boxes are arranged alphabetically.

Array Set: Fuse Experiment

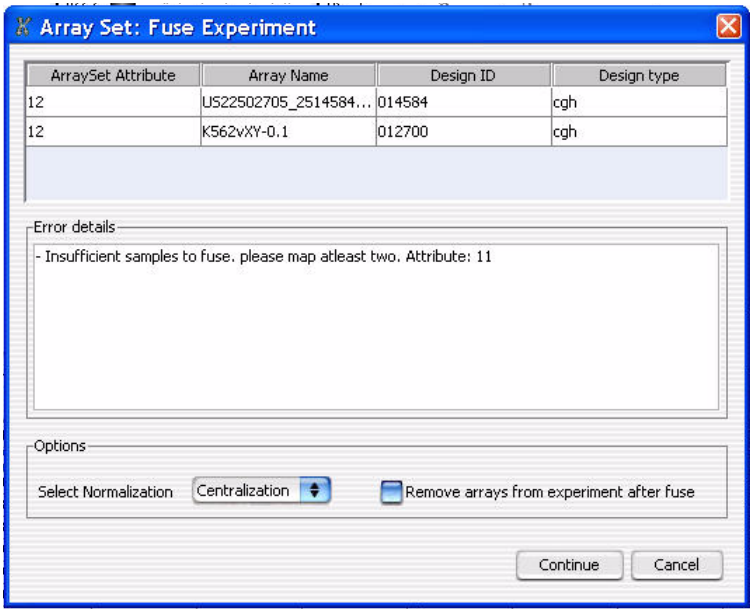


Figure 3-13 Array Set: Fuse Experiment dialog box

Select Normalization Allows selection of Centralization algorithm.

Remove Arrays after Fusion You can check the box for 'Remove arrays from experiment after fuse' that may aid in reducing duplication of information.

Continue Click on **Continue** to fuse experiments with selected options:

Cancel Cancels any selections and closes the dialog box.

Create Experiment

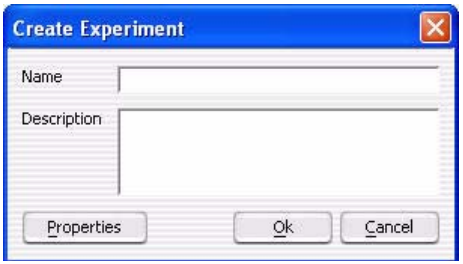


Figure 3-14 Create Experiment dialog box

Name Type a name for your new experiment.

Description Briefly, describe your experiment with information that will help you identify it.

Properties Click to access the Experiment Properties dialog box where you can select microarrays to populate your new experiment. See Figure 3-34 on page 131.

NOTE Do not click **OK** until you have populated your experiment in the Experiment Properties dialog box or you will have an empty experiment.

OK Accept your new experiment and add it to your list of experiments in the Experiments folder.

Cancel Close the dialog box without change.

Microarray Properties - Application Attributes

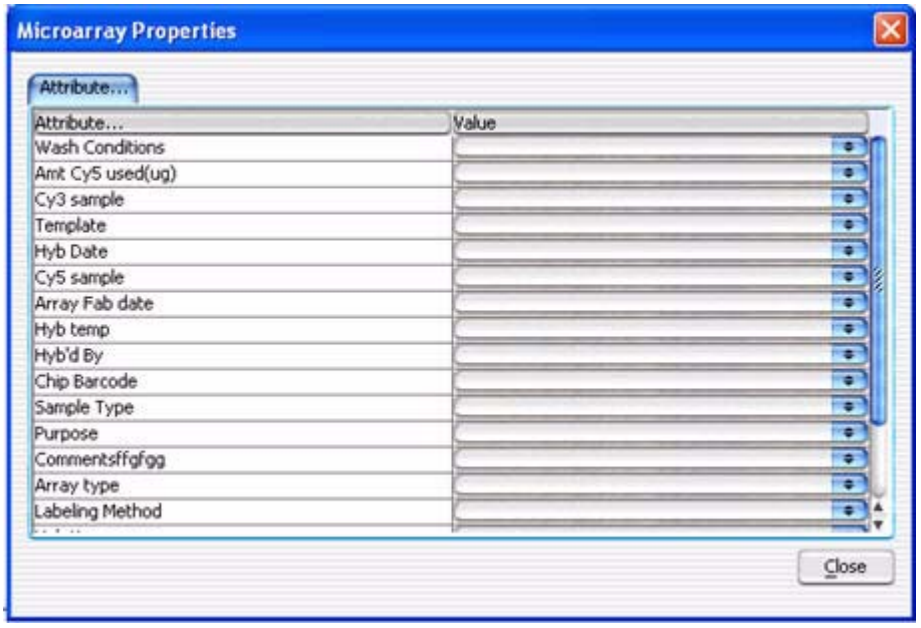


Figure 3-15 Microarray Properties dialog box showing attributes available in the application

Attribute Tab

- **Attribute** — Lists the attribute name.
- **Value** – Indicates the values for each attribute, if any, for each array.

Close Closes the dialog box.

Microarray Properties - Attributes

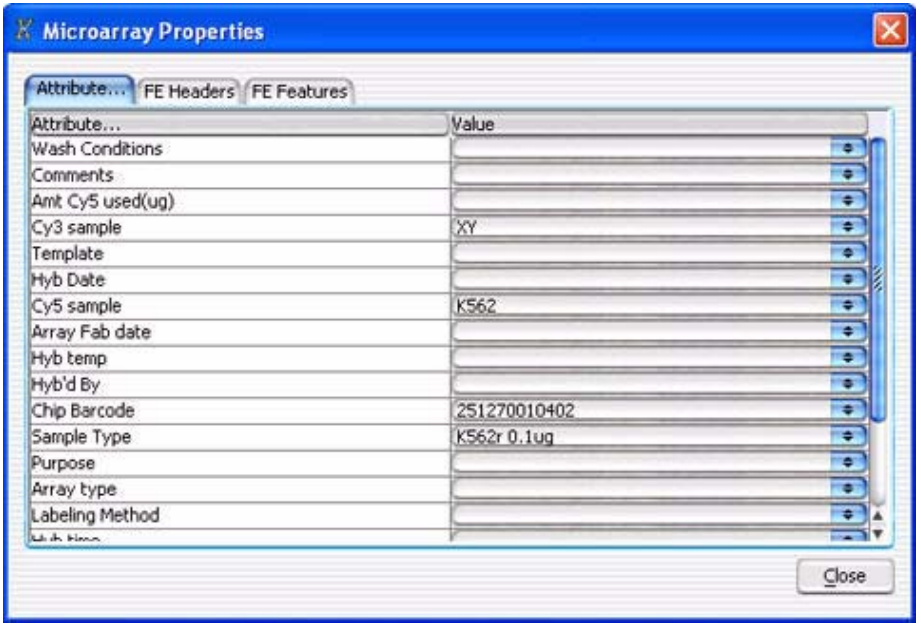


Figure 3-16 Microarray Properties dialog box listing Attributes and their values

Attribute Tab

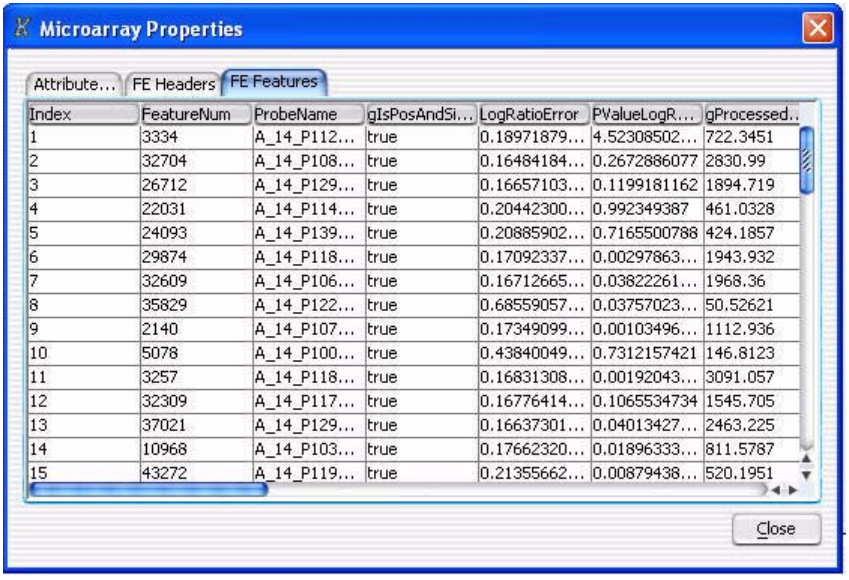
- **Attribute** – Lists the attributed in an array by name.
- **Value** – Indicates the values, if any, for each array.

Close Closes the dialog box.

For the FE Headers tab options, see Figure 3-18 on page 111.

For the FE Features tab options, see Figure 3-17 on page 110.

Microarray Properties - FE Features



Index	FeatureNum	ProbeName	gIsPosAndSi...	LogRatioError	PValueLogR...	gProcessed..
1	3334	A_14_P112...	true	0.18971879...	4.52308502...	722.3451
2	32704	A_14_P108...	true	0.16484184...	0.2672886077	2830.99
3	26712	A_14_P129...	true	0.16657103...	0.1199181162	1894.719
4	22031	A_14_P114...	true	0.20442300...	0.992349387	461.0328
5	24093	A_14_P139...	true	0.20885902...	0.7165500788	424.1857
6	29874	A_14_P118...	true	0.17092337...	0.00297863...	1943.932
7	32609	A_14_P106...	true	0.16712665...	0.03822261...	1968.36
8	35829	A_14_P122...	true	0.68559057...	0.03757023...	50.52621
9	2140	A_14_P107...	true	0.17349099...	0.00103496...	1112.936
10	5078	A_14_P100...	true	0.43840049...	0.7312157421	146.8123
11	3257	A_14_P118...	true	0.16831308...	0.00192043...	3091.057
12	32309	A_14_P117...	true	0.16776414...	0.1065534734	1545.705
13	37021	A_14_P129...	true	0.16637301...	0.04013427...	2463.225
14	10968	A_14_P103...	true	0.17662320...	0.01896333...	811.5787
15	43272	A_14_P119...	true	0.21355662...	0.00879438...	520.1951

Figure 3-17 Microarray Properties dialog box listing FE Features and associated data

List Box Displays columnar listing of FE features and the associated data. The fields are:

Index	FeatureNum	ProbeName
GIsPosAndSignif	LogRatioError	PValueLogRatio
gProcessedSignal	rProcessedSignal	gMedianSignal
rMedianSignal	gBGSubSignal	rBGSubSignal
gIsSaturated	rIsSaturated	gIsFeatNonUnifOL
rIsFeatNonUnifOL	gIsBGNonUnifOL	rIsBGNonUnifOL
rIsPosAndSignif	gIsWellAboveBG	rIsWellAboveBG

Microarray Properties - FE Headers

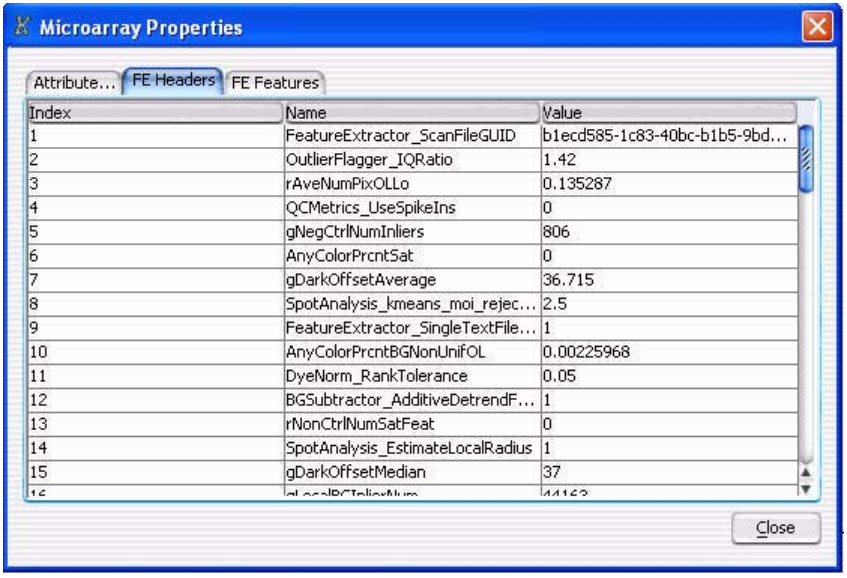


Figure 3-18 Microarray Properties dialog box listing FE Headers their values

Index Displays a sequential index to help identify FE properties.

Name Displays each microarray's name.

Value Displays the value for each microarray.

Close Closes the dialog box.

Set Genome Build

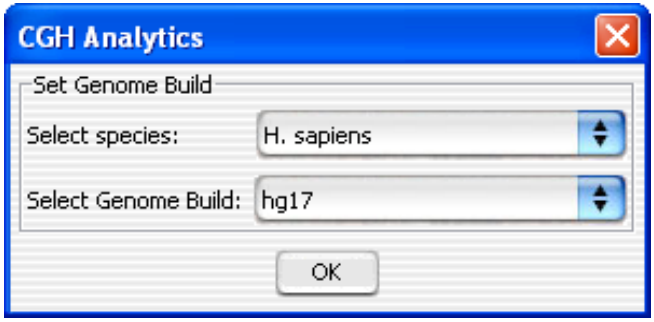


Figure 3-19 Set Genome Build dialog box

Select Species Selects the species for a genome build. There are three options:

- **M. musculus**
- **H. sapiens**
- **R. norvegicus**

Select Genome Build Specifies the species specific build to use.

OK Applies the Genome Build to the selected experiment with multiple array designs.

Analysis and Visualization

This section provides illustrations of the major dialog boxes that you will encounter when analyzing and visualizing your CGH data. The dialog boxes are arranged alphabetically.

Choose Gene List Color

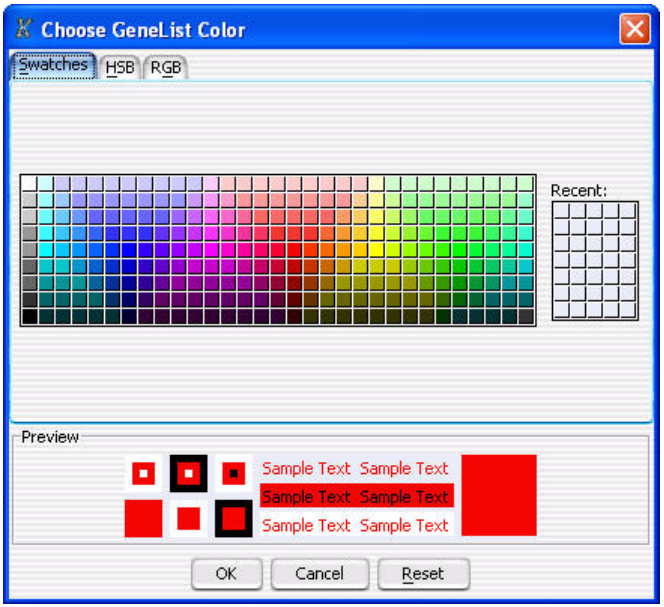


Figure 3-20 Choose GeneList Color dialog box

Swatches Tab Display colors based on color samples (swatches).

HSB Tab Display colors based on an HSB schema (Hue, Saturation, and Brightness or Value).

RGB Tab Display colors based on an RGB schema (Red-Green-Blue).

Recent: Display recent color selections.

Preview Panel Display what the results of the current selections would be.

OK/Cancel Click **OK** to accept the new color selections or **Cancel** to return to existing parameters.

Reset Click **Reset** to return colors to the default color selection.

Common Aberration

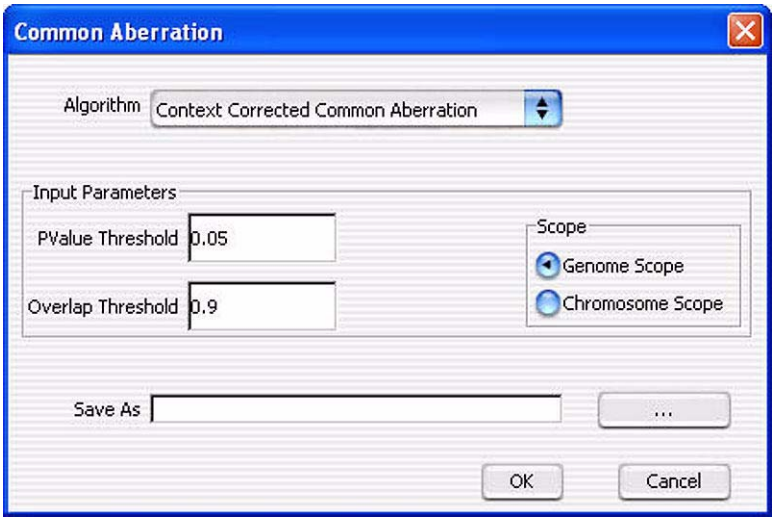


Figure 3-21 Common Aberration dialog box

Common Aberration algorithms are used to try to identify common aberrations over an entire genome.

Algorithm: Select the algorithm to use for calculating the common aberration. The output from these algorithms are almost the same.

- **Context Corrected Common Aberration** – The recommended algorithm to use.
- **T-Test Common Aberration** – Provides results that are less complete, but will work on all samples.

Input Parameters Specify the input parameters for multiple-sample experiments.

- **PValue Threshold** – Sets the pValue limits of the aberrations called.

- **Overlap Threshold** – Sets the limit of overlapping. A threshold of 0.9 covers the entire genome.

Scope Specify the scope of the analysis for Context Corrected Common Aberration algorithm. (By default, the scope of the T-Test Common Aberration algorithm is the entire genome.)

- **Genome Scope** – Results displayed for the entire genome.
- **Chromosomal Scope** – Results displayed on a chromosome-by-chromosome basis.

Save As Type in the name of the report under which the results will be saved, then click [...] to browse to the directory where the file is saved.

OK/Cancel Click **OK** to accept the new parameters or **Cancel** to return to the previously existing parameters.

Common Aberration Details

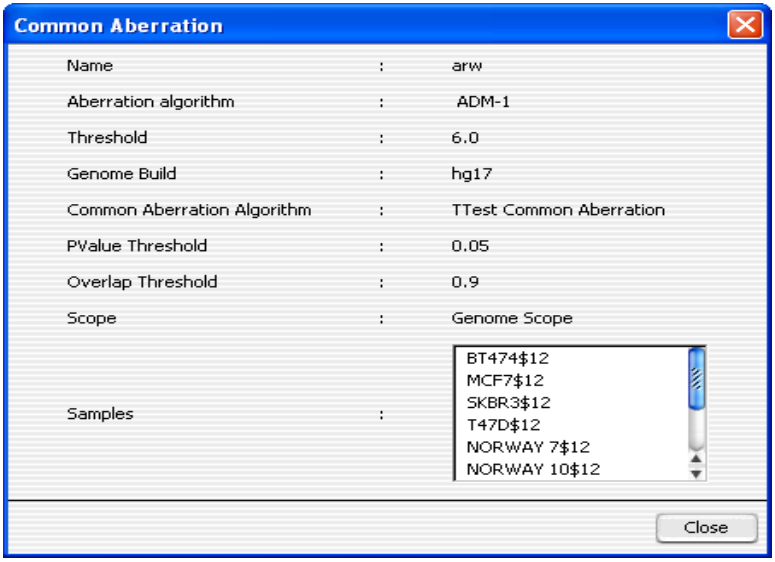


Figure 3-22 Common Aberration details dialog box

Displays an analysis summary about the common aberration. The information displayed includes:

- Name of common aberration node
- The aberration calling algorithm used to create the common aberration
- The threshold applied to the aberration calling algorithm
- The genome build of the samples used in analyzing the common aberration
- The common aberration algorithm used
- The p-value threshold applied to the common aberration algorithm
- The overlap threshold applied to the common aberration algorithm
- The scope of the common aberration analysis
- The sample names used in the common aberration analysis

Close Closes the dialog box and returns to the main window.

Common Aberration Gene List

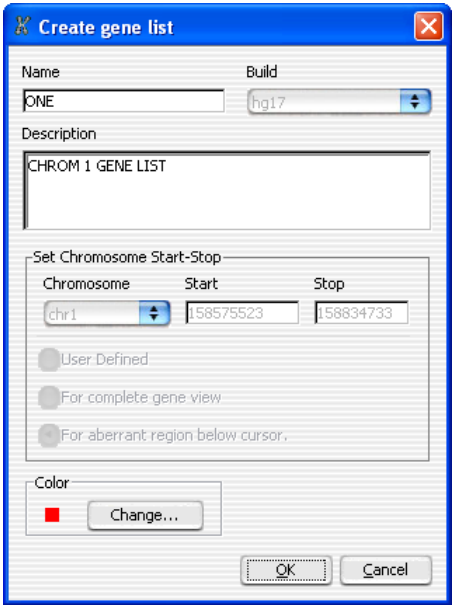


Figure 3-23 Common Aberration Create Gene List dialog box

Common aberrations in shared genomic intervals across samples can be mapped to proximal genes and persisted in a gene list.

Name Input the name of the new common aberration gene list.

Build Shows the current build specified. If Agilent CGH design files are used, this field is protected from changes.

Description Allows descriptive text to be attached to the gene list for reference.

Set Chromosome Start-Stop Allows user defined chromosomal information to be attached to the gene list including chromosome number and relative start and stop positions.

Color Select a color to differentiate gene list from other such lists.

OK Generate the common aberration gene list with parameters.

Cancel Exit the dialog box without saving any parameters.

Correlation Analysis Setup

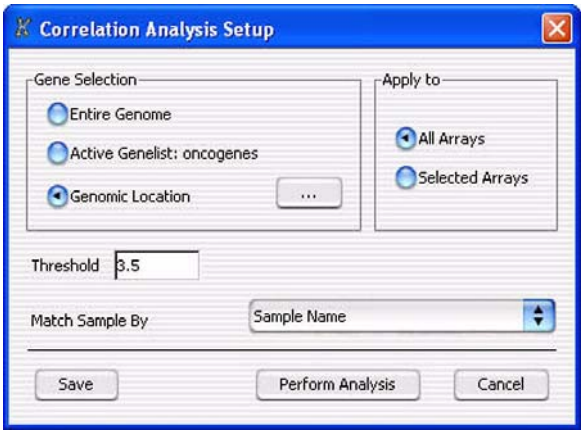


Figure 3-24 Correlation Analysis Setup dialog box

Gene Selection – Specifies the scope of the analysis, and analyze data from:

Entire Genome All of the genome is analyzed.

Active Genelist: oncogenes Arrays from the active gene list which is displayed by default. In this case, **oncogenes**.

Genomic Location Display a Select Chromosome Interval dialog box where you name the chromosome or genome, and their starting and stopping locations. See Figure 3-52 on page 154.

Apply to – Specifies the arrays to be analyzed

All Arrays Apply the correlation algorithm to all microarrays in the experiment.

Selected Arrays Apply the correlation algorithm only to the selected arrays listed in the Tab view.

Threshold Specify the limits between significant and insignificant correlations. The default is 3.5.

Match Sample By Select the attribute that will be used for matching the sample arrays. The default is **Sample Name**.

Perform Analysis Perform the analysis. A Matched Sample dialog box appears showing the Sample Name and the corresponding CGH and gene expression arrays. See Figure 3-42 on page 143.

Save/Cancel Save all changed parameters, or cancel all changes and returns to the previously set parameters.

Correlation Results

Figure 3-25 Correlation Results dialog box

Show entire genome Display the results for the entire genome.

Sort sample by the selected gene Select check box to change the order of the samples based on the expression data.

Show Plot Labels Select check box to show plot labels.

Zooming icons Three buttons control your view: Zoom In, Zoom Out, and Reset Zoom.

Arrow buttons Click the left-pointing arrow to move to the previous chromosome and the right-pointing arrow to move to the next chromosome

Enrichment Analysis Click to go directly to the Enrichment Analysis Setup dialog box. See Figure 3-33 on page 130.

Panes

Gene Expression and Aberration values – The aberration values plotted in relation to a 0.0 no-aberration y-axis. The x-axis on the left denotes Expression Level and on the right Aberration Level.

Correlation result for Amp-nonAmp t-test – The correlation scores displayed will change depending on the selection in the Tab view.

Heatmap for Expression Data – Displays the gene expression values from the expression arrays in a standard heatmap view.

Aberration Summary – Displays the CGH aberration calls from the CGH arrays.

Tab view Displays a tabular view of the output scores resulting from the algorithm.

- Chromosome, Gene Name, Start and Stop:
- Amp-nonapmp t-test – The Student-t-test results for amplification and non-amplification.
- Del-nondel t-test – The Student-t-test results for deletion and non-deletion.
- Pearson p-value test results.
- Max Score – The maximum value displayed from the three other test scores.

Edit Aberration Filters

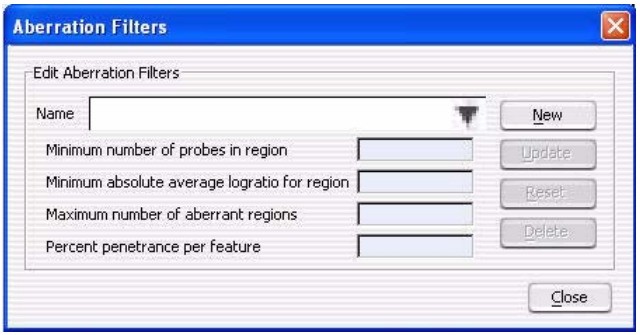


Figure 3-26 Edit Aberration Filters dialog box

Name Name your filter if it is a new filter. If you are changing the parameters of an existing filter, scroll down the list box to find and select the existing filter.

Minimum number of probes Define the minimum number of probes that should be present in an aberrant region. The filter is selected at the application level and applied to the selected experiment.

Minimum absolute average logratio for region Limit the minimum absolute level of average $\log_2$ ratios for a region. In effect, a filter ignores aberrations within a region centered on a $\log_2$ ratio of 0. The extent to either side of 0 is determined by the value entered.

Maximum number of aberrant regions Limit the maximum number of aberrant regions per chromosome. For example, you can set the limit to the three aberrations that have the highest statistical significance.

Percent penetrance per feature Specify the minimum percentage of penetrance per feature across the set of selected arrays. The filter eliminates aberrations that have less than that minimum penetrance across the selected arrays.

New Click to indicate that this is a new filter. It will be added to the list of filters.

Update Click to update an existing filter after changing its parameters.

Reset Click to reset the parameters to default values.

Delete Click to delete this filter.

Close Click to exit this dialog box

Edit Array Color

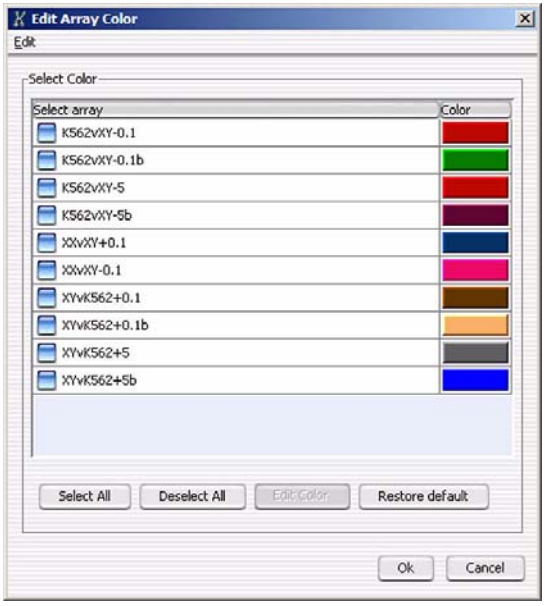


Figure 3-27 Edit Array Color dialog box

Select array Displays a list of the arrays in the sample. Each row in the table represents information on one array. Click the check box to select an array to be changed.

Color Displays the color currently assigned to each array.

Select All Selects all arrays in the list.

Deselect All Cancels the selection of any selected arrays in the list.

Edit Color After one or more arrays are selected, click to access the Select Color dialog box. See Figure 70 on page 129. Once there, select a color to apply to the array's components, and click **Continue**. You are returned to the dialog box where the effect of your changes can be previewed.

NOTE If you **Select All** arrays then select a color in the Select Color dialog box, all components of all arrays will carry the same color.

Restore default Restores the arrays to their default color selections.

OK Click to accept the selections.

Cancel Click to close the dialog box.

Edit Array Level Filters

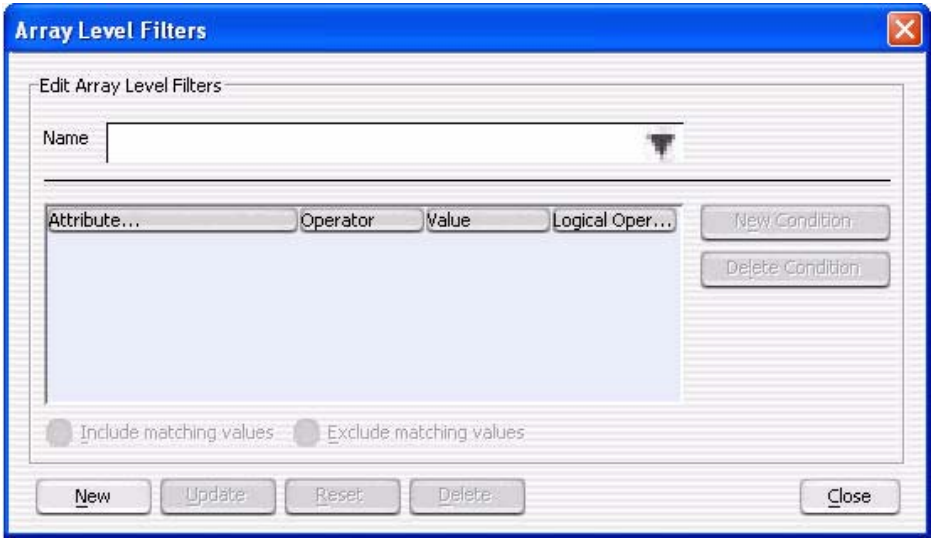


Figure 3-28 Edit Array Level Filters dialog box

Name If this is a new filter, type a name for it. If it is an existing filter, find and select it from the list box.

Attribute List Box

- **Attribute** – A list of named attributes associated with the filter.

Type appropriate entries for Operator, Value, and Logical Operator in their respective fields, or select the values from their drop-down lists, if available.

- Operator – Operators are =, <, <=, >, >=.
- Value – Values are True/False.
- Logical Operator – Logical Operators are AND/OR.

New Condition Add a blank entry to an existing list of attributes, then allows you to select an attribute from the drop-down list.

You can type appropriate entries for Operator, Value, and Logical Operator in their respective fields, or select the values from their drop-down lists, if available.

Delete Condition Select a condition and click to delete it from the filter.

Include matching values Include arrays that have matching attribute values.

Exclude matching values Exclude arrays that have matching attribute values.

New Specify that this is a new filter. It is added to the list of filters.

Update Update a filter after making changes in its parameters

Reset Reset a filter to its default values.

Delete Delete a filter.

Close End the filter-editing procedure.

Edit Array Order

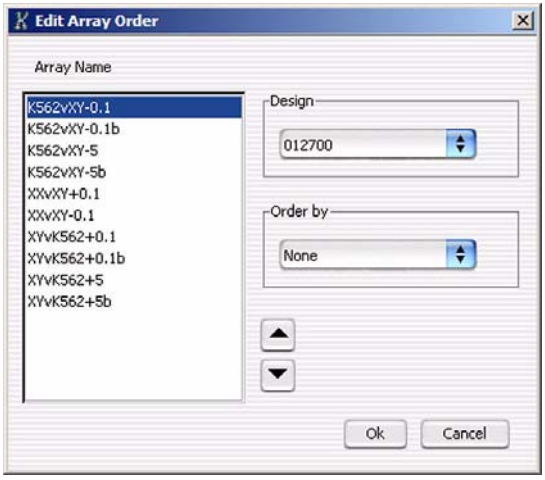


Figure 3-29 Edit Array Order dialog box

Array Name The arrays in the selected design listed in the order used in the Navigator area.

Design Select a design from the scroll-down list.

Order by A list box showing the attributes related to the arrays. Select an attribute to use for ordering the list. The arrays are re-ordered based on their respective values for that attribute. For example, in the arrays displayed, the list box showed:

- None
- Array type
- Chip Barcode
- Polarity
- QC Metric Status
- Sample
- Wash Conditions

OK Click **OK** and the arrays are re-arranged in the Navigator and Tab views.

Cancel Click **Cancel** to close the dialog box with the previous order unchanged.

Edit Attributes

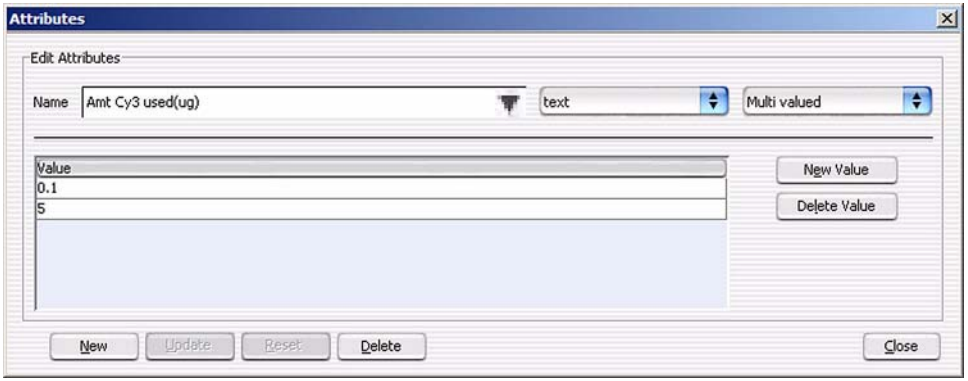


Figure 3-30 Edit Attributes dialog box

Name Type a name if this is a new attribute, or find and select an existing attribute from the list box.

Text List Box Select type of text:

- int – integer
- double – a high-decision floating point number
- boolean – True or False
- text – Plain text

Value List Box Select the type of value:

- Multi valued – Having more than one value.
- Single valued – Having one value
- Dynamic valued – Value at run time

Value Text Box Displays list of attribute: values.

New Value Add a new attribute value

Delete Value Delete an existing attribute value.

New Accept the attribute as new and add it to the list of attributes.

Update Update existing attributes to include any changes you made.

Reset Return a selected attribute to its default values.

Delete Delete a selected attribute.

Close Close the dialog box.

Edit Feature Level Filters

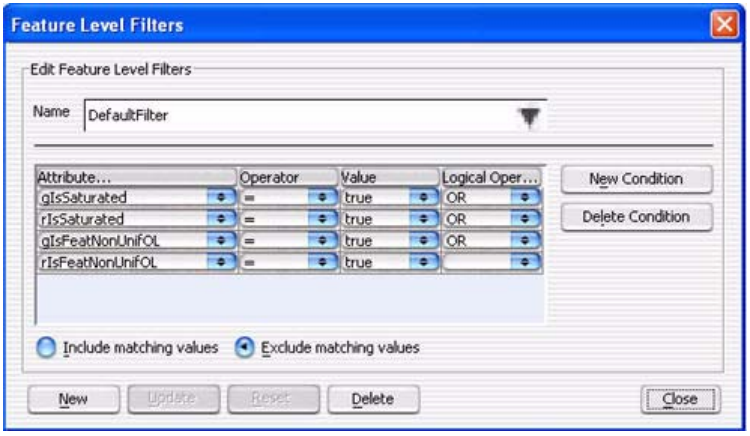


Figure 3-31 Edit Feature Level Filters dialog box

Name If this is a new filter, type a name for it. If it is an existing filter, find and select it from the list box.

Attribute List Box

- **Attribute** – A list of named attributes associated with the filter.

Type appropriate entries for Operator, Value, and Logical Operator in their respective fields, or select the values from their drop-down lists, if available.

- Operator – Operators are =, <, <=, >, >=.
- Value – Values are True/False.
- Logical Operator – Logical Operators are AND/OR.

New Condition Add a blank entry to an existing list of attributes, then allows you to select an attribute from the drop-down list.

	You can type appropriate entries for Operator, Value, and Logical Operator in their respective fields, or select the values from their drop-down lists, if available.
Delete Condition	Select a condition and click to delete it from the filter.
Include matching values	Include arrays that have matching attribute values.
Exclude matching values	Exclude arrays that have matching attribute values.
New	Specify that this is a new filter. It is added to the list of filters.
Update	Update a filter after making changes in its parameters
Reset	Reset a filter to its default values.
Delete	Delete a filter.
Close	End the filter-editing procedure.

Enrichment Analysis Result

Figure 3-32 Enrichment Analysis Result dialog box

Show Entire Genome Displays all of the chromosomes in the genome.

Show Plot Labels Displays labels for each of the plot windows. Even when selected, the axis values might not display in window that is not expanded.

Heatmap for Expression data Displays the expression data in a standard heatmap view where the magnitude of the log ratio is displayed by changing the intensity of the color.

Enriched Intervals Displays the intervals that are enriched.

Tab view
- Tabs for all microarrays in the sample
- Enrichment of high scores
- Enrichment of low scores

Enrichment Analysis Setup

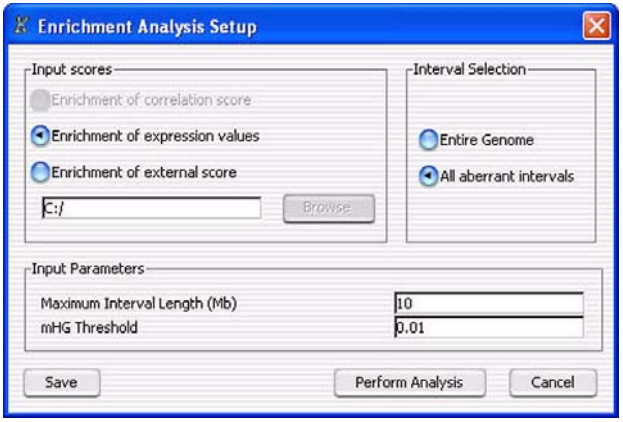


Figure 3-33 Enrichment Analysis Setup dialog box

Input scores
- Enrichment of correlation score
- Enrichment of expression values
- Enrichment of External score

Interval Selection
- Entire Genome
- All aberrant intervals

Input Parameters
- Maximum Interval Length (Mb)
- mHG Threshold

Perform Analysis Starts the process for detecting enriched intervals.

Save/Cancel Saves the parameters used for the analysis or cancels the analysis.

Experiment Properties

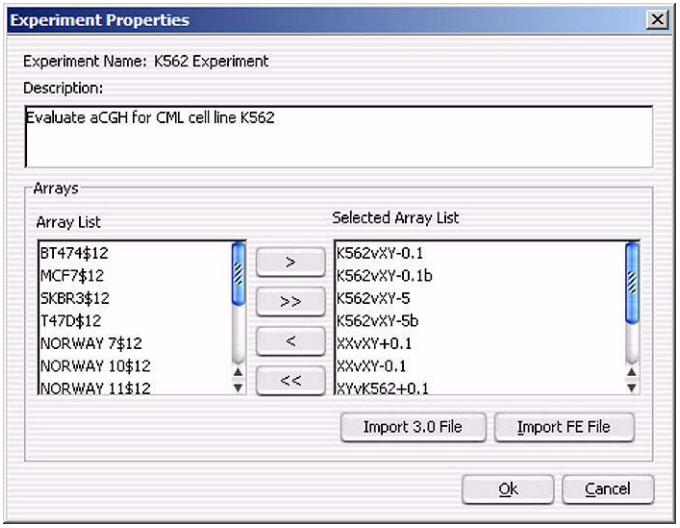


Figure 3-34 Experiment Properties dialog box

Experiment Name: The name of the selected experiment is displayed automatically.

Description Text Box: Displays the description of the experiment that was entered when the experiment was created.

Arrays

- Array List – A list of arrays that are available for this experiment.
- Selected Array List – A list of the arrays that you have selected for this experiment.
- Operators used to move files between Array List and Selected Array List list boxes:
  - > – Move selected file(s) from left to right list box.
  - >> – Move all files from left to right list box.
  - < – Move selected file(s) from right to left list box.
  - << – Move all files from right to lef3.1t list box.

- Import 3.1 File – Click to import CGH 3.1 files. A Confirm dialog box appears affirming that you opted to add a 3.1 array to a non 3.1 experiment and doing so will disable intra array replicates and remove any applied feature-level filters. Click OK and the Import CGH 3.1 File dialog box displays. See Figure on page 97.
- Import FE File – Click to import Feature Extraction files. The Import FE File dialog box appears. See Figure 3-3 on page 95.

OK Accept the selections and changes.

Cancel Leave the properties unchanged and close the dialog box.

Go To Gene/Genomic Location

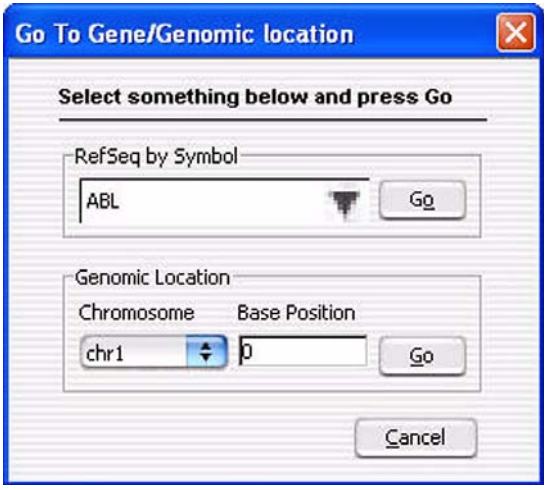


Figure 3-35 Go To Gene/Genomic location dialog box

RefSeq by Symbol Displays the Reference Sequence accession symbol from NCBI.

Genomic Location

- Chromosome – The chromosome number.
- Base Position – The position on the chromosome

Cancel Closes the dialog box.

Graphical Aberration Summary

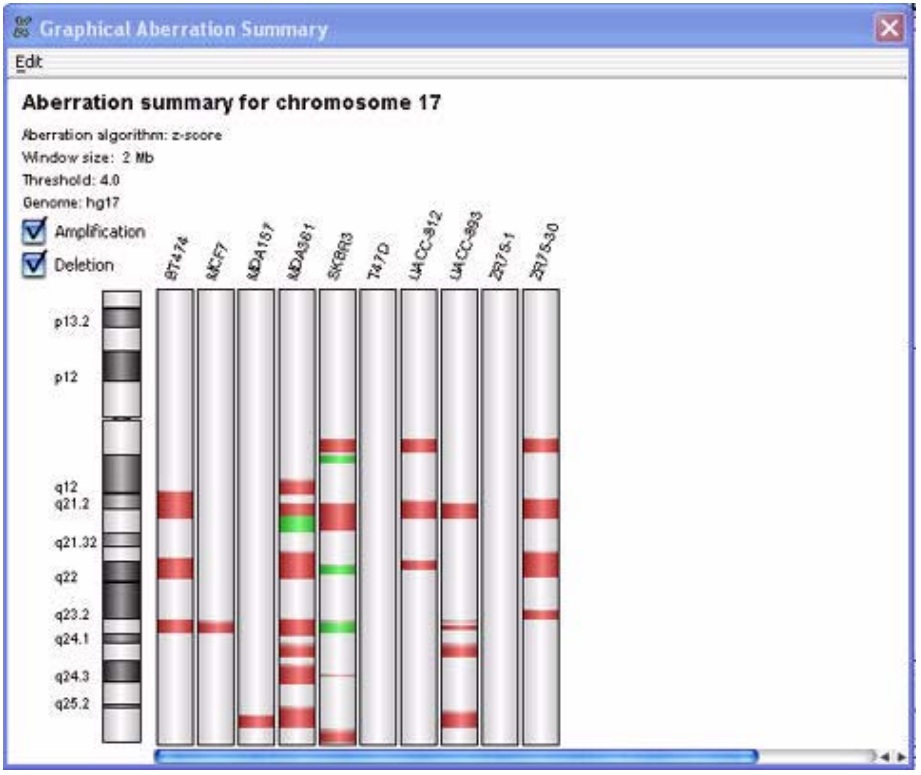


Figure 3-36 Graphical Aberration Summary pane

Rather than showing line or scatter plots, this summary colors significant aberrations in heat-map fashion (red for putative amplifications and green for putative deletions). The scoring is the same as it is in the main display. See "Z-Scoring for Aberrant Regions" on page 161.

Edit Copies the Graphical Aberration Summary to the clipboard (Windows only).

Amplification Select the check box to display amplified chromosomal regions in the samples analyzed in the selected arrays.

Deletion Select the check box to display deleted chromosome regions in the sample of the selected arrays.

Graphical Common Aberration

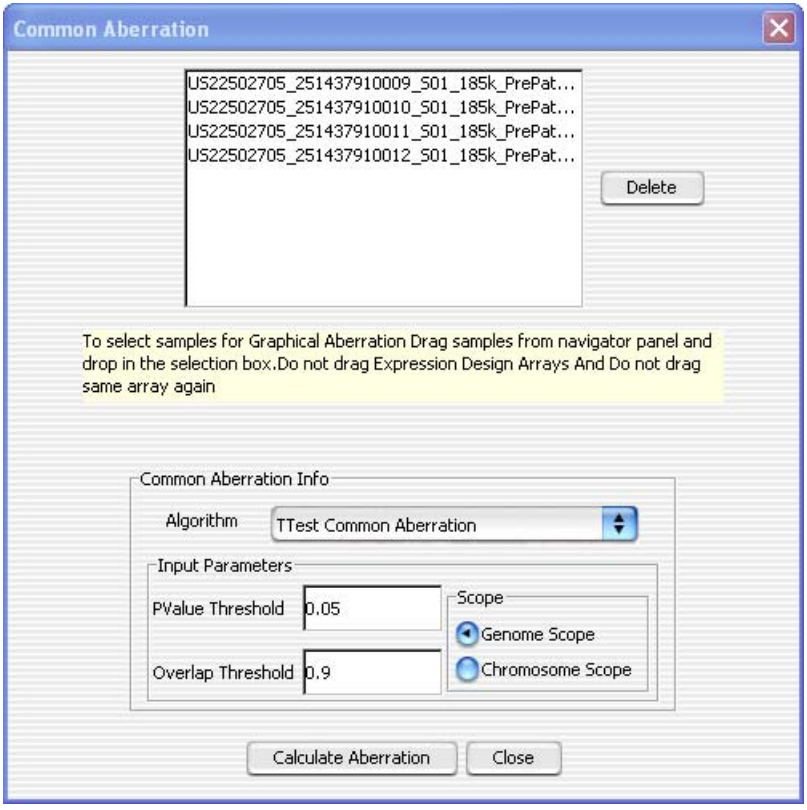


Figure 3-37 Graphical Common Aberration pane

This allows selection of arrays and parameters for inclusion in a **Graphical Common Aberration Summary**.

Main Window Drag and drop at least two arrays into the main window for inclusion in the **Graphical Aberration Summary**.

Delete Select any array(s) and click **Delete** to remove them from the array list.

Algorithm: Select the algorithm to use for calculating the common aberration. The output from these algorithms differ in sensitivity to large aberrations.

- **Context Corrected Common Aberration** – The recommended algorithm to use.

- **T-Test Common Aberration** – Provides results that may not be sensitive to large aberrations. The T-Test Common Aberration algorithm is a measure of the likelihood of detection of a given aberration. Small aberrations may be better detected using this algorithm due to their decreased likelihood score.

Input Parameters Specify the input parameters for multiple-sample experiments.

- **PValue Threshold** – Sets the pValue limits of the aberrations called.
- **Overlap Threshold** – Sets the limit of overlapping. A threshold of 0.9 covers the entire genome.

Scope Specify the scope of the analysis for Context Corrected Common Aberration algorithm. (By default, the scope of the T-Test Common Aberration algorithm is the entire genome.)

- **Genome Scope** – Results displayed for the entire genome.
- **Chromosomal Scope** – Results displayed on a chromosome-by-chromosome basis.

Calculate Aberration Click **Calculate Aberration** to proceed with the Graphical Common Aberration Summary.

Close Click **Close** to exit the dialog box and return to the Main Window.

Graphical Common Aberration Summary

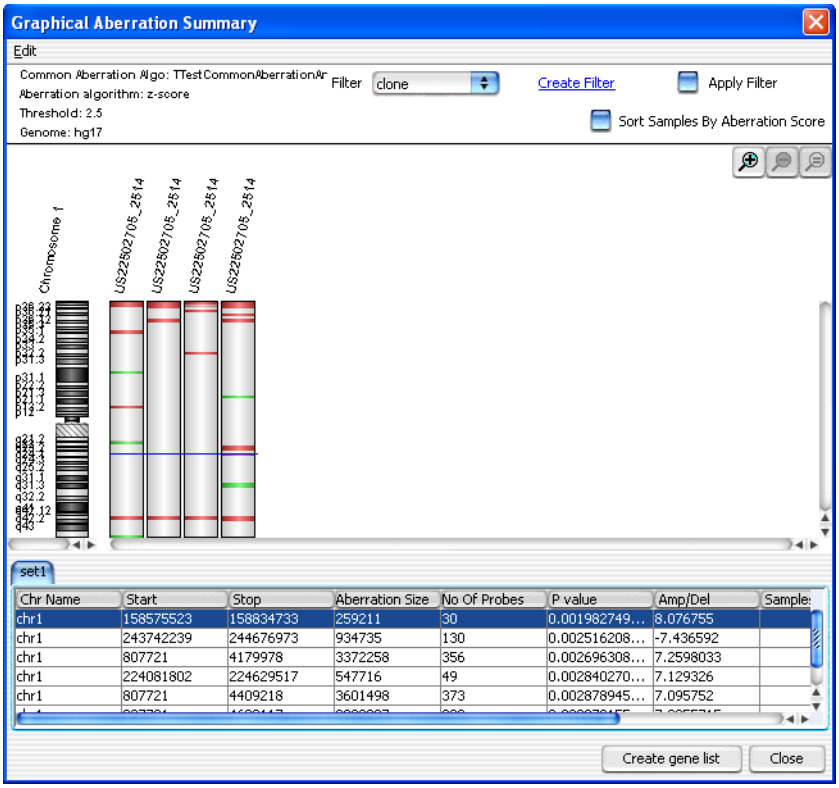


Figure 3-38 Graphical Common Aberration Summary pane

The graphical and tabulated common aberration summary pane. As in the **Graphical Aberration Summary** pane, this summary colors significant aberrations in heat-map fashion (red for putative amplifications and green for putative deletions). The scoring is the same as it is in the main display. In addition, this pane allows the creation and visual application of interval filters, sorting by aberration score, and creation of lists of genes proximal to a common aberration shared across samples.

Filter Selects an existing interval filter to apply to the graphical common aberration summary.

Create Filter Allows creation and specification of new interval filters to the graphical common aberration summary pane. Launches the Interval Filter setup dialog box.

Apply Filter Applies a selected filter to the graphical common aberration summary.

Sort Samples by Aberration Score Sorts the samples in ascending order by aberration score. This function will be unavailable if the T-Test Common Aberration algorithm was used to generate the common aberrations.

Set 1 / Set 2 /... Tab Select the differential subset of samples for tabulated display of sample attributes.

Create Gene List Generates a list of genes within a genomic interval containing a common aberration shared across sample sets.

Close Exit the dialog box without saving any parameters.

Graphical Common Aberration Summary Setup

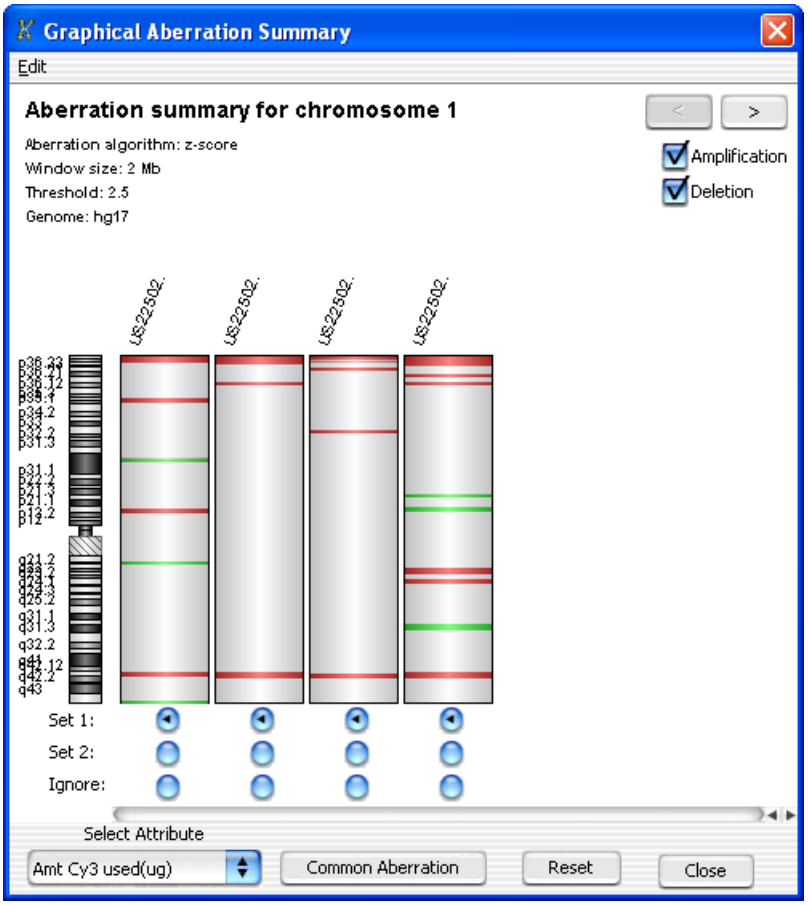


Figure 3-39 Graphical Common Aberration Summary Setup pane

This is analogous to the **Graphical Aberration Summary** pane, but allows for common aberration visualization with specified sample exclusion. It also allows for increment or decrement views by chromosome. As in the **Graphical Aberration Summary** pane, this summary colors significant common aberrations in heat-map fashion (red for putative amplifications and green for putative deletions).

Left and Right Arrow	Decrements or increments, respectively, the chromosome displayed.
Amplification	Select the check box to display amplified chromosomal regions in the samples analyzed in the selected arrays.
Deletion	Select the check box to display deleted chromosome regions in the sample of the selected arrays.
Set 1/Set 2	Select the samples for inclusion in up to two differential common aberration analysis sets.
Ignore	Select the samples for exclusion from any common aberration analysis set.
Select Attribute	Select the groupings of samples by attribute.
Common Aberration	Generates a graphical aberration output summary by set assignment using the chosen parameters.
Reset	Returns all parameters to the default settings.
Close	Exit the dialog box without saving any parameters.

Graphical Penetrance Summary

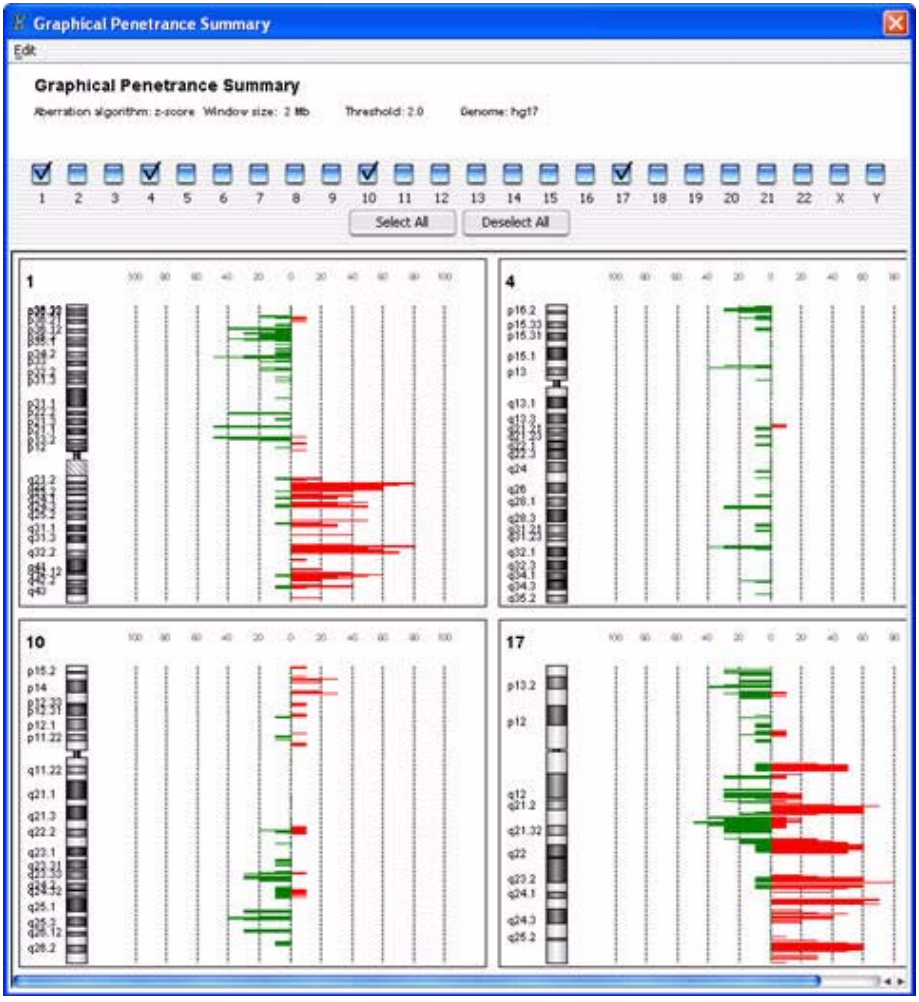


Figure 3-40 Graphical Penetrance Summary pane

Penetrance plots display the percentage of selected arrays that have an aberration at each probe position on the array for the selected chromosomes. Arrays with a deletion aberration are counted and displayed in green on the left side of the X axis. Arrays with amplification aberrations are displayed in

red on the right side of the X axis. Penetrance plots display penetrance for aberrations as they are currently displayed in the main windows with any filters applied.

Select Chromosomes The row of check boxes across the top of the window represent the chromosomes. This allows you to create plots for multiple chromosomes (in this example, four chromosomes: 1, 4, 10, and 17.

Select All Select all chromosomes.

Deselect All Clear the selection of any selected chromosomes.

Interval Filter Setup

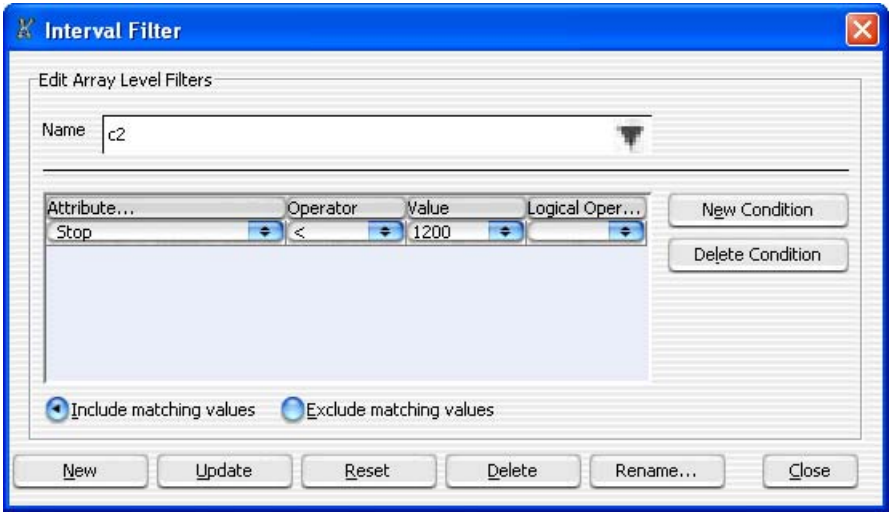


Figure 3-41 Interval Filter dialog box

Graphical common aberration summaries can be filtered by genomic interval to focus on or exclude chromosomal regions. This dialog box allows specification of such filters.

Name Displays a list of existing interval filters by name. Choosing an existing filter populates the main window of the dialog box with the existing filter parameters.

Filter Description Window This is the main window of the dialog box. Complete the Interval Filter dialog box by assigning attributes, operators, values, and logical expressions to match intervals. Each line comprises one set of criteria for the filter.

New Condition Allows the addition of an additional set of criteria to the filter.

Delete Condition Removes the selected set of criteria from the filter.

Include / Exclude Matching Values Application of an interval filter reflects intervals which pass a boolean conditional across all criteria. Select to include or exclude such intervals.

New Set the name and criteria for a new filter. Prompts for a name for the new filter, and then returns to the Interval Filter dialog box.

Update Applies changes in the set(s) of criteria to the filter.

Reset Returns all parameters and criteria to a default, unspecified state.

Delete Removes the filter from the list of available interval filters.

Rename Identifies the filter using a different name in the list of available interval filters.

Close Exit the dialog box without saving any parameters.

Matched Sample

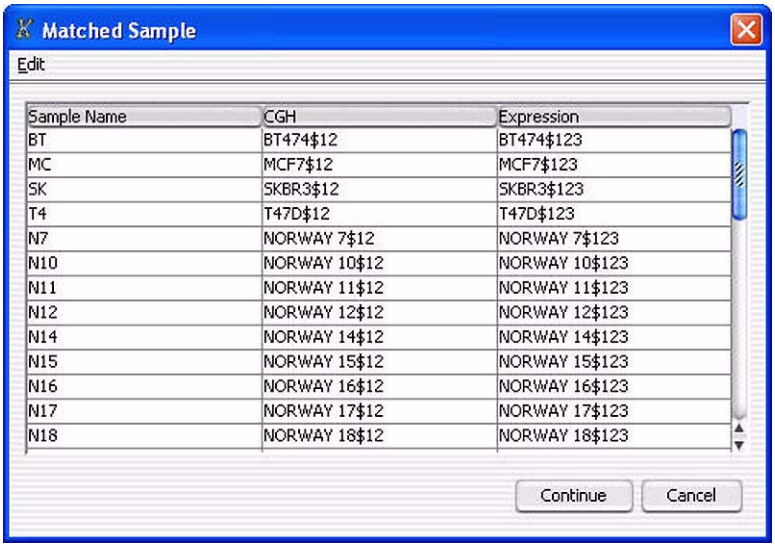


Figure 3-42 Matched sample dialog box

Edit Save the Matched Sample table data.

List Box

- **Sample Name** – Lists the sample names that had matching CGH and Expression microarrays. This column name changes depending on the attribute that is selected to match the microarray data.
- **CGH** – Identifies the CGH microarray in the sample that matched.
- **Expression** – Identifies the EXP microarray in the sample that matched.

Continue Opens the Correlation Results dialog box. See Figure 3-25 on page 119.

Cancel Close the dialog box.

Model System Attributes

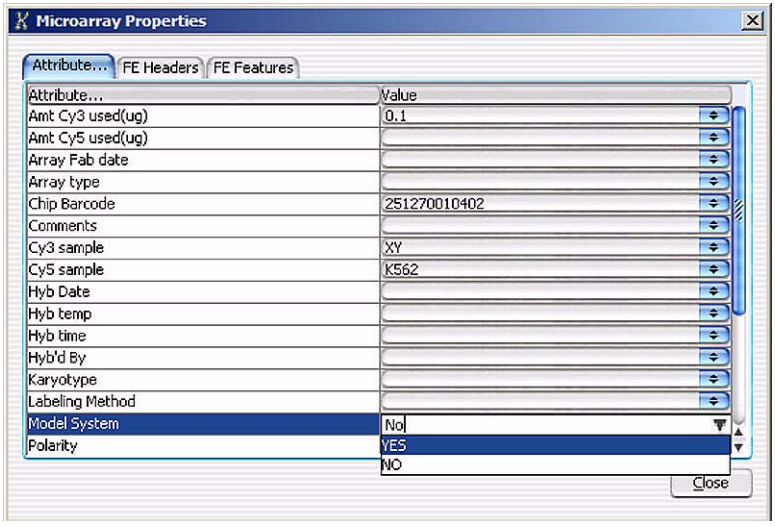


Figure 3-43 Microarray Properties dialog box

Displays the attributes in a sample by name and value. In this example, the Model System attribute was changed from NO to YES.

MovAvg Example Parameters

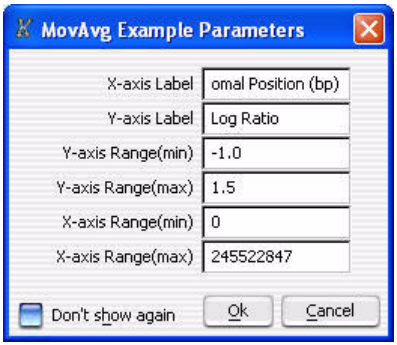


Figure 3-44 MovAvg Example Parameters.

This dialog box shows an example of the parameters used to conduct a Moving Average data-smoothing operation. For more information, see "MovAvg Example.pl" on page 194.

QC Frequency Distribution

Plots the frequency of a metric as a bar graph by selecting the check box in the QC Metrics Table dialog box (see Figure 3-49 on page 150), then clicking **Show Frequency Distribution**. To display the frequency of all metrics in one graph, click **Select All** before clicking **Show Frequency Distribution**. Figure 3-45 displays the frequencies of all metrics for the selected microarrays.

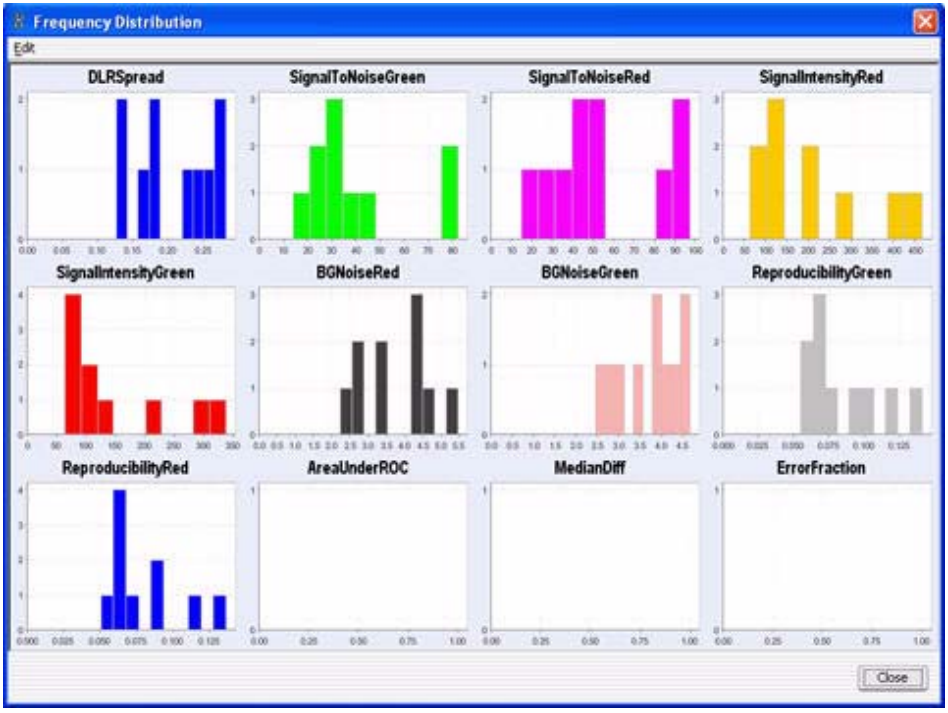


Figure 3-45 QC Frequency Distribution displaying multiple metrics in one graph

NOTE

The last three metrics are only applicable for use in a male versus female (XY versus XX) model systems. To calculate the AreaUnderROC, MedianDiff, and ErrorFraction, go to the Experiment in the Navigator area. Right-click the array, then **Show Properties > Attribute**. Set the **Model System** attribute to **Yes**, if appropriate, and click **Close.** See Figure 3-43 on page 144.

QC Metrics Graph

Displays a QC Metrics Graph showing the distribution of all of the metrics. You can display the distribution as a bar graph, line graph, or both. See Figure 3-46. The example displays all metrics in a combined bar graph-line graph format. Click an individual bar to display a pop-up window where you can change an array's QC Status.

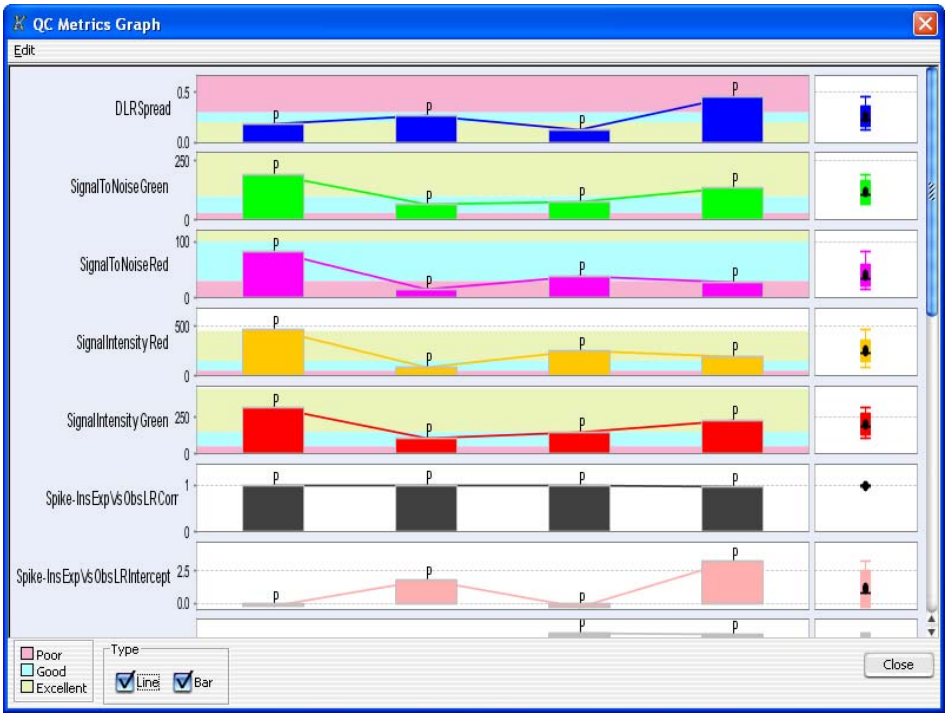


Figure 3-46 QC Metrics Graph showing distribution curves as combined bar graph and line graph

You can change the QCMetricStatus attribute (**Pass/Fail/Marginal/NA**) of each array from this QC Metrics Graph by right-clicking at the top of that particular bar. See Figure 3-47.

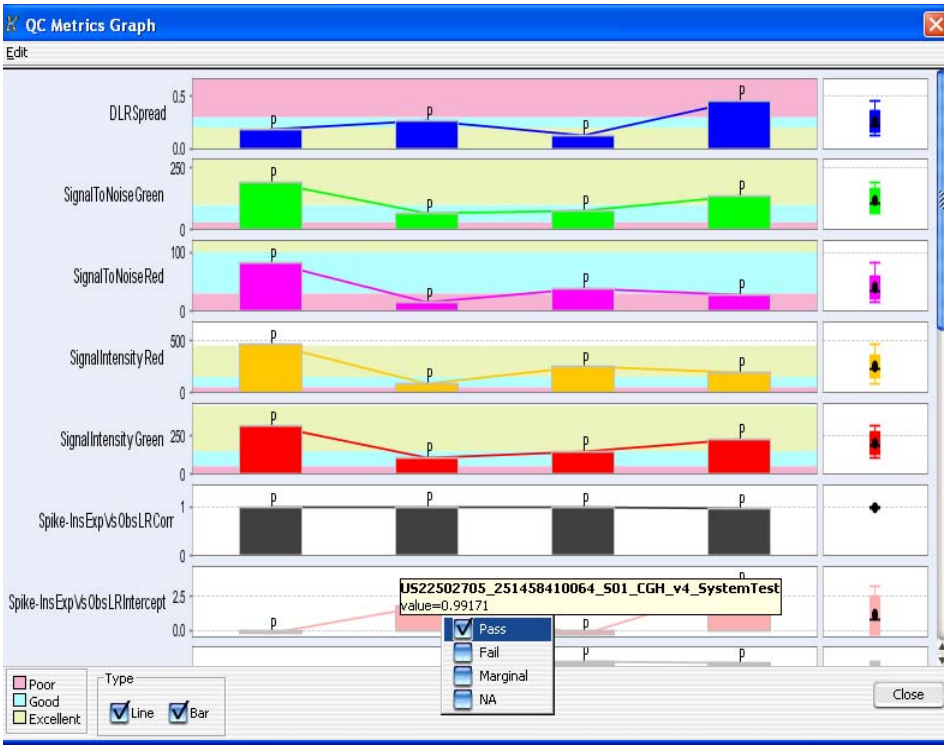


Figure 3-47 QC Metrics Graph enabled for changing QCMetricStatus attribute

The attributes that you see in the far right column of the graph for each metric are commonly known as “Box and Whisker plots” (see Figure 3-48) and carry the following meanings:

- The lower and upper edge of the box represents the 25th and 75th percentiles.
- The black horizontal line in the box is the median.
- The black dot is the mean.
- The "whiskers" indicate the farthest points that are not outliers (points that are within 3/2 times the 25th and 75th percentile range.
- If a marker arrow is present, it indicates that there are far-out values that are more than 3/2 times the interquartile range from the end of a box.

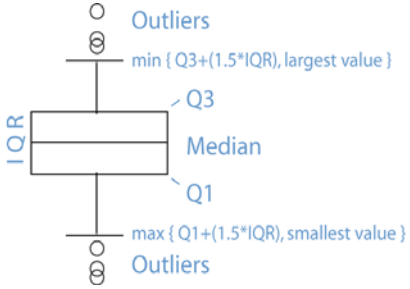


Figure 3-48 Box and Whisker plot (Tukey, 1977).

QC Metrics Table

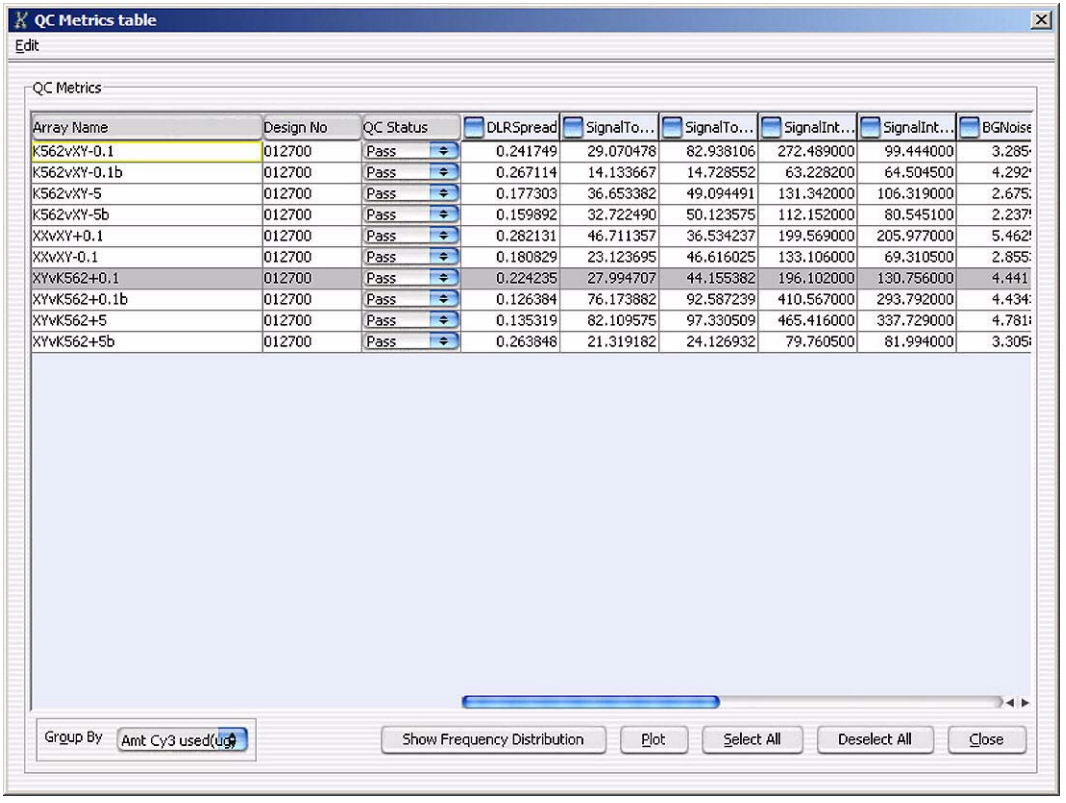


Figure 3-49 QC Metrics Table dialog box

In this table and the associated plot, you can evaluate the quality of your aCGH microarray results and assign a QC status to each microarray based on this evaluation.

Table

- **Array Name** – Lists the microarrays in the experiment.
- **Design No** – Identifies the design containing the microarray.
- **QC Status** – Status can be Pass, Fail, Marginal, or NA.
- **Metrics** – The metrics are named in the headers of columns containing the data. The columns can be repositioned by dragging-and-dropping the headers to help in comparing the results for different microarrays.

- **DLRSpread**
- **SignalToNoiseGreen/Red**
- **SignalIntensityRed/Green**
- **Expected vs. Observed Spike-in Plot Correlation**
- **Expected vs. Observed Spike-in Plot Intercept**
- **Expected vs. Observed Spike-in Plot Slope**
- **BGNoiseRed/Green**
- **ReproducibilityGreen/Red**
- **AreaUnderROC**
- **MediaDiff**
- **Error Fraction**

Group By Select the Attribute to use in grouping microarrays for analysis.

Show Frequency Distribution Graphically displays the distribution of the metrics in bar graph format.

Plot Graphically displays the distribution of the metrics as a line plot.

Select All Select all metrics' check boxes.

Deselect All Clear all selected metrics' check boxes.

Close Close the QC Metrics Table dialog box.

For more information on metrics, see "QC Metric" on page 71 under the Tools menu.

QC Metrics Thresholds - Recommendations (Plot and Table)

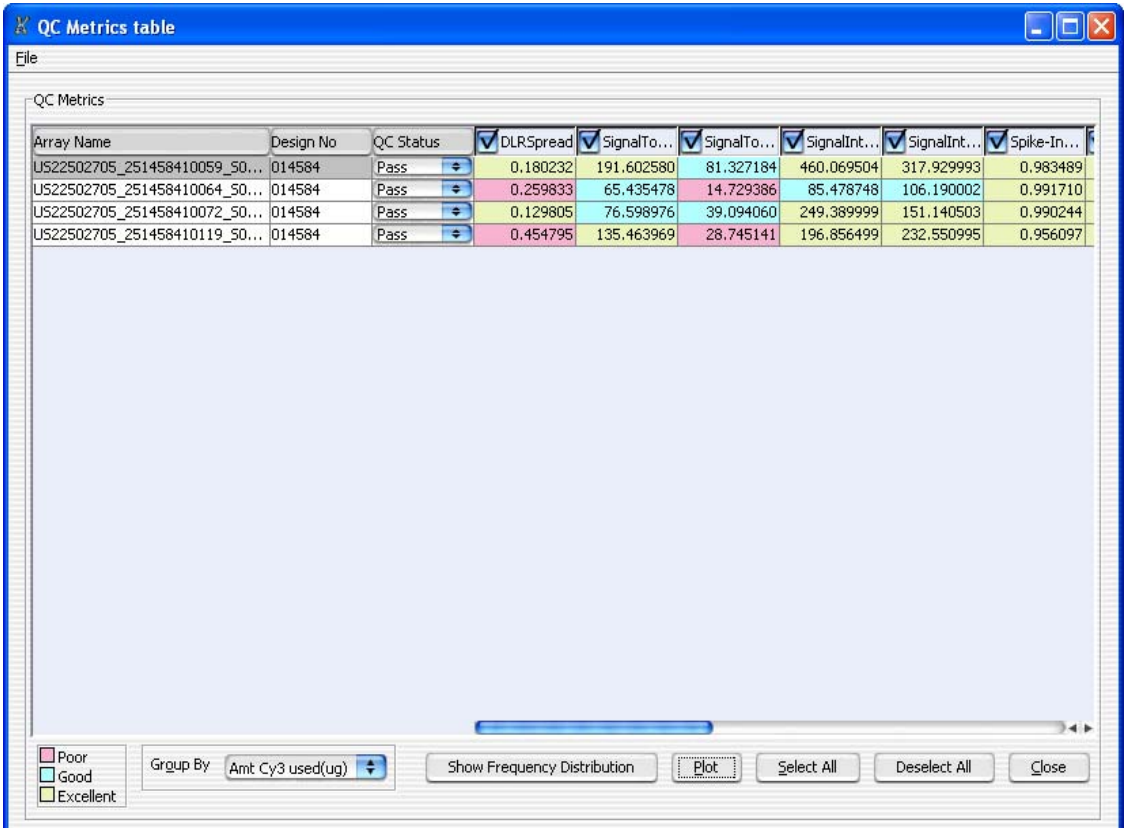


Figure 3-50 QC Metrics Table dialog box represents three color-coded threshold regions (Excellent, Good, Poor)

The QC Metrics thresholds that appear in the table in Appendix B of this User's Manual now have been grouped by color and are rated as Yellow=Excellent, Turquoise=Good, and Pink=Poor. These three colors highlight the appropriate data in the QC Metrics dialog box and provide a quick assessment of the quality of your results.

Color Legend The Color Legend is located in the lower left corner of the QC Metrics table and graph dialog boxes. The colors (yellow, turquoise, and pink) represent threshold ratings of Excellent, Good, and Poor, respectively.

- In the QC Metrics table dialog box, select a metric to plot (**DLR Spread**, **Signal to Noise Green,** etc.**)**

Show Frequency Distribution Button

- **Click on the Show Frequency Distribution** button on the bottom of the dialog box. The QC Metrics Graph dialog box appears with then selected frequency distribution(s) plotted on the graph.

Close Close the QC Metrics Table dialog box.

For more information on metrics, see "QC Metric" on page 71 under the Tools menu.

Scroll to Column

Figure 3-51 Scroll to Column dialog box

Select column Select the column that you want to move to.

OK Move to that column.

Cancel Close the dialog box.

Select Chromosome Intervals

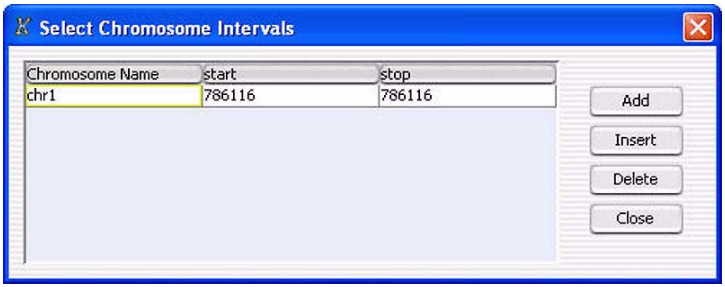


Figure 3-52 Select Chromosome Intervals dialog box

List box Chromosome and their intervals are listed based on

- Chromosome Name – Lists all chromosomes under study and each interval, if previously identified.
- Starting point – The location on the chromosome where the interval begins.
- Stopping point – The location on the chromosome where the default interval ends.

Add Click to add a blank entry to the list.

a Click the blank entry in the **Chromosome Name** column and select the chromosome from the drop-down list. The interval displayed covers the entire chromosome.

b Click the blank entry in the **Start** column and type in the interval's starting point.

c Click the existing entry in the **Stop** column to select it, then change the entry to show the interval's stopping point.

Insert Select any column and click Insert to add a duplicate row of that table to the list.

Delete To delete a row from the list.

Close To exit the dialog box.

Select Color

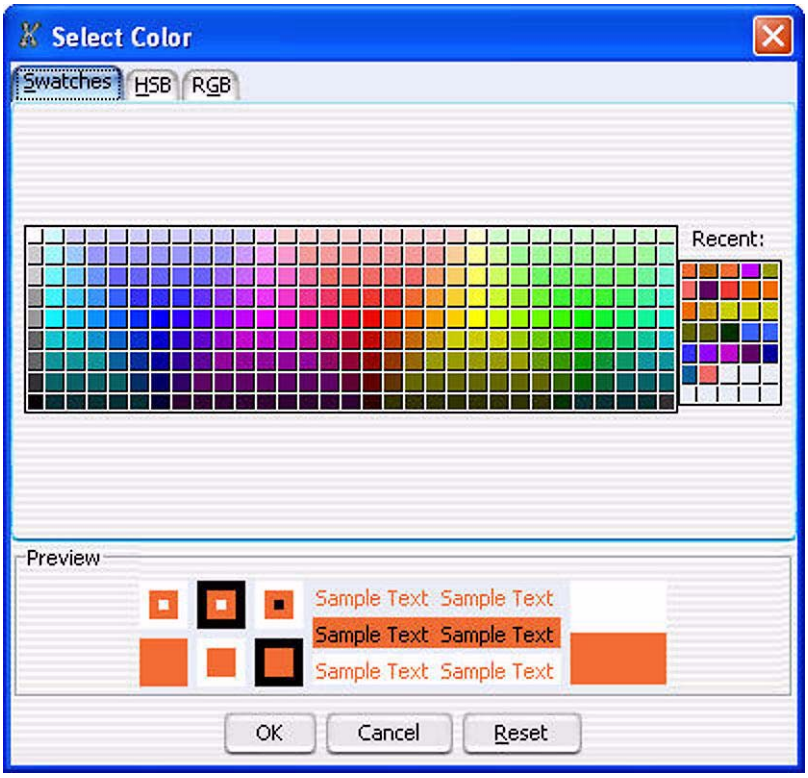


Figure 3-53 Select Color dialog box displaying swatches

Swatches Tab Display colors based on color samples (swatches).

HSB Tab Display colors based on an HSB schema (Hue, Saturation, and Brightness or Value).

RGB Tab Display colors based on an RGB schema (Red-Green_Blue).

Recent: Display recent color selections (in Swatches view only).

Preview Panel Display what the results of the current selections would be.

OK/Cancel Click **OK** to accept the new color selections or **Cancel** to return to existing parameters.

Reset Click **Reset** to return colors to the default color selection.

See Figure 3-54 and Figure 3-55 for the HSB and RGB tab views.

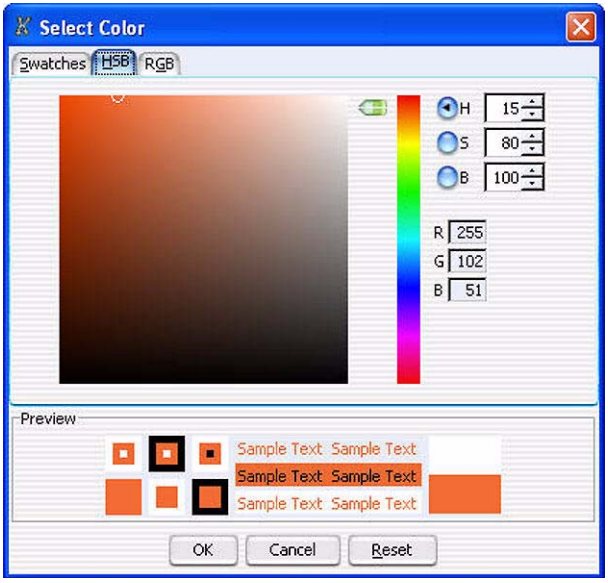


Figure 3-54 Select Color dialog box displaying HSB schema

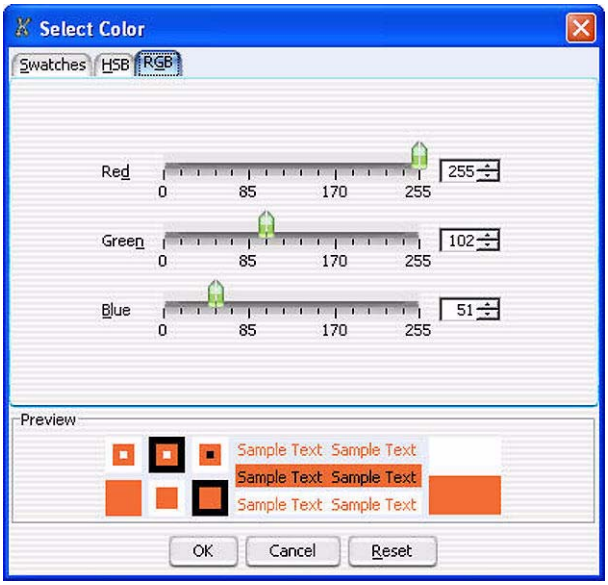


Figure 3-55 Select Color dialog box displaying RGB schema

Spike In Plots

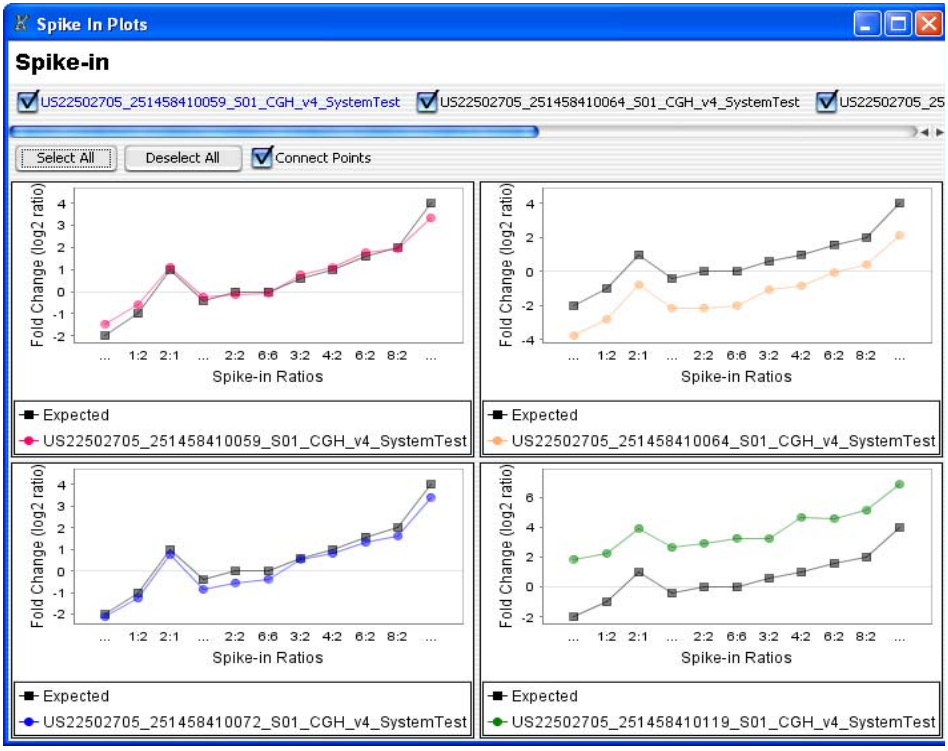


Figure 3-56 Spike In Plots dialog box

Select All Clicking on the **Select All** button produces a graph for each spike-in ratio calculated. The x-axis represents the spike-in ratios and the y-axis represents the median $\log_2$ value of the sample in the experiment.

Deselect All Removes all graphs from the screen.

Connect Points Produces a line connecting all points.

Spike In Panel Plots

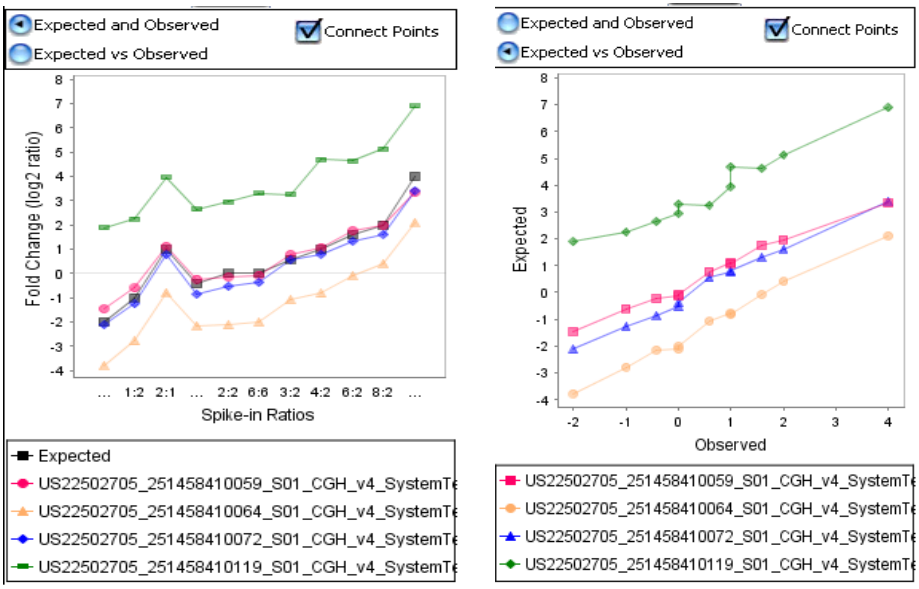


Figure 3-57 Spike In Panel Plots dialogs

Expected and Observed Produces a graph for each spike-in ratio calculated. The x-axis represents the spike-in ratios and the y-axis represents the median $\log_2$ value of the sample in the experiment.

Expected vs. Observed Produces a graph for each spike-in ratio calculated. The x-axis represents the observed ratios and the y-axis represents the expected ratios of the sample in the experiment.

Connect Points Clicking on the box next to **Connect Points** produces a line connecting all points.

Spike In Ratios Plots

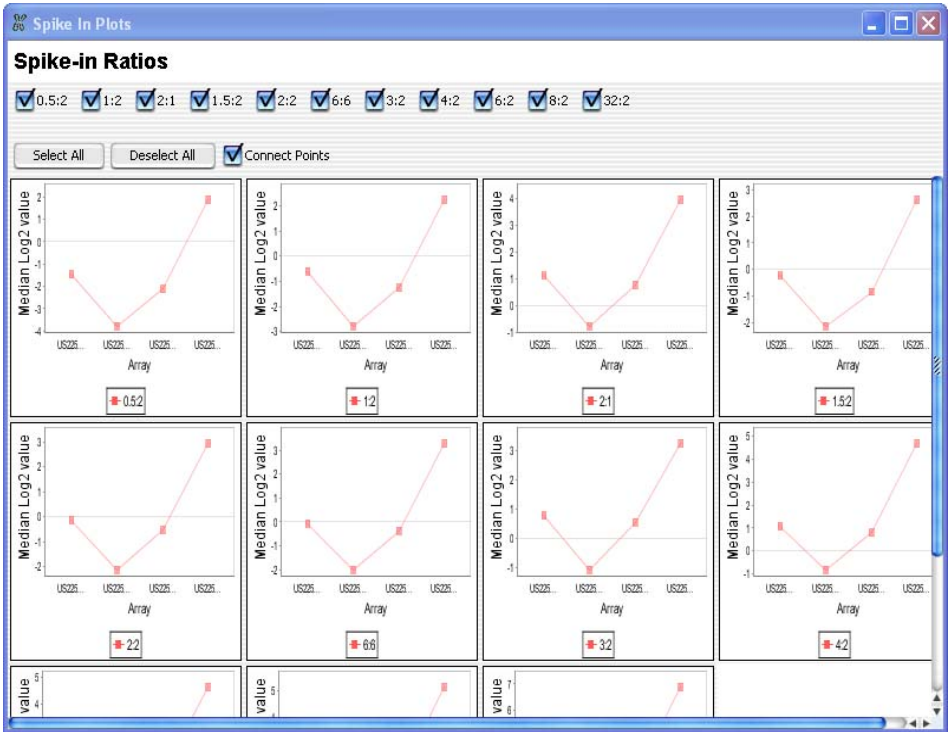


Figure 3-58 Spike In Ratios dialog box

Select All Clicking on the **Select All** button produces a graph for each spike-in ratio calculated. The x-axis represents each array in the specific experiment and the y-axis represents the median $\log_2$ value of the spike-in ratio tested in the experiment.

Deselect All Clicking on the **Deselect All** button deletes all graphs on the screen.

Connect Points Clicking on the box next to **Connect Points** produces a line connecting all points.

Report Creation and Export

This section provides illustrations of the major dialog boxes that you will encounter when analyzing and visualizing your CGH data. The dialog boxes are arranged alphabetically.

Export

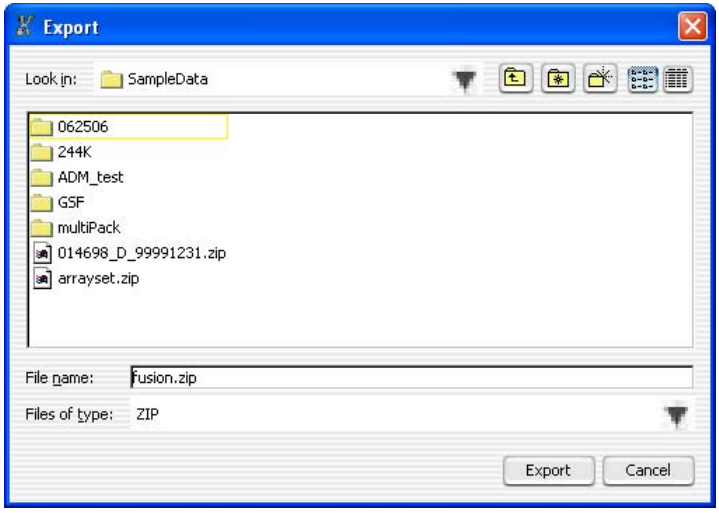


Figure 3-59 Export dialog box

Look in Displays the name of a selected folder to export files. Click the down arrow to select a different folder from your system's directory.

File Navigation icons Click to move to a higher level folder, return to the desktop, create a new folder, or change the way folder and file names are displayed in the in the list box.

List Box Displays the names of folders and files of the indicated type contained in the selected folder. Click to select a folder or file.

File name Displays the name of the selected file.

Files of type Displays the type of files displayed in the list box.

Export Saves the selected file.

Cancel Cancels your selections and close the dialog box.

Export Experiments

Figure 3-60 Export Experiments dialog box

Select Experiments to export Displays the available experiments to export.

Select All Select all of the available experiments to export.

Deselect All Cancel the selection of any selected experiments.

OK Accept your selections.

Cancel Close the dialog box.

Export Filters

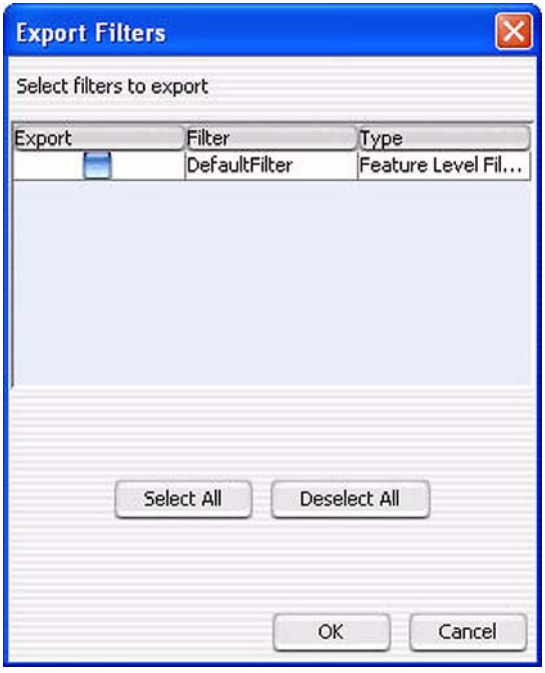


Figure 3-61 Export Filters dialog box

List Box
- Export – Displays list of available filters.
- Filter – Displays name of filter.
- Type – Displays type of filter.

Select All Select all filters in the list.

Deselect All Cancel selection of all selected filters.

OK Accept the selections and import those filters.

Cancel Cancel the selection and close the dialog box.

Gene List

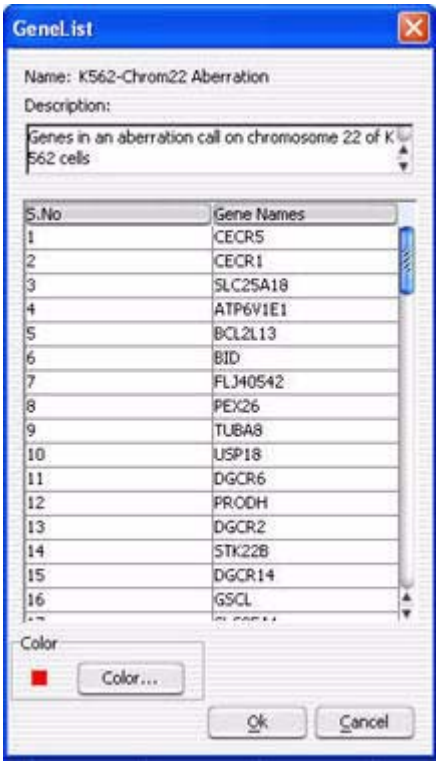


Figure 3-62 GeneList dialog box

Name Displays the experiment's name, automatically.

Description Displays the description entered when the experiment was created.

SNo Displays the sort order number.

Gene Name Displays a list of the gene's names.

Color Click the Color button to access the Select Color dialog box where you can change a gene's color code. See Figure 3-54 on page 156.

OK/Cancel Accept any changes or Cancel to return the list to its previous selections.

Text Aberration Summary Setup

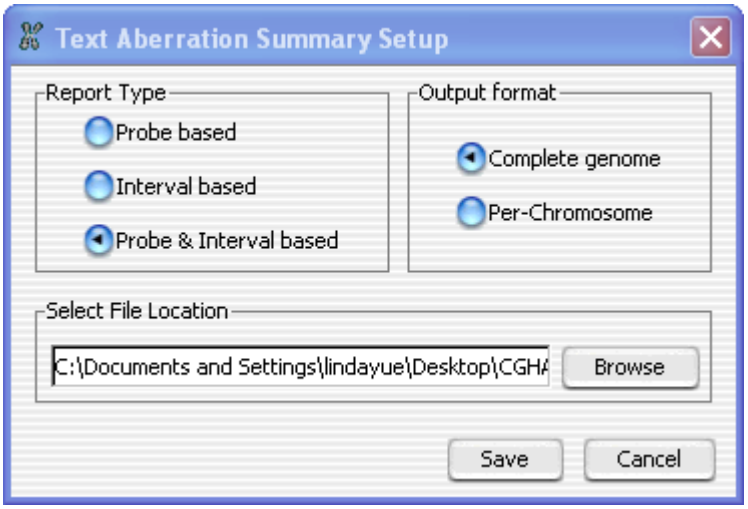


Figure 3-63 Text Aberration Summary Setup dialog box

Tabulated aberration reports can be generated for any number of samples.

Report Type: Select the report type to use for the aberration output file. The Report Types are:

- **Probe Based** – Reports by each probe on specified array(s).
- **Interval Based** – Reports by genomic intervals.
- **Probe & Interval Based** – Reports by probe and interval.

Output Format Specify the output format to use for the aberration output file.

- **Complete Genome** – Create a report for the entire genome.
- **Per-Chromosome** – Create a report on a chromosome-by-chromosome basis.

Select File Location Type in the name of the report under which the results will be saved, then click **Browse** to specify the directory where the file is saved.

Save/Cancel Click **Save** to accept the new parameters and generate a report file or **Cancel** to return to the main window.

Preferences

This section provides illustrations of the major dialog boxes that you will encounter when setting preferences. The dialog boxes are arranged alphabetically.

Preferences - Gene Symbols

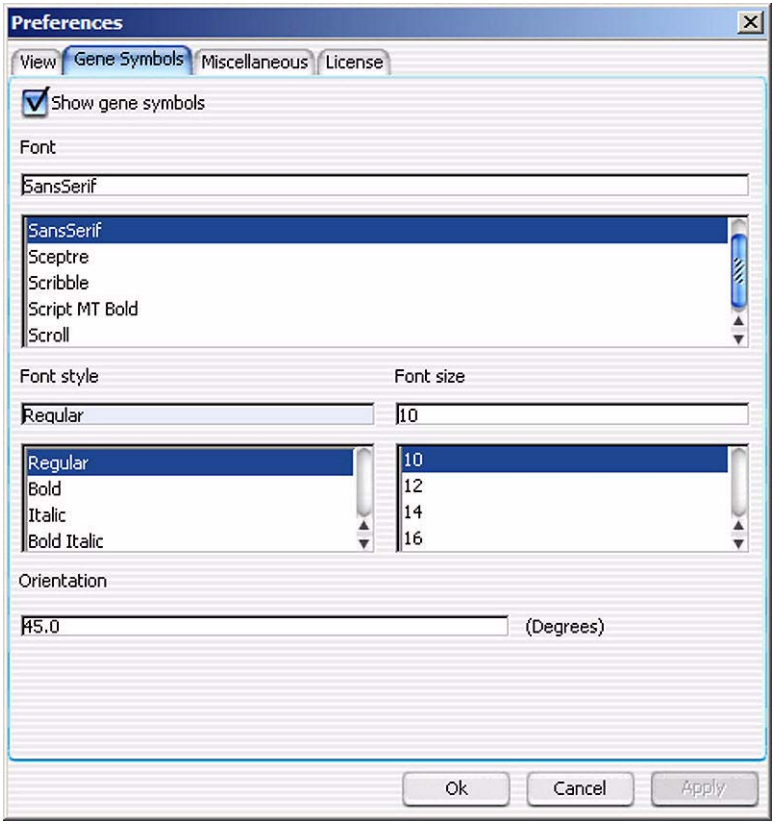


Figure 3-64 Preferences dialog box displaying Gene Symbols tab options

Show gene symbols Select the check box if you want to show gene symbols.

Font Select the font from the list box showing all available fonts. The default is Sans Serif.

Font style Select the fonts style from the list box. Choices are: Regular, Bold, Italic, and Bold Italic. The default is Regular.

Preferences - License

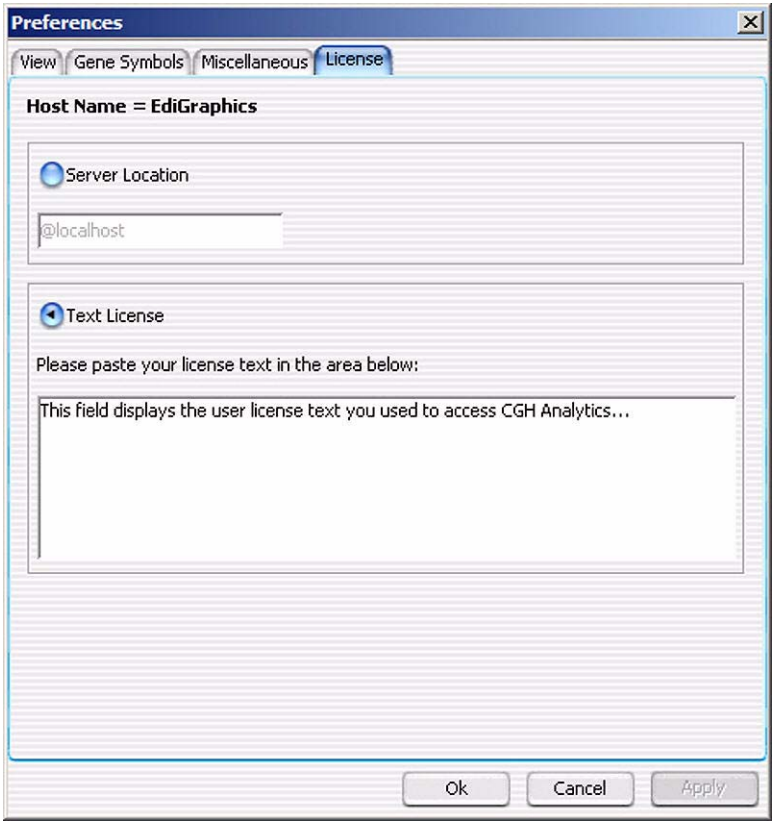


Figure 3-65 Preferences dialog box displaying License tab options

Host Name Displays the host name, automatically.

Server Location Server location should be selected if you have a concurrent user license. If appropriate, click to enable and type in the name of your license server. The default is @localhost. Replace localhost with the name of the computer used as the license server. **Text License** is disabled (grayed) if **Server Location** is enabled.

Text License Text licenses are used if you have a workstation license. If you do have a workstation license, paste your license in the text box. If you have entered a license previously, it is displayed in the text box. **Server Location** is disabled (grayed) when **Text License** is enabled.

Apply Apply your changes to the parameters.

OK/Cancel Accept your changes and exit, or cancel all changes and return to the previously selected parameters.

Preferences - Miscellaneous

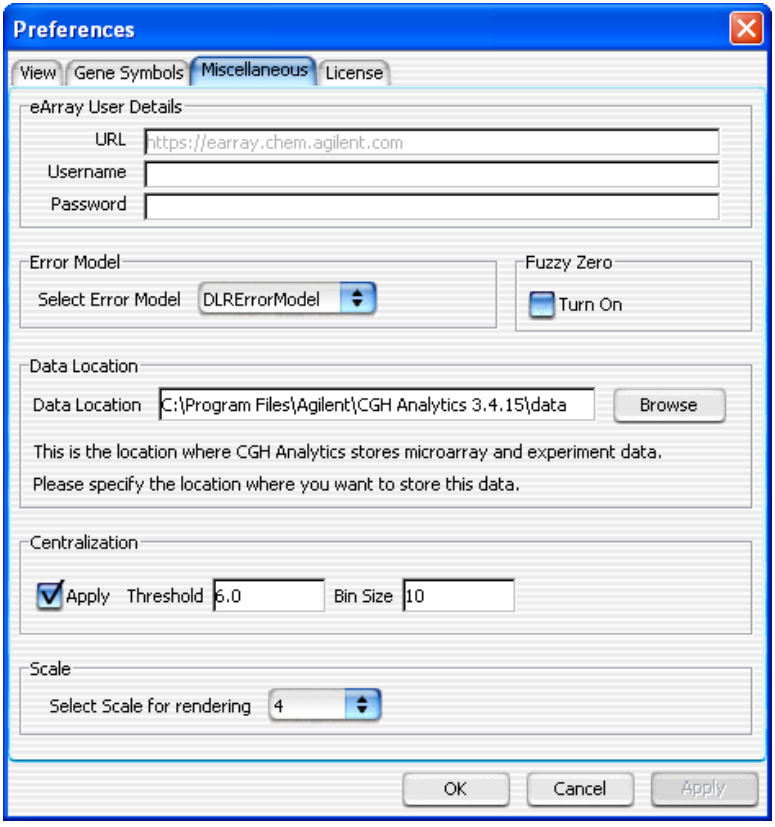


Figure 3-66 Preferences dialog box displaying Miscellaneous tab options

eArray User Details Set user details for logging on to the Agilent eArray site where you can update Agilent microarray annotation information.

- **URL** – Currently, https://earray.chem.agilent.com.
- **Username** – The name registered on the eArray site.
- **Password** – The password registered on the eArray site.

Error Model Set the error model used by CGH Analytics. In version 3.3 there is only one choice, **DLRErrorModel**.

Fuzzy Zero Select **Turn On** to apply Fuzzy Zero correction. Fuzzy Zero correction is turned off by default.

Data Location Type in the path to the folder where you want CGH Analytics to store the microarray and experiment data.

Centralization Select to toggle application of Centralization, an algorithmic method for redefining the log ratio view by shifting the zero point. To enable, select the check box and type in the desired threshold value. For more information, see the "Centralization Algorithm" on page 171. Centralization is applied by default.

NOTE Depending on the density of the array, centralization processing may take up to 1 minute per array. This can lead to extensive processing time and should be planned for or controlled accordingly. For example, ff you have an experiment which has 50 arrays with centralization turned on, the first time you run the application it may take 50 minute.

Scale Change the scale for rendering your data. This toggles the maximum $\log_2$ ratio displayed between 4 and 5 (16- and 32-fold).

Apply Apply your changes to the parameters.

OK Accept your changes and close the dialog box.

Cancel Cancel all changes and close the dialog box.

Preferences - View

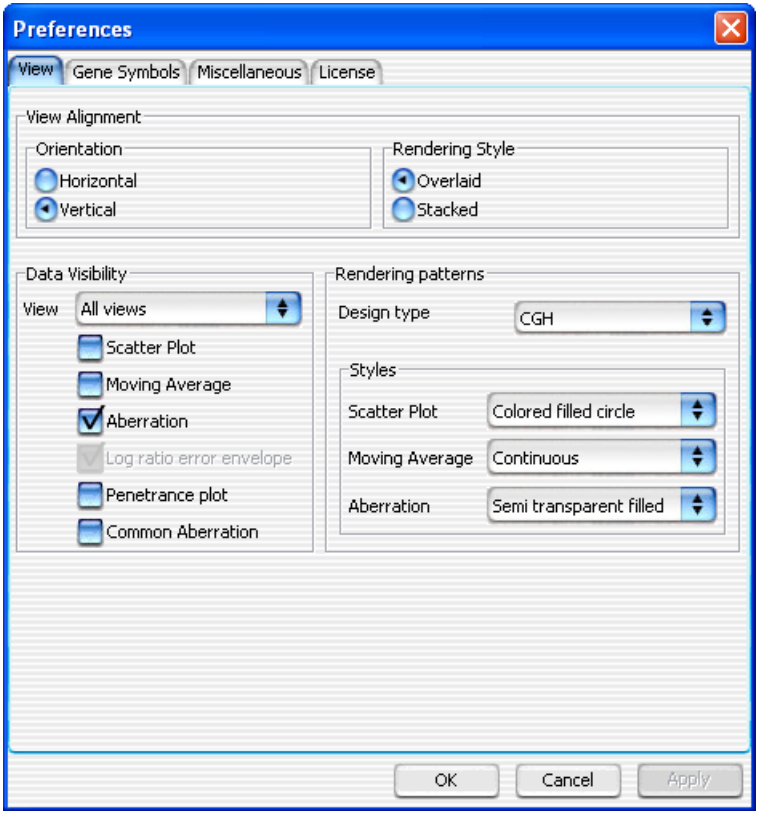


Figure 3-67 Preferences dialog box displaying View tab options

The View tab options control the general characteristics of how data are displayed on your monitor.

View Alignment

- Orientation – Selects the orientation of three views on the Main view screen:
- Horizontal – Re-orient three views to a horizontal aspect in the order of Gene, Chromosome, and Genome views, top to bottom. The Navigator and Tab view orientation remains unchanged. See Figure 2-8 on page 65.

- Vertical – Display all views in a vertical aspect, left to right: Navigator, Genome, Chromosome, and Gene views. This is the default display. See Figure 2-1 on page 46.
- Rendering Style – Select the way Consomme and Gene data are rendered on your screen.
  - Overlaid – Display data from multiple arrays superimposed one on top of another.
  - Stacked – Display data from each array in a separate plot.

Data Visibility

- **View** -- Choose what features you want to display for the Genome, Chromosome, and Gene views, either individually or together. Select one or more check boxes:
  - **Scatter Plot**
  - **Moving Average**
  - **Aberration**
  - **Log ration error envelope**
  - **Penetrance plot**
  - **Common Aberration**

Rendering patterns

- **Design type** -- Specify the type of design to which you are applying these patterns: **CGH, Expression, or Other.**
- **Styles** -- Set up the parameters for displaying your data.
  - **Scatter Plot** -- Specify how to display individual data points as: **Color filled circles (ellipses), + signs,** or **x signs.**

NOTE

Rendering scatter plots for more then 10 high density arrays in the chromosome view may take significant time. It is advisable to not select more than 10 arrays if the scatter plot is turned on for chromosome view. Selecting ellipses as the rendering style for CGH scatter plots can also decrease performance. Please change the rendering style for CGH data from ellipse to the plus (+) or cross hair sign.

- **Moving Average** -- Specify how to display moving averages: as **Continuous, Dashed, or Dotted** lines, or you can choose **Do not show area.**
- **Aberration** -- Specify how to display aberrations: as **Semi transparent** or **Hatched** areas, or you can choose **Do not show area**.

Apply Apply your changes to the parameters.

OK/Cancel Accept your changes and exit, or cancel all changes and return to the previously selected parameters.

4 Statistical Algorithms

This section provides implementation details for several algorithms to facilitate the statistical analysis and calibration of aberrant regions. Algorithms are listed in alphabetic order. If you are planning a CGH analysis, the next page provides an overview of the aberration algorithms.

Aberration Algorithms Overview

CGH Analytics provides algorithms for aberration analysis and visualization. The aberration algorithms include procedures for estimating noise in the data and representing aberration boundaries. Figure 4-1 provides a brief overview of the focus, options, and dependencies for the aberration algorithms.

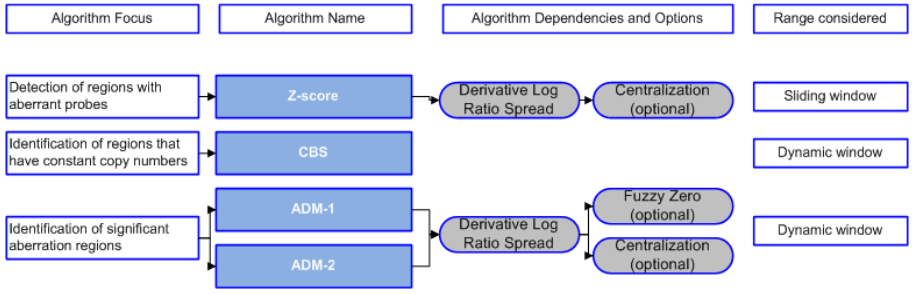


Figure 4-1 Overview of CGH Analytics Aberration algorithms

Derivative Log Ratio Spread (*dLRsd*)

To make aberration calls, CGH Analytics software needs a measure of log ratio noise. Therefore, a measure of the minimum log ratio difference is necessary to make reliable amplification or deletion calls. The *dLRsd* algorithm is a robust method of estimating noise from the sample array alone by calculating the spread of the log ratio differences between consecutive probes along all chromosomes. See 'Derivative Log Ratio Spread' on page 204 for more information.

Z-score

The Z-score algorithm is a quick method of detecting aberrant regions. It assesses genomic intervals with an over or under abundance of probes with log ratio scores that deviate significantly from baseline. The Z-score scores intervals using sliding window of fixed size, specified by the user. Results from the Z-score can suggest aberrant intervals by identifying regions of high probe log ratio change. See 'Z-Scoring for Aberrant Regions' on page 215 for more information.

CBS

Circular Binary Segmentation (CBS) is a method for identifying all genomic change points where the mean log ratio score changes between intervals. CBS therefore provides a reliable report on all intervals that themselves contain constant copy numbers. CBS automatically determines interval sizes which

contain constant copy numbers. Since the output is comprehensive and not ranked, ADM-1 and ADM-2 are preferred methods for aberration classification. See 'CBS' on page 184 for more information.

ADM-1 The Aberration Detection Method 1 (ADM-1) algorithm identifies all aberrant intervals in a given sample with consistently high or low log ratios based on the statistical score. The ADM algorithms automatically determine the optimal size of a statistically significant aberration. If you are trying to find aberrant genomic intervals, the ADM algorithms are the preferred method. See 'ADM-1' on page 178 for more information.

ADM-2 The Aberration Detection Method 2 (ADM-2) algorithm generates a similar statistical score to that produced by ADM-1 analysis, but ADM-2 incorporates quality information about each probe measurement. Use of the probe log ratio error in addition to the log ratio values makes ADM-2 more robust than ADM-1 when data is noisy or there are many aberrations. See 'ADM-2' on page 182 for more information.

Centralization Many statistical algorithms for aberration detection assume that the log ratio values are centered around zero if no aberration occurs, reflecting no change between the reference and sample channels. In samples with a high aberration percentage, this assumption may lead to erroneous results as the measured center of the data may deviate from a log value of zero. The centralization algorithm re-centers the data by finding a constant value to add to or subtract from all log ratio measurements, ensuring that the zero-point reflects the most-common-ploidy state. See 'Centralization Algorithm' on page 193 for more information.

Fuzzy Zero ADM-1 and ADM-2 scores may identify extended aberrant segments with low absolute mean ratios. Often such aberrations represent noise, and are detected because of high number of probes in the region. If long, low aberrations are detected in your analysis, you can apply the fuzzy zero algorithm to correct for the reliance on segment probe number. See 'Fuzzy Zero' on page 209 for more information.

ADM-1

Aberration Detection Method 1 (ADM-1 or “adam-one”) is an aberration algorithm that identifies all aberrant intervals in a given sample with consistently high or low log ratios based on the statistical score. The ADM algorithms search for intervals in which the average log ratio of the sample and reference channels exceed a user specified threshold. In contrast to the Z-score algorithm, the ADM algorithms do not rely upon a set window size, instead sampling adjacent probes to arrive at a robust estimation of the true range of the aberrant segment. The output differs from that of the CBS algorithm by reporting statistically significant aberrant regions, allowing rapid genomic assessment.

The ADM-1 statistical score is computed as the average normalized log ratios of all probes in the genomic interval multiplied by the square root of the number of these probes. It represents the deviation of the average of the normalized log ratios from its expected value of zero.

The ADM-1 score is proportional to the height h (absolute average log ratio) of the genomic interval, and to the square root of the number of probes in the interval. Roughly, for an interval to have a high ADM-1 score, it should have high height or/and include large number of probes.

Before calling any aberration detection routine, the log ratios are normalized in the following way:

Normalization

In the normalization step, the expected average μ is subtracted from all log ratios $v_{i,}$ and then these modified log ratios are divided by the estimated variance σ. This transforms the log ratio scores into a normal distribution with a mean of 0 under the null model assumption. You can calculate μ and σ from one or more microarrays.

$$v_i \rightarrow \frac{v_i - \mu}{\sigma}$$

NOTE

These parameters should be computed for samples that do not contain genetic anomalies, so that μ and σ represent the distribution of a non-diseased sample. Those arrays are usually marked as calibration arrays in an experiment and are typically male/female arrays without genetic anomalies. In some cases, however, self-self arrays are also considered for calibration.

If no array is selected for calibration, then the variance s is computed as derivative log ratio spread of the sample array itself. In that case, the mean μ is considered to be zero.

Once the data are transformed the following score is assigned to each interval *I*:

$$S(I) = \frac{\sum_{i \in I} v_i}{\sqrt{|I|}}$$

S(I) represents the number of standard deviations that the sum of values in *I* deviates from its expected value, under the null model of 0.

Maximization

A call to the ADM-1 function starts a recursive process. The first step is to identify the interval *I* for which *S*(*I*) is maximal and exceeds a predefined threshold parameter, *t*, specified in the user interface. Then the process is called on the interior of this interval, using the interval median as a mean for re-centering the values, as well as on the two intervals, one to the left and the other to the right flanking *I*, towards the two ends of the chromosome.

Algorithm

The overall recursive structure of the algorithm is:

Given a data vector for a single sample, single chromosome, and a statistical threshold value:

1. Find the most significant interval *I* in the chromosome.
2. If *score(I)* >= *t*, you found a significant interval.
3. Add *I* to the list of intervals.

Search recursively for more intervals (a) to the left of *I*, (b) to the right of *I*, and after normalizing, (c) inside *I*.

NOTE A text report can also be generated which reports the p-value corresponding to each interval. The p-value is calculated using the normal probability distribution function and the score of that interval.

Interpreting the results

The recursion stops when no interval with $S(I)$ exceeding t is found. All intervals found in this process are reported, and a plot is generated as output. The intervals are rendered as steps in the visualization panel. The height of each step is equal to the average log ratio of that interval. Steps are also extended halfway to the neighboring probes on each side of an interval (other than at the end of the chromosome and centromere). See Figure 4-2.

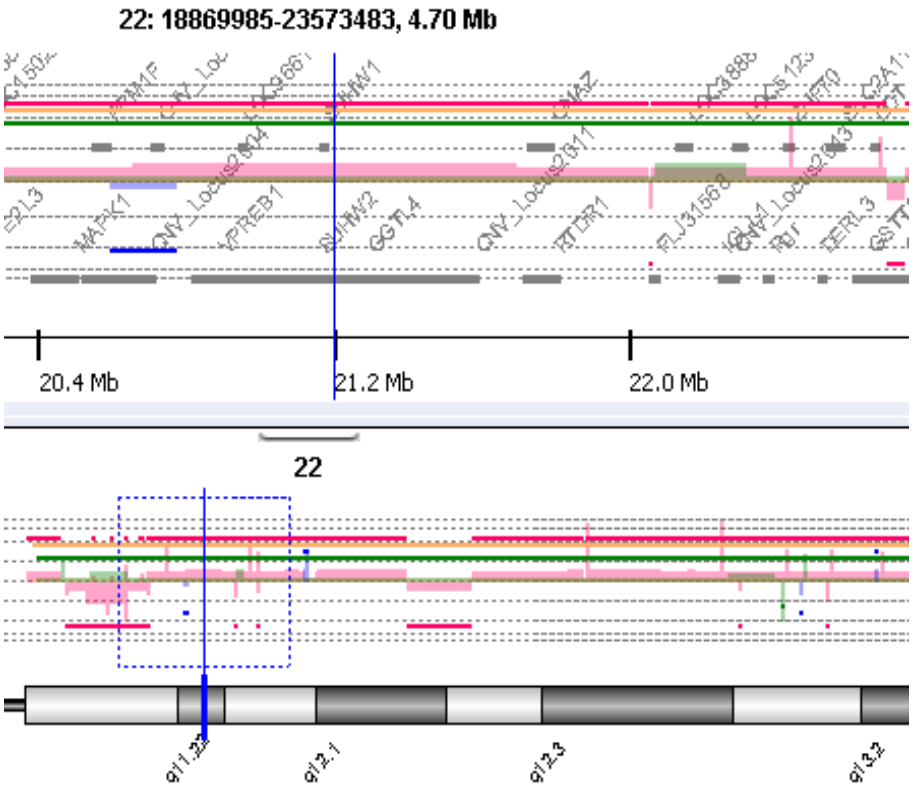


Figure 4-2 An example of ADM-1 output. Four samples (indicated by the colors red, orange, blue and green) were analyzed. All intervals are rendered as steps where the height of each step is equal to the average log ratio for that interval.

ADM-2

The ADM-2 algorithm identifies all aberrant intervals in a given sample with consistently high or low log ratios based on the statistical score. ADM-2 uses the same iterative procedure as ADM-1 to find all genomic intervals with the score above a user specified threshold. In ADM-2, the score represents the deviation of the weighted average of the normalized log ratios from its expected value of zero. This score is similar to the statistical score used in ADM-1 analysis, but ADM-2 incorporates quality information about each probe measurement.

Algorithm

The Quality-Weighted Interval Score algorithm (ADM2) computes a set of aberrations for a given sample. The overall recursive structure of the algorithm is the same as it is in ADM-1.

Log ratio error model

The only difference between ADM-1 and ADM-2 is in the definition of the score of the interval. ADM-1 considers only the log-ratios, while in ADM2 you also consider the log-ratio error information, hence the name Quality-Weighted Interval Score.

The following describes the ADM-2 score:

- Input is a vector of pairs $(v_1,q_1),(v_2,q_2),\ldots,(v_n,q_n)$, where
  - v_i is the log-ratio signal for the i-th probe
  - q_i is the log-ratio error for the i-th probe ordered
- Define $w_i = 1/(q_i)^2$

  Assume that under the null model, $v_i \sim N(0, 1/w_i)$ and the different v_i are independent of each other.
- Consider the weighted sum, for an interval I:

$$s(I) = \sum_{i \in I} w_i v_i$$

- The variance of *s(I)*:

$$var[s(I)] = var(\sum w_i v_i) = \sum (w_i^2 var\, v_i) = \sum w_i$$

- The ADM-2 interval score is

$$\varphi(I) = \frac{\sum w_i v_i}{\sqrt{\sum w_i}}$$

This score reflects the deviation of the sum *s(I)* from its expected value (0) in units of standard deviation. If the quality weight of each probe is the same then the score will be same as ADM-1.

CBS

CGH Analytics provides the Circular Binary Segmentation[5] (CBS) algorithm developed by *Olshen et. al.* for identification of genomic locations with significant adjacent mean copy number changes. While similar in scope to the ADM algorithms, CBS differs from aberration calling algorithms in general by partitioning the set of probes on the array corresponding to a given chromosome into subsets that share the same copy number. This yields a robust map of chromosomal copy number change points and is useful for putative aberration characterization, copy number estimates, and downstream analysis.

Change point identification

An effective approach to ensure that all change points are considered during segmentation is to recurse over any given segment length. Recursion within any given segment ensures that the relative locations of change points are preserved. Consider the following situation as illustrated in Figure 4-4:

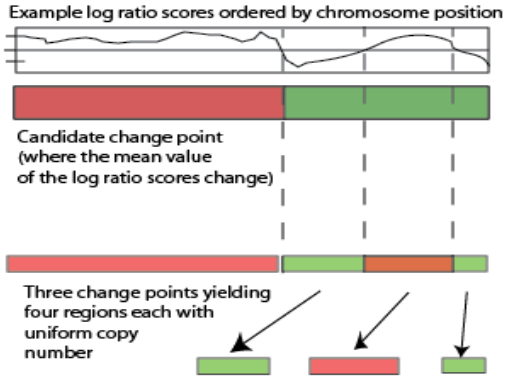


Figure 4-3 Change point identification by recursive segmentation. The first candidate change point results in the creation of two new segments. Within each segment further change points can be assigned, allowing the process to iterate on each subsegment until no additional change points can be found.

Binary Segmentation

The approach of chromosomal segmentation for aCGH analysis is to recursively split a position-specific set of log ratio intensities into two candidate subsets and compute the difference of means between the subsets as shown in the first candidate split in Figure 4-4.

Let i be an index of probe location across a chromosome of length n such that $1 \le i \le n$ with any subset denoted by $\{s,t\}$ such that $1 \le s \le i \le t \le n$. Additionally, let the score X_i be the log ratio intensity of probe i. If we let the mean score for any subset be proportional to the copy number, any position i may denote a change point log ratio means, and therefore in copy number. Such an assignment will create two syntenic regions with a difference in mean scores, where the range of the first subset is $\{s..i\}$ and the second is $\{i+1..t\}$. The respective ranges become the new segment boundaries for a new nested candidate breakpoint. This definition is applied recursively in each subset until no further segment can be split. Figure 4-4 illustrates the application of binary segmentation of a given chromosome.

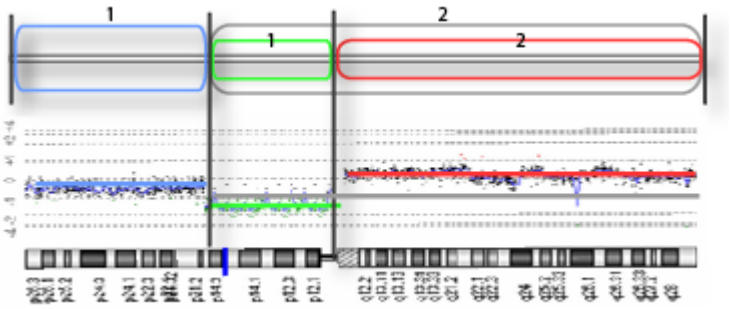


Figure 4-4 Binary Segmentation of a chromosome. The middle scattergram and lower chromosome panels show log ratio scores and chromosomal position, respectively. The top panel illustrates successive iterations of binary segmentation. In the outermost rounded rectangles, segmentation of log ratio means (indicated by matching colored bars in the middle panel) leads to two daughter regions (demarcated by the first three panel vertical bar and indicated by outer labels *1* and *2*), each potentially having uniform copy number. In the second iteration (indicated by the nested rounded rectangles), the segment originally labelled as segment *2* is reassigned by a new change point into two new segments (*1,2*), each with potentially uniform copy numbers. This process is iteratively applied until no more change points are found.

Circular Binary Segmentation

A potential complication of binary segmentation is that the algorithm as described may be insensitive to abrupt copy number changes where the range of such a change is minimized relative to the parent $\{s,t\}$ segment[5]. Under binary segmentation, such an event is modeled according to the relative means

of candidate subsets of the partition, which, for very small variances, may not pass threshold. This problem is associated with the strategy of identifying one change point at each iteration, and the decreased resolution becomes limiting in CGH copy number change map analysis. A more robust model can be achieved by specifying the presence of two change points defined as the start and end of a region of differential mean value.

Circular Binary Segmentation extends the concepts of binary segmentation to allow both binary and ternary segmentation of a segment $\{s,t\}$. Let $\{i..j\}$be a range of probe locations across a chromosome of length N such that $1 \le s \le i < j \le t \le N$. This definition allows the creation of a nested candidate subset $\{i..j\}$ with two flanking candidate regions. If the split is ternary in this way, the two end candidate subsets are joined together, treating the data as if it were a circle and allowing a mean value comparison between the $\{i..j\}$ subset and the joined $\{s..i,j..t\}$ subsets. Figure 4-5 illustrates the application of binary segmentation of a given chromosome.

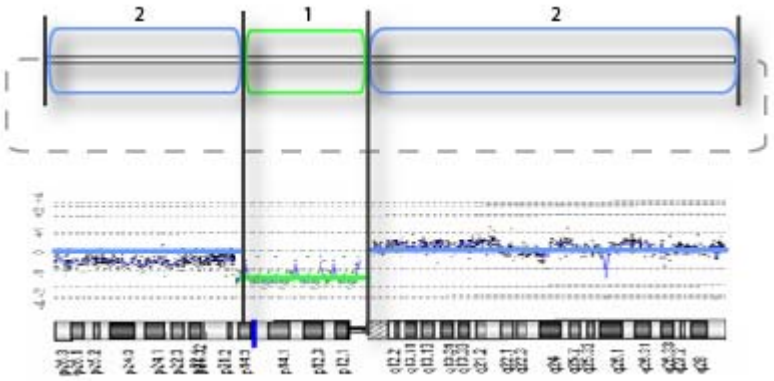


Figure 4-5 Circular Binary Segmentation of a chromosome. The middle scattergram and lower chromosome panels show log ratio scores and chromosomal position, respectively. The top panel illustrates a first iteration of the CBS algorithm. The center green rounded rectangle denotes a candidate subrange segmentation of log ratio means (indicated by a matching colored bar in the middle panel). Creation of this candidate segment leads to two candidate flanking regions (demarcated by the first three panel vertical bar and indicated by outer labels *2*). The dashed line in the upper panel connecting the end points of the chromosome indicates that both subsegments (labeled *2*) are joined together to evaluate a combined mean log ratio score (indicated by equivalent blue lines in the middle panel). This process is iteratively applied until no more change points are found.

CBS test statistic

CBS measures a difference of means between two adjacent candidate regions, one of which may be the result of a join of flanking regions resulting from a ternary split. If we assume the underlying model for copy number changes in partitioned subsets is based on the mean alone[6] (each region therefore having a common variance), a statistical test of the difference in means between syntenic regions formed by candidate partitioning is sufficient. The means of any two adjacent candidate regions suggested by the segmentation process can be computed from the partial sums of the member log ratio scores normalized

by the length of the respective subsets (where that length may result from two joined regions flanking the putative change points). A statistic $Z_{i,j}$ is calculated for a range given by $\{i,j\}$:

$$Zi,j = \left\{ \frac{1}{(j-i) + \frac{1}{(n-j+i)}} \right\}^{-\frac{1}{2}} \left\{ \frac{(Sj - Si)}{\frac{(j-i)-(Sn-Sj+Si)}{(n-j+i)}} \right\}$$

The candidate Z statistic Z_C is assigned to the $\{i,j\}$ segment such that

$$Z_C = max_{1 \le i \le j \le n} |Z_{ij}|$$

The null hypothesis of no change between adjacent regions created by assignment Z_C is tested under the assumption that all Z_{ij} follow some unknown distribution that can be estimated from the data. If Z_C exceeds the upper αth quantile, Z_C will split the parent region. Once a change point is detected in this manner, the procedure is repeated recursively in the newly created segment $\{i,j\}$ until no changes are detected in any of the segments.

Estimation of the log ratio score distribution

CBS generalizes the test statistic to non-parametric assumptions about the underlying distribution of the log ratio scores X^*_i by randomly permuting all X_i within $\{s,t\}$. If Z^*_C is the maximum Z^*_{ij} obtained from the shuffled log ratio scores, threshold parameters can be estimated without an assumption of normality. Generation of a robust p-value threshold can be achieved through a high number of permutations but is computationally expensive. A reduction strategy is therefore applied by dividing all X_i into K overlapping windows. These windows are then used as the candidates for change points.

Edge effect correction

CBS allows evaluation of candidate change points that are close to the parent segment boundaries. If the start change point for a given segment within $\{s,t\}$ given by the maximum Z_{ij} is close to s or the stop change point is close to t, a binary split may be the true result in place of the suggested ternary split. Such an event may occur due to proximity of local noise to an established change point and will result in a false positive identification. To correct for this, all ternary splits are evaluated for viability in the context of overlapping segment data (allowing the start position i to be equal to s or the end position j to be equal to t). If any given i is not a viable change point by itself for the segment $\{s..j\}$, the change point is removed from further consideration. The same test is applied for the overlapping segment $\{i..t\}$.

Change point pruning

Local data trends may lead to false positive identification. Following segmentation, a list of *C* change points is evaluated. For each segment bounded by consecutive *C* values, a segment log ratio score average is calculated. The sum of squared deviations *SS(C)* around this average are used to compute the best set of change points *SS(1)..SS(C-1)*. These change points are used to compile the subset which minimizes a function c under a pre-defined constant γ:

$c'=\min\{c:[SS(c)/SS(C)-1]<\gamma\}$

where γ=0.05 or 0.10. The change points that give this minimum $SS(c')$ are retained.

Interpreting the results

The recursion stops when no additional change points can be found. All intervals with constant copy number bounded by these change points found in this process are reported, and a plot is generated as output. The intervals are rendered as bars in the visualization panel. Identification of these intervals can be used for further exploration of aberration regions or for copy change analysis.

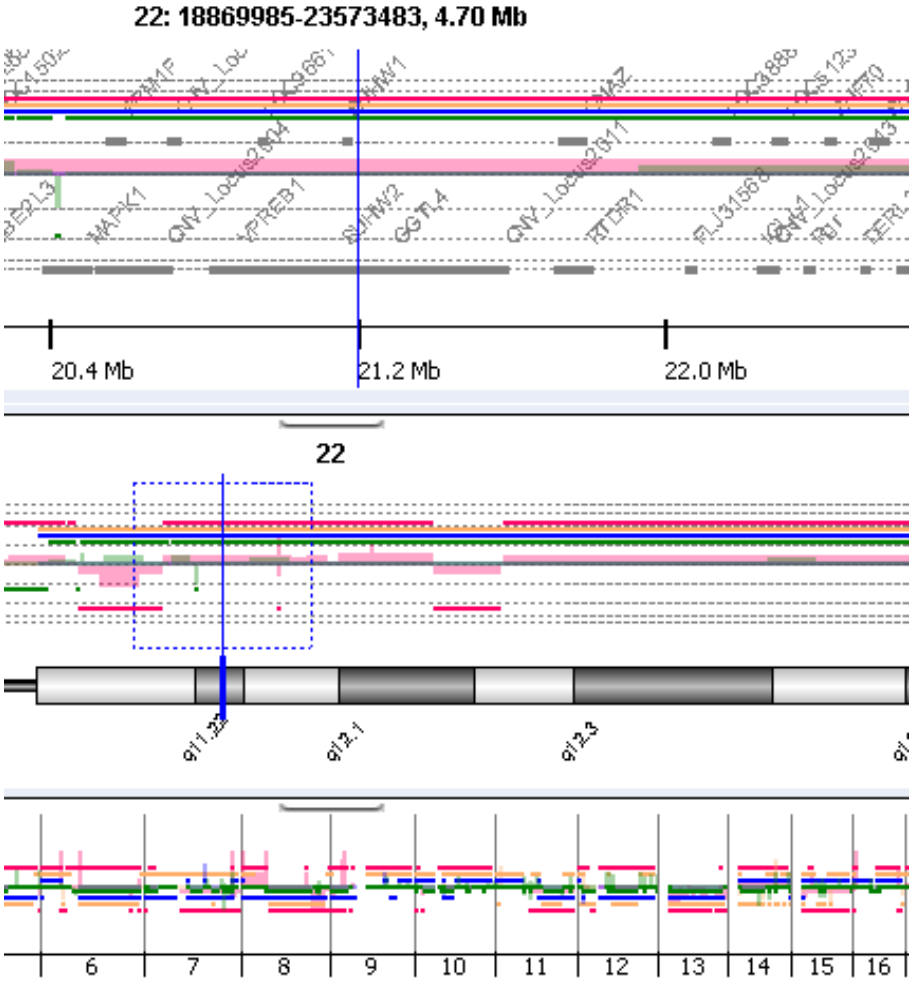


Figure 4-6 An example of CBS output. Four samples (indicated by the colors red, orange, blue and green) were analyzed. All intervals are rendered as bars indicating regions of constant copy number.

Parameters used

The default parameters follow the R-language implementation of this algorithm[7] with the exception of nperm, which was changed to 100 permutations. The values are listed below with a brief summary of their function within CBS:

Parameter	Default	Used	Description
alpha	0.01	0.01	The significance level for the statistical test.
nperm	10000	100	The number of permutations to run.
p.method	"hybrid"	"hybrid"	The method used for computation of the *p*-value.
kmax	25	25	The maximum width of segments for permutation using the "hybrid" method.
nmin	200	200	The minimum length of data for which the approximation of maximum statistics is used under the hybrid method.
window.size	NULL	NULL	Size of window used to speed up computations when segment size is too large. Not applicable using the "hybrid" method.
overlap	0.25	0.25	Proportion of data that overlap for adjacent windows.
trim	0.025	0.025	Proportion of data to be trimmed for variance calculation for smoothing outliers and undoing splits based on the standard deviation.

Parameter	Default	Used	Description
undo.splits	"none"	"none"	A character string specifying how change-points are to be undone, if at all.
undo.prune	0.05	0.05	The proportional increase in sum of squares allowed when eliminating splits if **undo.splits** is applied with parameter "prune".
undo.SD	3	3	The number of standard deviations between means to keep a split if **undo.splits** is applied with parameter "sdundo".
verbose	1	1	The level of verbosity for monitoring the program's progress. The value 1 specifies that the program will print the current sample's progress.

Centralization Algorithm

Many statistical and algorithmic approaches to aberration that call aCGH data assume that the data points are distributed around some zero value if no aberration occurs. Typically, aCGH data fluorescence ratios for each array are normalized by setting the average log fluorescence ratio for all array elements to zero. This may lead to erroneous aberration calls for highly-aberrant genomes such as those found in tumor samples. Given a data vector for a single sample or entire genome, this step attempts to find the best way to center the data by adding or subtracting the same constant to or from all log ratio measurements. Doing so will make the most-common-ploidy the new zero-point.

Specifically, define a score S for a possible centralization value c where $S(c)$ equals the number of probes that are not included all aberrations as called by aberration finding routines applied to the original log-ratios, shifted by c.

Try to find the value of c that minimizes score $S(c)$. That is, a value that minimizes the number of probes that are called aberrant. The search for the optimal value of c can be time consuming because you must run ADM-1 on each possible value.

In order to speed up the computation, without affecting the performance, contiguous probes are binned across the genome. In the user interface, you can choose a bin size for this algorithm (the default bin size is 10). In the default case each ten contiguous probes are averaged to reduce the number of probes used in the centralization procedure.

In the following example for the given array, the log ratio values are plotted in Figure 4-7. The plot of score $S(c)$ for different values of c is generated, and the plot is shown in Figure 4-8. The centralization algorithm defines the new zero where the center of the highest peak lies in Figure 4-8.

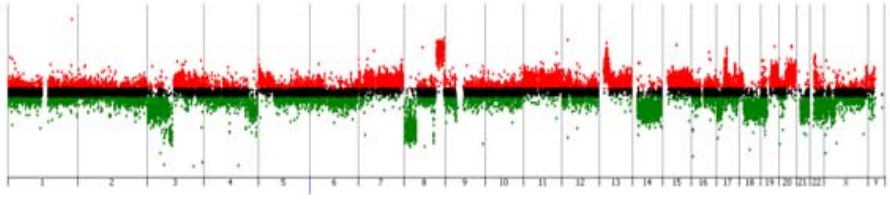


Figure 4-7 Log ratio values of an HT29 cell line in Genome view

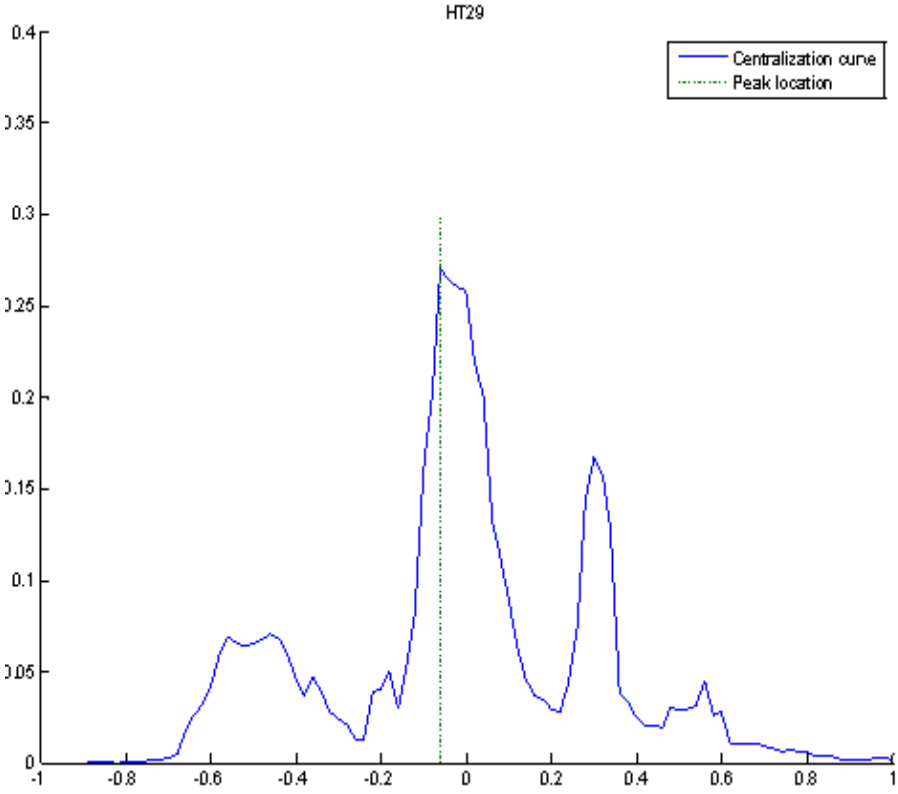


Figure 4-8 The plot of the score S(c) and the location of the peak for this centralization curve. For this example the log ratios will be shifted after centralization by 0.06.

CGH-Expression Analysis

The following table shows the type of analysis this software provides for performing a joint analysis of CGH arrays and corresponding Gene Expression arrays.

Input \ Output	Gene-centric	Aberration-centric
Single sample	-	Enrichment of extreme exp values, single sample
Multiple matched samples	Regional correlation scores	Enrichment of high-scoring genes, multiple matched samples
Multiple unmatched samples	-	Enrichment of high-scoring genes, unmatched samples

Enrichment of extreme expression values in a single sample

CGH-Expression Analysis can be computed for any sample in which both CGH and expression are measured. The output is a list genomic intervals for each sample with a statistical score showing the enrichment of over- and under-expressed genes in that interval. This analysis can be performed either on all aberrant intervals or on the whole genome.

NOTE When genes have multiple representatives in the expression data, these multiples have to be unified into a single representative for each gene before performing an enrichment analysis. You must choose Intra Array replicate combine for those expression arrays.

The following outlines the steps in this analysis.

Consider a single sample, S:

1. Rank all genes according to their expression values in S: g_1 (highest level of expression in S), g_2, ... , g_n.
2. Consider a genomic interval, I.
3. Compute the enrichment of over-expressed genes in I using the following process:

- Compute an occurrence vector

$$v(i) = \begin{cases} 1 \\ 0 \end{cases}$$

If $g_i \in I$, otherwise

- Let $B(v)$ equal the number of 1s in v.
- The mHG score [1, 2] for the vector v represents the rank imbalance of genes from I in the ranked list $g_1, g_2, \ldots, g_n$:

$$(*) \quad mHG(I) = mHG(v) = min_{1 \le i \le r} HGT(b_i(v); n, B, i)$$

where

$$HGT(b; n; B, i) = \sum_{k=b}^{min(i,B)} \frac{\binom{i}{k}\binom{n-i}{B-k}}{\binom{n}{B}}$$, r is a hard-coded boundary on the threshold optimization range and

$$b_i(v) = \sum_{k=1}^{i} v(k).$$

4 Using a symmetric procedure ranking genes in reverse order, intervals are found enriched with under-expressed genes.

5 Report intervals I with significant mHG(I), as determined by comparing to the user-defined threshold. Since the intervals are reported together with -log(mHG(I)), large numbers represent significant enrichment events.

Regional correlation scores with multiple-matched CGH and Expression samples

These scores are computed for all genes for which expression was measured and there is a well-defined known genomic location.

For this analysis the designs for the CGH and the expression arrays need not be the same, but the sample set must contain at least six pairs of matching CGH and EXP arrays.

It is assumed that a sample-by-sample aberration calling was run. The output will be transformed CGH data that, for each sample *S* and each genomic location *x,* there is an average value of log(R/G) signals in the interval that spans *x*, as determined in the aberration calling step. This processed CGH data is maintained in a matrix, denoted by *A*. which is in fact, the basis for the aberration summary graphics. Rows of *A* represent genomic locations and columns represent samples.

Consider a gene *g*, and the vector of expression levels of *g* measured over the entire set of samples: *E = E(g)*.

Now, consider a row of *A* that corresponds to the genomic location of gene *g*. This row forms a vector of DNA copy numbers at *g*, as denoted by *C = C(g)*.

The correlation between expression and DNA copy number at *g* is evaluated using three different scores:

- Student-t-test approach.

  Let P be the set of samples for which a region spanning g is called amplified. Let Q be the set of samples for which a region spanning g is called deleted. Let U denote the entire set of samples.

  Let

  - μ_P be the average expression value of g, for samples in P
  - μ_Q be the average expression value of g, for samples in Q
  - μ_{U-P} be the average expression value of g, for samples in U-P (the complement of P)
  - μ_{U-Q} be the average expression value, of g, for samples in U-Q (the complement of Q)

  Student-t statistics and the corresponding p-values are computed for one-sided alternative hypotheses:

  - $\mu_P > \mu_{U-P}$ (g's expression levels in amplified samples are larger than those in non-amplified samples).
  - $\mu_Q > \mu_{U-Q}$ (g's expression levels in amplified samples are larger than those in non-amplified samples).
- Pearson correlation of C and E and a corresponding p-value.

All three scores are reported for every gene, g. Output formats include graphical output as well as a text report reflecting the bottom panel table.

Enrichment of CGH/expression correlation

This can be computed for a set of samples for which both CGH and expression were measured. The output consists of a list of genomic intervals and a statistical score of the enrichment of CGH/expression correlation in each interval.

NOTE When genes have multiple representatives in the expression data these have to be unified into a single representative for each gene prior to performing the enrichment analysis.

Consider a set of samples for which regional correlation was computed as described above.

1. Rank all genes according to their correlation statistical score: g_1 (most significant regional CGH/expression correlation), g_2, ... , g_n.
2. Consider a genomic interval, I.
3. Compute the enrichment of correlated genes in I using the following process:
   a. Compute an occurrence vector

$$v(i) = \begin{cases} 1 \\ 0 \end{cases}$$

   If $g_i \in I$, otherwise

   b. Let $B(v)$ equal the number of 1s in v.

- The mHG score [1, 2] for the vector v represents the rank imbalance of genes from I in the ranked list $g_1, g_2, \ldots, g_n$:

$$(*) \quad mHG(I) = mHG(v) = min_{1 \le i \le r} HGT(b_i(v); n, B, i)$$

where

$HGT(b;n;B,i) = \sum_{k=b}^{min(i,B)} \frac{\binom{i}{k}\binom{n-i}{B-k}}{\binom{n}{B}}$, r is a hard-coded boundary on the threshold optimization range and

$b_i(v) = \sum_{k=1}^{i} v(k)$.

c Report intervals I with significant mHG(I), as determined by comparing to the user-defined threshold. Since the intervals are reported together with -log(mHG(I)), large numbers represent significant enrichment events.

Enrichment of external gene annotation.

You can compute this annotation for any external quantitative information about genes. Samples do not have to be matched. The output consists of a list of genomic intervals and, for each, a statistical score of the enrichment of the external quantitative annotation in that interval.

Example: The external quantitative annotation can be differential expression scores, between tumor and normal samples, as measured in an independent study.

NOTE When genes have multiple representatives in the expression data these have to be unified into a single representative for each gene prior to performing the enrichment analysis.

1 Rank all genes according to their external quantitative annotation: g_1 (highest value, e.g. most over-expressed or most under-expressed in tumor), $g_2, \ldots, g_n$.

2 Consider a genomic interval, I.

3 Compute the enrichment of high scoring genes in I using the following process:

a Compute an occurrence vector

$$v(i) = \begin{cases} 1 \\ 0 \end{cases}$$

If $g_i \in I$, otherwise

b Let $B(v)$ = the number of 1s in v.

c The mHG score [1, 2] for the vector v represents the rank imbalance of genes from I in the ranked list $g_1, g_2, \ldots, g_n$:

$$(*) \quad mHG(I) = mHG(v) = min_{1 \le i \le r} HGT(b_i(v); n, B, i)$$

where

$$HGT(b; n; B, i) = \sum_{k=b}^{min(i,B)} \frac{\binom{i}{k}\binom{n-i}{B-k}}{\binom{n}{B}},$$ r is a hard-coded boundary on the threshold optimization range and

$$b_i(v) = \sum_{k=1}^{i} v(k).$$

4 Report intervals I with significant mHG(I), as determined by comparing to the user defined threshold. The intervals are reported together with -log(mHG(I)). Therefore large numbers represent significant enrichment events.

Common Aberration Analysis

Common aberration analysis is used with multiple samples to identify genomic intervals that have statistically significant common aberrations. The framework of this type of analysis has the following steps:

1. Apply one of aberration detection algorithms (ADM-1or ADM-2) to a set of samples, S_1, S_2, ... S_n, to identify a set of aberrant genomic intervals in each sample that have a score above a user-specified threshold.
2. Construct a set of candidate genomic intervals for common aberration analysis using all intervals identified in Step 1, and all possible intersections of those intervals.
3. For each candidate interval A, and each sample S, compute a statistical score representing the significance of aberration A in sample S.
4. Compute a combined common aberration score for each candidate interval A, using scores computed in Step 3, to test the hypothesis that A is a common aberration for samples S_1, S_2, ... S_n or a subset of these samples.
5. Report candidate intervals with common aberration scores above a user-defined threshold and the corresponding supporting samples. You can choose to report only the most significant common aberrant interval from each set of overlapping intervals to make the report more concise.

Steps 3 and 4 are implemented in different ways in the 't-test' and in the 'context corrected' common aberration analysis, as described below.

T-test analysis

In Step 3 of the common aberration detection framework, the significance of aberration A in sample S, $P(A,S)$ is computed applying interval scoring of the algorithm selected in Step 1.

In Step 4, two analysis are performed. One analysis tests for common amplification, and the other for common deletion.

To compute the statistics T_n, let $\bar{\mu}$ and S^2 be the estimated mean and variance of per-sample statistical scores *{P(A, S_1), P(A, S_2), ..., P(A, S_n)}*:

$$\bar{\mu} = \frac{1}{n}\sum_{m=1}^{n} P(A, S_m) \ ,$$

$$S^2 = \frac{1}{n-1}\sum_{m=1}^{n} (P(A, S_m) - \bar{\mu})^2 \ .$$

Then T_n is

$$T_n = \frac{\bar{\mu}}{S/\sqrt{k}} .$$

The corresponding common amplification p-value, p_{amp}, is calculated as *Prob{t>=T_n}* based on Student's t-distribution with n-1 degrees of freedom. Score of candidate interval *A* is *(-log10(p_{amp}))*. To test for common deletion, the corresponding common deletion p-value, p_{del} is calculated as *Prob{t<=T_n}* based on Student's t-distribution with n-1 degrees of freedom. Score of candidate interval *A* is *(-log10(p_{del}))*.

Context corrected analysis

In Step 3 of the common aberration detection framework, the significance of aberration *A* in sample *S, P(A,S)* is computed using the chromosomal or whole genome context of aberrations distribution in sample *S*. Namely, for each candidate interval *A*, and each aberrant interval *I* (longer and higher than *A*) in sample *S*, compute the probability that *A* overlaps with *I*:

$$P(A \subseteq I) = \frac{|I| - |A| + \varepsilon}{|G| - |A| + \varepsilon},$$

Where $|I|$ denotes the size of interval *I* in Mb or in number of probes, $|G|$ is the size of whole chromosome (chromosomal context) or the size if the whole genome (whole genome context), and ε is a small constant added to avoid dividing by zero, ε is proportional to $1/|G|$.

Context specific score of candidate interval *A* in sample *S* is computed as a probability that *A* is amplified in *S*, $P_{amp}(A, S)$ or as a probability that *A* is deleted in *S*, $P_{del}(A, S)$. To compute P_{amp}, consider the set $I = \{I_m\}$ of aberrant intervals in *S* with height (height is the ADM computed height of the interval)

and length greater or equal than the height and length of the interval A, where height of A is the minimum height of all aberrant intervals in S that intersect A.

Then

$$P_{amp}(A, S) = \sum_{1} P(A \subseteq I_m)$$

$P_{del}(A, S)$ is defined is a similar way, using the set I of aberrant intervals in S with height less than or equal to the height of the interval A, and length greater than or equal to the length of the interval A. In this case, the height of A is the maximum height of all aberrant intervals in S that intersect A.

In Step 4 of the common aberration detection framework, you test for common amplification and common deletion. To test for common amplification, scores $P_{amp}(A, S_i)$ are sorted in increasing order. The probability that interval A is aberrant in all k samples, $S_{i1}, S_{i2}, ..., S_{ik}$, is estimated by applying Chernoff bound for each k between 1 and n. By Chernoff bound,

$$\text{Probability } (k \text{ or more aberrations}) \leq \frac{e^{(k-\bar{\mu})}}{(k/\bar{\mu})^k} = P_k,$$

where $\bar{\mu} = nP_{amp}(A, S_{ik})$

The score of candidate interval A is $max_k(-log10(p_k))$, and the corresponding set of supporting samples is the set $S_{i1}, S_{i2}, ...S_{ik}$ for k that yields this score.

To test for common deletion, apply this analysis to deletion scores $P_{del}(A, S_i)$.

Derivative Log Ratio Spread

To make aberration calls, CGH Analytics software needs a measure of log ratio noise. Therefore, a measure of the minimum log ratio difference is necessary to make reliable amplification or deletion calls. In the previous version of the software, CGH Analytics 3.1, a set of calibration arrays consisting of a normal sample vs. normal sample was used. These arrays have extensive stretches of chromosomes along which the genomic copy number should remain constant.

The shape and the spread of the log ratios are calculated from the calibration arrays and those statistics are passed to sample arrays with chromosomal abnormalities to make amplification or deletion calls. Experience shows that this method often underestimates the noise of the log ratio for the sample array. It is better to estimate the noise from the selected array itself. Observations show that instead of calculating the standard deviation of the log ratio, a more robust estimate of noise is attained by calculating the spread of the log ratio differences between consecutive probes (*dLRsd*) along all chromosomes, divided by sqrt(2) to counteract the effect of noise averaging.

If you estimate the noise using the derivatives of the log ratio as described above, you can also eliminate the need for the calibration array.

Even highly anuploid samples have chromosomes with extensive stretches along which the genomic copy number is constant, or nearly so. In such constant-copy-number regions, the true log ratios are constant, although not necessarily zero. Estimations of log ratio error, and therefore the minimum log ratio difference required to make reliable amplification or deletion calls, is based on observations of the variation in such constant-copy-number regions. The dLRsd metric is an attempt to quantitate such "eyeball" estimates.

For normal samples, this metric is the width of a self-self distribution, and should be below 0.2 log units—closely approximating the spread of log ratio in the calibration array. It will be somewhat greater for abnormal chromosomes because (a) the width of regions of constant copy number different from two will include both noise and the variable log ratio compression observed for many probes, and (b) constant-copy-number regions can include single probes or small regions where the copy number varies.

To make this metric more robust and a true measure of noise, you must remove the outliers from the constant-copy-number regions. To remove outliers, you can use IQR (Inter Quartile Range) with appropriate scaling to calculate the spread of the distribution instead of calculating the standard deviation of the derivative of log ratio directly.

By default, CGH Analytics uses individual arrays for determining the mean. However, if you have specific calibration arrays, such as self-self microarrays, or just arrays without CGH anomalies, they can be used for accurately estimating array statistics for arrays without aberrations. If a calibration array is used for this purpose, the underlying assumption is that the distribution of noise is the same in both the calibration array and the sample array. For most applications, Agilent recommends that you not use a calibration array in determining log ratio noise.

To select arrays for calibration, right-click a microarray header in the Probe table to invoke a shortcut menu. Calibration is one of the choices available from the menu.

The metric can be calculated from either the calibration array or the sample array, and the value of this metric (*dLRsd*) is used to compute Z-normalized values of the log ratios which are input to both the aberration algorithms. If you select a calibration set, then the algorithm follows the first option, otherwise it follows the second option.

- First Option (using calibration set): Calculate the spread (*dLRsd'*) and the mean for each calibration array. The mean should be close to zero. Then count *R*, *R'*, and *N* from the Z-normalized data of that array. If there is more than one array in the calibration set then compute *R*, *R'*, and *N* for each array, and finally compute the sums *Rtotal*, *Rtotal'*, and *Ntotal* to pass the ratios *Rtotal*/*Ntotal* and *Rtotal'*/*Ntotal* to the Z-score aberration algorithm. Next, calculate the derivative log ratio spread (*dLRsd*) for the sample array, and use it as the spread of log ratio noise in the Z-score and ADM-1 algorithms. In this case, assume that the spread of the log ratio in a calibration array and a sample array (array with chromosomal aberration) is the same.

- Second Option (no calibration set): If you do not select a calibration array for determining aberration calls, then compute the spread of the derivative of log ratio (*dLRsd*) from the sample array. Take mean = 0. In the previous release (CGH Analytics 3.1) if no calibration arrays were selected, you computed the above statistics for the sample arrays, and expected any genetic anomalies to cause only a small perturbation. Since only some of these chromosomes will show anomalous behavior, you can expect only a small contribution to amplifications and deletions when compared to the overall global behavior. Further investigation revealed that this is not a very safe assumption so the algorithm has been changed to compute the statistics from the derivative of log ratio instead of the log ratio from the sample array. Compute R/N and R'/N for the Z-score algorithm from the derivative of the log ratio after proper Z-normalization of that vector.

Error Model and Combining Replicates

Import the log ratio and the log ratio error values for each feature from the Feature Extraction output text file. For combining the log ratios of replicated features, assign a quality to each feature. Instead of directly averaging the log ratios of replicated features, put a weight on each feature before combining. The weight is proportional to quality, and quality is defined as the inverse of square of log ratio error.

If you combine replicated probes in an array (intra array) or within replicated arrays (inter array), then the log ratio and the log ratio error is combined as follows:

Define a weight w_i for *each* probe to be $w_i = 1/e_i^2$

That is, the noisier a given probe is, the smaller is its weight. The error, e_i, is defined as the maximum between the log ratio error, lre_i, of that probe and the spread of derivative of log ratio, *dLRsd*, for that array, i.e. e_i = $max(lre_i, dLRsd)$. The *dLRsd* is described under See 'Aberration Algorithms Overview' on page 176 for more information..

Define the quality-weighted average log ratio for replicated probes as

$$LR_{combined} = \frac{1}{W}\sum_{iel} w_i l_i \quad \textbf{(1)}$$

where

$$W = \sum_{iel} w_i$$

Next, estimate the Log Ratio Error of the above mean. The combined Log Ratio Error of the combined log ratio is captured in

$$\sigma_{combined} = \left(\sum_{iel} 1/e_i^2\right)^{-1/2} = 1/\sqrt{W} \quad \textbf{(2)}$$

When combining dye-swapped arrays, first combine any replicates of each polarity using to the above equations. Then take the average log ratios of each probe present in dye swaps without error weighting. The probe should be filtered out of the combination only after the probe is completely filtered out in either polarity.

Recompute the Log Ratio Error average after completing any averaging or dye swapping. Also, estimate the Log Ratio Error of each probe in each polarity using Equation 2.

Once you have the Log Ratio Error for each probe and polarity, add the Log Ratio Error of the two polarities in quadrature and assign the value as the combined Log Ratio Error of the dye-swapped pair(s).

$$\sigma_{combined}(dyeSwap) = sqrt((\sigma^2_{polarity1} + \sigma^2_{polarity2})/2) \quad \textbf{(3)}$$

NOTE

In the current release of this software, you cannot combine virtual arrays (i.e. nested combines of arrays). So if you create two or more virtual arrays by inter-array combining using a particular attribute from the user interface and then pick a second attribute for inter-array combining, the software decouples the first set of virtual arrays then recombines them using the second attribute chosen.

Fuzzy Zero

The aberration-calling algorithms ADM-1 and ADM-2 identify aberrant genomic intervals whose copy numbers are significantly different between the sample and reference channels - that is, their measured log ratios are significantly different from zero. An interval is considered to be aberrant if the mean of the log ratios of probes in the interval differs from zero by more than a user-specified number of standard deviations.

The ADM-1 and ADM-2 algorithms estimate the standard deviation of the mean log ratio of an interval using a statistical error model that treats probe to probe errors as independent. In many samples, the assumption that the log ratio errors of successive probes are independent is not in fact valid. The errors of the probes are often correlated over wide genomic intervals, and the ADM algorithms therefore underestimate the error for long intervals. Long aberrations with low average log ratios are thus often incorrectly deemed significant.

Fuzzy zero correction applies a "Global error model" to all aberrant intervals identified in ADM-1or ADM-2 analysis. The global error model uses a more realistic error model to avoid erroneous aberration calls when the errors are correlated.

For the global error model, we assume that there are two independent sources of noise contributing to the total noise of the intervals. A local probe-to-probe noise, σ^1_{Local}, which is not correlated between different probes along the interval as described above, and a global noise, σ_{Global}, which is correlated between probes in an interval. The global noise component, σ_{Global}, is calculated as the variation of the average log ratios in large genomic intervals. As local probe-to-probe noise, σ^1_{Local}, is not correlated between different probes, when k probes are averaged, we assume that the local noise is reduced by a factor of $1/\sqrt{k}$. Thus,

$$\sigma^k_{Local} = \sigma^1_{Local} \Big/ \sqrt{k} \qquad \textbf{(4)}$$

The score of interval I under the global error model, $S_g(I)$, is

$$S_g(I) = \frac{h}{\sqrt{(\sigma^k_{Local})^2 + \sigma^2_{Global}}} \qquad \textbf{(5)}$$

here h is the average log ratio of all probes in the interval I. If ADM-2 algorithm is used, h is the quality weighted average log ratio of all probes in the interval I.

Using α to denote $\left(\frac{\sigma_{Global}}{\sigma^1_{Local}}\right)^2$ we derive

$$\sigma^k = \sigma^1_{Local}\sqrt{\frac{1}{k}+\alpha} \quad \textbf{(6)}$$

and

$$S_g(I) = \frac{h}{\sigma^1_{Local}\sqrt{\frac{1}{k}+\alpha}} \quad \textbf{(7)}$$

Fitting the model parameters

For a given logratio vector v of length N (for a particular sample), σ^1_{Local} equals the Derivative Log Ratio Spread (*dLRsd*):

σ^1_{Local} =*dLRsd*(v)

See 'Derivative Log Ratio Spread' on page 204 for more information.

Then α is estimated using the following iterative procedure:

- Start with an initial estimate of α, α_0= 0.01, and set v_0=v.
- At each iteration i, repeat:
- Find all aberrant intervals I in v_i with the score $S_g(I)$ (**4**) above the user defined threshold, T.

  Note that the score $S_g(I)$ (**4**) depends on the current value of α_i. This set of aberrant intervals is considered as the signal component of the data.
- Compute the residual vector v^r.

  To compute v^r, we subtract from v_i the heights of each aberrant interval I. Namely, we subtract from each probe in I the height h of the aberration containing the probe. The resulting vector v^r represents the current estimate of the noise in the data.
- Estimate the combined noise σ_i^k from the residual vector v^r.

To estimate the combined noise σ_i^k, we bin consecutive probes into bins of size $k = \sqrt{N}$. Then we derive a binned vector u_k, where each element of u_k is the average log ratio of all probes one bin. We estimate σ_i^k by computing $dLRsd(u_k)$. To make this estimation more robust we repeat the binning using 10 different starting positions of the first bin. The final estimation of σ_i^k is the median of these 10 different estimations.

- Compute the new α_i+1 from equation (**6**) based on the current estimate of σ_i^k.
- Set v_i+1 =v^r.
- Continue the iterations until the process converges, i.e. $|\alpha_i - \alpha_{i-1}| < 0.001$, or 10 iterations were made.

Triangular Smoothing

When visualizing or analyzing aCGH data, it is common to smooth the data using a moving average. However, the moving average approach is not the optimal means of reducing the noise associated with each independent point, since it can minimize localized copy number changes and obscure individual points. A good compromise can be achieved between reducing the noise of the individual points, and still remain sensitive to true localized, or small-scale, variations in the data.

Agilent has introduced a triangular smoothing function that, like other shaped smoothing, has a peak (maximum weight) at the center and falls off to zero with increasing distance from the center. Rather than smoothing, the data at each point, triangular smoothing averages the value of a point of interest and a number of its neighboring points.

The number of neighboring points depends on the type of moving average. If we have point input (Pt input) from UI, the number of points that are averaged is kept fixed, say at 3, 5, 7, 9, or 11 points, and each point is given equal weight. In other cases, a window of constant width (specified from UI in Mb or Kb) moves across the data and centers on the point of interest. All points within its range are averaged to yield the moving average value for the point.

There are two potential problems in smoothing using fixed width.

- First, using a fixed window width causes a variable number of probes to be averaged at each smoothed point. Therefore the degree of averaging varies from probe to probe, depending on how many probes happen to be in the fixed-width window. Since varying numbers of measurements contribute to each smoothed point, the degree of statistical noise reduction also varies for each point. This can complicate the error analysis.
- The second problem with fixed window sizes arises on so-called "zoom-in" arrays, in which some genomic regions of interest are covered more densely than neighboring regions. For such arrays, the appropriate window size can vary greatly between the different genomic regions. Smoothing windows that are appropriate for sparsely tiled regions will obliterate all structure in densely tiled regions, whereas windows appropriate to densely tiled regions will perform practically no averaging at all in sparsely tiled regions.

These problems with fixed-size smoothing windows can be avoided in triangular smoothing by using smoothing windows containing a fixed number of probes, regardless of the total range of sequences those probes span. This respects the fact that nearby probes are more relevant than distant probes to the average at any point.

The concept is illustrated in Figure 4-9. Fixed count smoothing includes the same number of points in each average, but weights probes far from the averaged point as much as points near to it. In the new triangular smoothing with Pt input, a symmetric window around the averaged point is enlarged until it contains the number of points chosen for the fixed size window. These points may be on one side of the averaged point or on both, depending on the probe density around the averaged point. These points are then weighted appropriately for triangular smoothing function, depending on their distance from the averaged point.

When user inputs number of points from the user interface, it uses a variable window width, which is chosen to be the smallest window, symmetrical about the averaged point, which includes the specified number of points. Figure 4-9 illustrates the application of this method for Pt input.

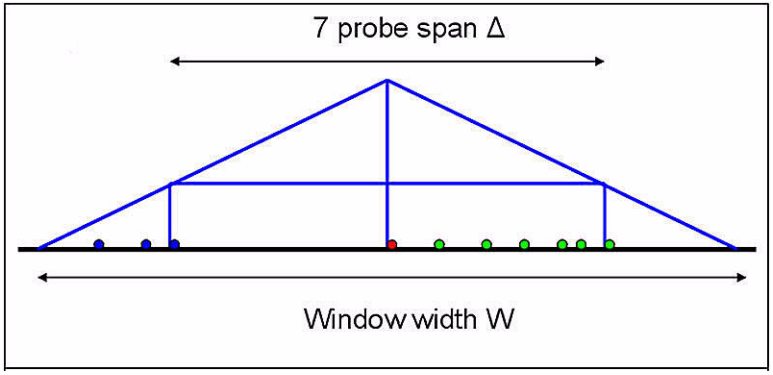


Figure 4-9 Triangular smoothing with Pt input.

Smoothing is applied to a region of varying probe density. The effective width of the smoothing window, W, depends on the length of the smallest symmetrical region (Δ) that includes the specified number of points. The weight given to each point is proportional to the height of the triangle at that point.

The weights assigned to the log ratio values of these probes are given by the equation:

$$w(x) = (W - |x|)/W^2 \quad \textbf{(1)}$$

where the effective window width, W, is determined by the span, Δ , of the symmetrical region spanning the specified number of points by the equation:

$$W = \Delta/(2 - \sqrt{2}) \quad \textbf{(2)}$$

Z-Scoring for Aberrant Regions

This method identifies all aberrant regions in a given sample using statistical analysis based on hypergeometric Z-scores. The scoring method has essentially two steps. In the first step it identifies the total number of probes with log ratios significantly different from zero in a normal sample (a sample with no genetic abnormalities). These probes are referred to as aberrant probes and the normal sample is referred to as the calibration sample. In the second step, the method determines if the actual sample of interest has a significantly higher proportion of aberrant probes in any given region of the genome than the proportion of total aberrant probes in the normal (calibration) sample used in the first step. If it identifies any such region, then this region gets a higher score and it is called as an aberrant region. The two steps are explained in more detail in the following section:

Step 1 Calibration

Each CGH ratio is converted to a log ratio and then Z-normalized by computing the usual formula:

$$Z(x) = \frac{x - \mu}{\sigma}$$

where x is the log of a measured CGH ratio, μ is the mean and σ is the noise level of the population of such log ratios. You can calculate μ and σ from an entire array or over the entire collection of microarrays. The averaging method used may depend on the context. Chromosomes X and Y are not included in the calibration of μ and σ since gender differences between arrays may offset the statistics even for arrays that were designed to be calibration arrays.

Furthermore, each Z-normalized value can be classified as significantly above or below the mean by using a Z cutoff, Z_C. This cutoff can be supplied as a user-specified value. In essence, you are simply stating that you consider log ratios greater than or less than Z_C to be outliers from the normal population of log ratios.

Important: Z_C is not a cutoff used to *filter* data. It is a cutoff for *classifying* data as being significantly above or below the mean. To avoid reinforcing the idea that this value filters Z scores, the UI refers to Z_C as threshold.

As part of the computation, also count the number of entries in three classes:

- R = the number above the positive cutoff (Z_C)

- R' = the number below the negative cutoff ($-Z_C$)
- N = the total number of measurements

These Z-normalized values and counts can be pre-computed and reserved for subsequent calculations in step 2. The values computed in step 1 would only need to be recomputed if a different Z_C were desired. Even so, μ and σ can still be reused without computation.

Ideally, these global statistics would be computed for samples that contain no genetic anomalies, so that μ and σ represent the distribution of a non-diseased sample. If no calibration array is selected, the statistics are calculated from the sample array itself. Compute R, R', and N from the derivative of the log ratio (after proper normalization) and using threshold from UI.

If you want to determine the pre-computed statistics more accurately, you can specify the specific arrays to be used in this calibration step.

Step 2 Computation

While computing a moving average, log ratios are averaged over a small subset of points. This moving average window, w, may be simply a number of adjacent measurements or it may be over a positional window (such as every megabase). For each of these windows there are w entries. The objective is to analyze the over- or under-abundance of log ratios that deviate significantly from the mean from step 1 and lie inside the window. For this smaller subset, compute the same three counts as in step 1 using exactly the same cutoff values, but in this case, only for the points within the averaging window, w:

- r = the number above the positive cutoff (Z_C) in w
- r' = the number below the negative cutoff ($-Z_C$) in w
- n = the total number of measurements in w

Now, compute an exact Z-score that measures the significance of this over abundance or under abundance in w of significant positive deviations as

$$Z(w) = \frac{\left(r - n\frac{R}{N}\right)}{\sqrt{n\left(\frac{R}{N}\right)\left(1-\frac{R}{N}\right)\left(1-\frac{n-1}{N-1}\right)}}$$

You can compute the same formula for r' to obtain a score for negative deviations. This score can be plotted in a manner analogous to a moving average. This would identify statistically significant groups of probes that appear to deviate from the typical distribution of values for the given microarrays. In this way, it provides some predictive power to call amplification or deletion events in CGH studies.

This computation was designed to be done easily in parallel with a moving average. It can also be done independent of any other calculations.

Essentially for each moving window, the Z-score algorithm takes the points within that window as a sample and computes a hypergeometric Z-score that measures the significant number of outliers – points that lie above (positive values) or below (negative values) the threshold or z-level. Note that a point that is slightly beyond the threshold is counted the same as a point that is considerably beyond the threshold.

Z-scores are plotted to indicate statistically significant groups of probes that appear to deviate from the typical distribution of values for the given microarrays. They provide some predictive power to call amplification or deletion events in CGH studies. You should adjust the cut-off, Z, should be adjusted appropriately based on your visual analysis of amplified and deleted regions in the chromosomes.

Also note that the score that is plotted has nothing to do with the log ratio, and you should not expect the values to necessarily line up with the log ratios. It simply represents a statistical measure of aberration that can be used to track the distribution of outliers, which you can usually see by comparing the scatter plot to the Z-score plot.

One final interesting point is that if you set the threshold too low (i.e. cut-off is too small), most of the data points on an array are outliers (i.e. very high values of R and R'), and you will probably get a Z-score of zero. If on the other hand, you set the threshold too high, none of the points are outliers and again, the Z-score will be zero. Usually a threshold of 2-3 is the best setting. You can go slightly higher if you want to look for very deviant aberrations, but going too high will show no aberration.

Visualization

When the final Z-scores are computed, they can be plotted as a line graph similar to the moving average. To enhance the visibility of the plots and to distinguish them from the moving average, the graphs are filled from the

origin. As a further refinement, the filling is alpha-blended for transparency. When plotting multiple microarrays simultaneously, this minimizes obscuring of the data and allows you to detect overlaps. For two or three simultaneous plots, it is often possible to distinguish the various intersections based on the color blending.

The Z-score is also reduced by a factor of 10, thus allowing you to read the actual underlying value by interpolating the location on the graph scale (2, 4) and simply multiplying that value by 10. All Z-scores are positive, and those shown in the negative direction are actually positive Z-scores for decreased copy number. They are shown on the same 1/10th scale, but the implied negative sign should be ignored in this case.

It can still be difficult to read small segments of significant Z-scores, particularly in the overview. So as a further aid, *side-bars* are provided. Usually, these side-bars are not overlapped and provide a quick means for locating interesting anomalies in single microarrays. They also provide better separation when looking at multiple microarrays. Normally, the side-bars are stacked, but if there is insufficient room they may overlap. This allows you to see clear areas of interest. You can always manipulate the display to enlarge the available space in order to separate the side-bars as needed.

5 Reference

Macintosh-Related Issues

Software Requirements

Starting with version 2.3a, CGH Analytics is supported on the Macintosh (Mac) platforms with the following restrictions:

- You must be using a Java 1.5 run-time.
- You must be running Mac OS X 10.4 (Tiger).
- You must have at least 10GB of free disk space.
- You should have at least 2GB of RAM.

Contextual Menus

Effort have been made to support proper Mac mappings of shortcut menus (also known as contextual menus). If your system uses a two-button mouse, use the right mouse button to invoke the Web Searching menus. In lieu of a two-button mouse, use the Control key and left or single mouse button to invoke these menus by holding down the Control key while you click the mouse button.

Edit Menu Items

Converting between Java image data types and the Mac environment is very different, so the current Mac version does not support copying the panels to the clipboard. The Edit menu is empty on the Mac version but not on the PC version.

Microarray QC Metrics

These metrics are appropriate only for Agilent catalog oligo CGH arrays. You can use them to assess the relative quality of the data from a set of microarrays in an experiment. In some cases, they can indicate potential processing errors that have occurred or suggest that the data from particular microarrays might be compromised.

Many factors can influence the normal range of these metrics including the biological source and quality of the starting DNA, processing protocols, scanner sensitivity, and image processing. The guidelines presented here represent the normal ranges that Agilent has observed when analyzing well established and characterized cell lines using standard Agilent protocols.

Metric	Excellent	Good	Poor
BGNoise	< 5	5 – 10	> 10
Signal Intensity	> 150	50 – 150	< 50
Signal to Noise	> 100	30 – 100	< 30
Reproducibility	< 0.05	0.05 – 0.2	> 0.2
DLRSpread	< 0.2	0.2 – 0.3	> 0.3

BGNoise

For each channel, this metric is calculated as the standard deviation of negative control probes after rejecting feature nonuniform outliers, saturated features, and feature population outliers.

If the noise is high, examine the array image for visible non uniformities. High background noise is often introduced during slide handling or from contaminated buffers.

Signal Intensity

For each channel, this metric is calculated as the median background-subtracted signal after rejecting nonuniform outliers and saturated features.

If the signals are too low, fail the array. If the signals are marginal, expect noisy results. Low signals can result from poor quality input DNA or from losses during labeling and cleanup

SignalToNoise

For each channel, this metric is calculated as the Signal Intensity divided by BGNoise.

If this ratio is low, fail the array. A ratio over 100 indicates that the DNA quantity is sufficient and that no significant error was introduced during hybridization, washing, or scanning.

Reproducibility

For each channel, this metric calculates the Median %CV of background-subtracted signal for replicate non-control probes after outlier rejection. You want to exclude from the calculation any probe sequences for which the average signal of a probe sequence is below the additive noise of that channel, i.e. Average(BGSubSignal) * Multiplicative error < Additive error / Dye Norm Factor. After rejecting nonuniform outliers and saturated features, at least three probes are required to calculate the CV for that sequence. Calculate the median of the CVs of the remaining sequences. If the number of sequences that pass the filter is less than 10, do not calculate this metric.

High scores on this metric may signal catastrophic failures (e.g. that the slide leaked or fell out of the rotisserie). Large bubbles cause moderate values on this metric, but do not compromise the results significantly.

DLRSpread

This metric calculates probe-to-probe log ratio noise of an array and hence of the minimum log ratio difference required to make reliable amplification or deletion calls. Noting that most pairs of probes adjacent to each other along a chromosome have the same true copy number, the large majority of differences between log ratios of adjacent probes are just noise. The metric is computed as: DLRSpread = IQR(dLR) / 4*erfinv(0.5), where dLR is an array of differences between log ratios of adjacent probes, erfinv is the Inverse Error Function and IQR is Inter Quartile Range.

Model System Metrics

Three metrics are applicable only to model systems. In CGH Analytics version 3.2, the only model system supported is a male (XY) vs. female (XX) comparison where the female sample is labeled with CY3 (green) and the male sample with CY5 (red). The model system metrics are meant to be a means of helping to validate the performance of the aCGH system initially and over time. The metrics are described in the following paragraph.

NOTE These metrics are *only* meaningful on arrays that are analyzing normal male/female samples with the correct labeling (female with CY3).

Metric	Excellent	Good	Poor
AreUnderROC	< 0.95	0.85 – 0.95	> 0.85
MedianDiff	> 0.9	0.8 – 0.9	< 0.8
ErrorFraction	> 0.05	0.05 – 0.1	< 0.1

AreaUnderROC

Sort the log ratios in ascending order for the entire array. Each log ratio in the data set comes from an X-probe or autosome. If it is an X-probe, it contributes to the number of True Positives (TP). If it is an autosome, it contributes to the number of False positives (FP). So for each log ratio, start from the lowest and continue incrementing either TP (if an X-probe) or FP (if an autosome). Then for each log ratio, plot FP/(total number of autosomes) vs. TP/(total number of X probes). The first metric, AreaUnderROC, is calculated using the trapezoidal rule to estimate the area under this curve. Please see the discussion below for more information regarding ROC curves

MedianDiff

This metric is the difference between the medians of the histogram of X-probes and autosomes. In the figure below the difference between the median of X-probes and autosomes is 0.67.

ErrorFraction

(FP/(total number of autosomes) + (1 – TP/(total number of X probe))/2. The metric reported is the minimum Error Fraction (minimum value of all the error fractions calculated).

Discussion of ROC Curves

In many research areas, it is necessary to segregate data into two groups that overlap in a measurement variable. In such cases, there are often many possible divisions or cut points that determine whether an item goes into one group or the other. The Receiver Operating Characteristic (ROC) curve analysis is a common method used to identify the optimal cut point and to determine the accuracy of the segregation. An ROC curve is a graphical representation of the trade off between the false positive and true positive rates for every possible cut off. By tradition, the plot shows the false positive rate on the X axis and true positive rate on the Y axis.

The following curve displays data on the log ratios of the probes on a microarray derived from biological samples taken from a normal male and a normal female. The expectation is that the probes that identify regions on the autosomes should have a median log ratio of zero while probes specific to the X-chromosome should have a median log(2) ratio of –1 when the ratio is Cy5/Cy3 and the male DNA is labeled with Cy5. The distribution of actual values for the autosome and X-specific probes overlaps, requiring a cut-point log ratio where probes with log ratios less than it are predicted to be X-specific and probes above it are predicted to be autosomal. With any cut point, some probes will be misassigned to the wrong category.

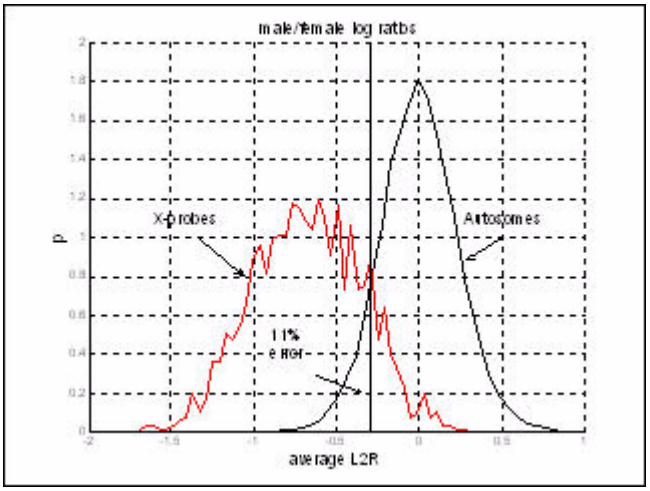


Figure 5-1 Graph of log ratio data

The second curve displays several possible ROC curves. Ideally, two distributions that are well separated will produce a curve similar to the green line. On the other hand, two distributions that overlap completely will produce a curve following the black diagonal (45 degree) line. The blue and pink lines depict more common curves similar to what would occur if the data from first graph were analyzed. The area under the ROC curve is a measure of how close to the ideal was achieved—perfect separation by the cut point. A score of 1 indicates perfect separation whereas an area of 0.5 (the 45 degree diagonal) indicates no separation effect.

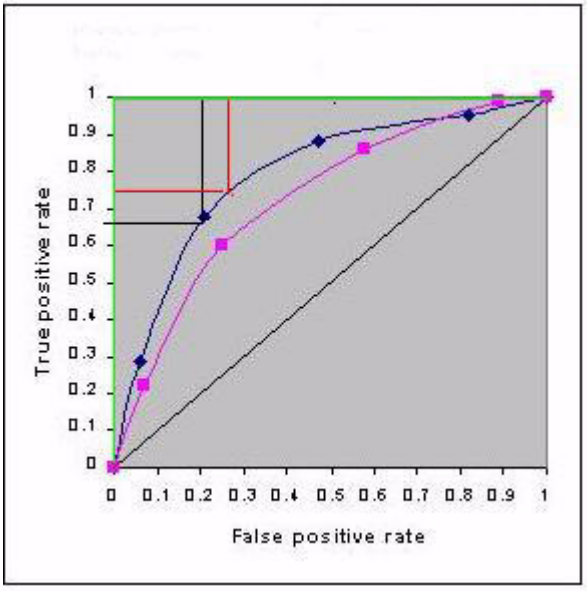


Figure 5-2 Graph showing possible ROC curves

Performance Tips

If you change the Statistic Type or Moving Average, select **All Arrays** again to save time in the future.

Your system may slow when it is trying to plot too many data points at once. You can minimize these delays by right-clicking the Genome, Chromosome, or Gene views and selecting Preferences. In the Preferences dialog box, turn off items in the Genome view that are graphics intensive, such as Scatter Plot and Moving Average.

Plug-Ins

To create a plug-in, you must write a standalone executable that can read standard input (STDIN), process the data, and write the results as standard output (STDOUT). The input format consists of a tab-delimited stream of text lines. The first line is a header that identifies the columns of incoming data. The format is always:

<ID>\t<CHROMOSOME>\t<START POS>\t<STOP POS>\t<ARRAY1>\t<ARRAY2>\t...\t<ARRAYN>

Since the header line can contain any text, do not make assumptions about strings in the header — the header will be used for labeling subsequent graphs. All microarray data will be in the format of $\log_2$ ratios. If the underlying raw data is not in that format, it will be converted to $\log_2$ ratio before sending it to the plug-in. In this way you can expect a consistent style of input data. The following is a very short sample of typical data for illustration:

id	chr	start	stop	log ratio
Hs.279061	17	647299	671372	0.907519155
Hs.178306	17	651554	653309	0.920187651
Hs.437447	17	687335	714056	0.952637998
Hs.434004	17	891508	1067881	0.972654947

The plug-in should read this data and then compute additional columns for each row based on the $\log_{10}$ ratio data. You can append as many columns of additional data as you desire.

The final result must be printed as standard output (STDOUT). You can filter out rows or add new ones. Potentially, you can even filter out specific columns that you do not wish to plot. However, you must maintain a consistent number of columns in all rows. At a minimum, the first four columns should be retained, since they are used to plot the remaining columns of computed data.

Plug-ins should be placed in the directory named plugins that is created in the CGH Analytics application directory. On Win32, this is usually **C:\program files\CGHAnalytics\plugins**. On the Macintosh, the directory is **/Applications/CGHAnalytics/plugins**.

When CGH Analytics is started, the plug-ins directory is queried for available plug-ins that are then displayed in the new Tools menu. If you add a plug-in, you must restart CGH Analytics before the plug-in will be recognized.

In this early version of the plug-in architecture, only data from the currently selected microarrays for the currently selected chromosome are sent to the plug-in. Later versions of CGH Analytics may have more elaborate selection mechanisms for processing plug-ins.

A couple of examples are provided in the plug-ins directory.

Echo Example.c

This is a very small C program that echoes STDIN back to STDOUT. Even though represents the simplest possible example of a plug-in, it can be quite useful. Since the result of this will be to plot all the selected microarrays, you can now create a stacked plot of selected microarrays for comparison.

MovAvg Example.pl

This is a short Perl program that computes a 10-point moving average of each column of microarray data. As the result, you will get a stacked plot of all the input data and all the moving averages. This program is a good example of how computed columns are processed. You must have Perl is installed on your computer to use this plug-in.

Note that the current plotting is very simple and options are not provided for altering formats, but the simple plug-in architecture will allow you to write your own computational methods and analyze data selected from browsing the CGH Analytics display. Some minimal formatting control is allowed by adding text to the column headers. You can force a line graph vs. scatter plot by adding the string -plotline to a column header. If the header name ends with -plotline , a line plot will be made in place of the default scatter plot. This is useful for moving averages and similar computations. You can also keep a column from being plotted by adding -noplot to the header text.

Note that this extra text is stripped from the header before it is displayed on the plot. This extra text will not show up in figures, and it is only used to control the plot's format. MovAvg Example.pl shows how column-naming can be used. As you read the first line (which contains the header text), you can add text to the existing headers or add text to the headers for your generated columns, as well, to give you a small amount of formatting control.

CGHSmooth Plug-in

The CGHSmooth plug-in provides several methods for performing moving-average-like smoothing functions on the data. This plug-in was written by Bo Curry based on work by Bo and Nick Sampas at Agilent Laboratories. There are currently five options for different weighting functions applied to the moving average. The general idea is to weight measurements close to the center position more than measurements distant from the center. Each is specified by number as follows:

0=None: Performs no smoothing, and basically returns the original data. In some cases, points that have exactly the same position will be averaged. In effect this is a window size of 0.

1=Rectangular: Performs a standard moving average. All points within the rectangle (the window) receive the same weight.

2=Gaussian: Applies a Gaussian weighting function.

3=Triangular: Applies a triangular weighting function.

4**=Lorentzian:** Applies a Lorentzian weighting function.

5=Biexponential: Applies a biexponential weighting function.

Sample Data

Two sample data files are provided with this release. Both of these files are provide only as examples of data sets and should not be used for determining any scientific or biological findings.

K562

This data set contains results from a series of Agilent microarrays used to analyze the K562 cell line DNA in one channel and reference male DNA in the other channel, as well as Male vs Female control samples. The data is replicated, dye swapped, and represents both amplified and direct-labeled samples.

CGH_EXP

This data set is from: Pollack, J. R., et al., 1999. Genome-wide analysis of DNA copy-number changes using cDNA microarrays. Nature Genetics 23:41–46.

This data set is compiled from the data supplied to the public with genomic annotations from the HG16 build of the human genome. Not all array features were mapped in this manner, so the data represents a subset.

NOTE: This data appears to have been in $\log_2$ format originally (but this is unconfirmed).

Spike In Reference DNA

The spike-in reference provides an external DNA reference standard that monitors the overall system through the standard CGH workflow for a given set of experiments. The reference DNA spike-in is supplied by Agilent.

The Agilent CGH Spike-In Kit furnishes positive controls for monitoring the microarray workflow process from sample digestion and labeling through hybridization. The kit contains two spike-in mixtures, A and B. The two mixtures contain a common set of 12 different synthetic double-stranded DNA molecules, SM_01 to SM_12, that are designed to hybridize only with specific spike measurement array probes without cross-hybridization to biological or other control probes on the microarray.

The concentration of each of the different synthetic DNA molecules varies within a spike-in mixture and between each of the two mixtures. The A and B spike-in mixtures are added to the experimental and control samples, respectively, prior to the restriction digestion and are carried through the experimental workflow, where each sample is labeled with a different dye and co-hybridized on a microarray. Spike mix A is always added to the experimental sample and spike mix B is always added to the reference sample. The final relative amount of each of the 12 spike-in DNAs in each of the two dye labels samples is shown in Table 1 on page 232.

Table 1 Probes and spike in targets and their ratios

Spike Name	Reference A	Reference B	A:B Ratio	Log2 of A:B
SM_01	0	2		n/a
SM_02	0.5	2	0.5:2	-2
SM_03	1	2	1:2	-1
SM_04	2	1	2:1	-1
SM_05	1.5	2	1.5:2	-0.4149
SM_06	2	2	2:2	0.0000
SM_07	6	6	6:6	0.0000
SM_08	3	2	3:2	0.584991538
SM_09	4	2	4:2	0.3010
SM_10	6	2	6:2	0.4771
SM_11	8	2	8:2	0.6021
SM_12	32	2	32:2	1.2041

Spike cocktail "A" will always be added to the **experimental** sample.
Spike cocktail "B" will always be added to the **control** sample.
A/B ratio = the ratio of the mass of a given spike target between each of the 2
"+" orientation = cy5/cy3

Using the Agilent CGH spike-in reference results in spike measurement probes on the array displaying specific intensity ratios between the two dye channels for a given spike DNA molecule. For the CGH Analytics process, the spike-in hybridizations results are captured in the QC report as a plot of the expected versus observed $\log_2$ ratios for each spike-in reference. You can use the data in this plot to monitor the system for linearity, sensitivity, and accuracy.

User Interface Elements

Several features of the Main display are designed to make your navigation easier.

Expansion Arrows

Between the main window panes, you will find opposing "arrowheads" that facilitate expanding or contracting the size of the panes they border. You can use them in two ways:

- Click on an arrowhead to expand the adjoining window pane in the direction of the arrow. For example, clicking on the left-pointing arrowhead between the Genome view and the Navigator area, will expand the three right most panes to fill the screen. Afterwards, clicking on the right-pointing arrowhead will return the panes to the original view.
- Alternatively, click and drag an arrowhead to expand the adjoining window only partially if you want a better view of the data in the window.

Similarly, you can manipulate the up and down arrowheads to expand or contract the Tab view area upwards or downwards.

Toolbar Isolation

A wide, multi-shaded blue bar delimits the right and left ends of the toolbar. You can isolate the toolbar for better viewing by clicking on the bar at either end and drag the toolbar away from its location at the top of the Main window. To return the toolbar to its original location, press **Alt + F4** or right-click on the blue bar across the top of the isolated toolbar and click **Close**.

Toggle Buttons

A toggle button is centered at the top of each pane of the Main display. Clicking a toggle button will expand and isolate that window pane in a separate, titled window.

Web Searching

You can do a simple Web search directly from within CGH Analytics. Right-click a cell in the detailed table to display a drop-down menu with several search options.

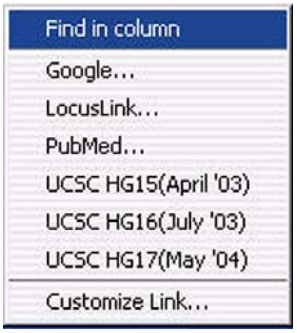


Figure 5-3 Web search drop-down menu

If you select **Find in column...**, a new Find in column dialog box appears.

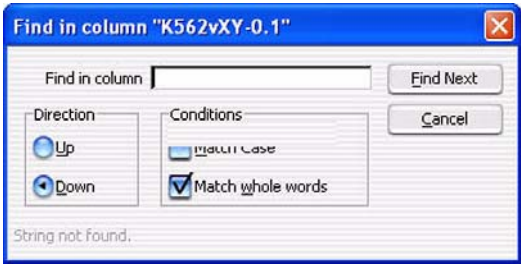


Figure 5-4 Web Search Find in Column dialog box

You can type in a search word and decide to match the whole word or the case, and the direction of the search. The software searches the column you selected for the first match that meets your criteria. If you select any other option, a Web query is generated using the cell that you clicked (in this case, K562vXY-0.1). A browser window displays the results, if any. If you click on a cell whose value is not valid for the site you are searching, the results may not be useful.

If you select **Customize Link...**, the Customize Search link dialog box appears. There you can specify the URL of another search site on the Web. You can add this link to the Web Search menu.

Figure 5-5 Customize Search Link dialog box

References

1. Chaya Ben-Zaken Zilberstein et al., *Lecture Notes in Computer Science: Regulatory Genomics: RECOMB 2004 International Workshop, Revised Selected Papers* in Eskin E. and Workman, C. (eds.), Springer Berlin / Heidelberg 2005(3318). ISBN: 3-540-24456-5

2. Hedenfalk,I. et. al. “Molecular classification of familial non-BRCA1/BRCA2 breast cancer” Proc. Natl. Acad. Sci. USA, 2003 Mar 4;100(5):2532-7.

3. Kincaid, R., A. Ben-Dor, and Z. Yakhini. "Exploratory visualization of array-based comparative genomic hybridization." Information Visualization, 2005, 4:176-190. doi:10.1057/palgrave.ivs.9500095.

4. Lipson, Doron, Yonatan Aumann, Amir Ben-Dor, Nathan Linial, and Zohar Yakhini. "Efficient Calculation of Interval Scores for DNA Copy Number Data Analysis." Proceedings of Recomb, 2005.

5. Olshen, A. B., Venkatraman, E. S., Lucito, R., and Wigler, M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, **5**, 4, 557-572.

6. JV Braun, HG Muller (1998), “Statistical methods for DNA sequence segmentation", Statistical Science, 13(2):142-162.

7. E. S. Venkatraman and Adam B. Olshen. (2005). DNAcopy: A Package for Analyzing DNA Copy Data. *The Bioconductor Project*: http://www.bioconductor.org.

www.agilent.com

In This Book

The User Guide presents detailed information identifying the components of CGH Analytics 3.4 software and their use.

First Edition, July 2006

Agilent Technologies