affymetrix

User Manual

Affymetrix® Genotyping Console 4.2 User Manual

P/N 702982 Rev. 3

For research use only.
Not for use in diagnostic procedures.

Trademarks

Affymetrix®,  , GeneChip®, NetAffx®, Command Console®, Powered by Affymetrix™, GeneChip-compatible™, Genotyping Console™, DMET™, GeneTitan™, Axiom™, and GeneAtlas™ are trademarks or registered trademarks of Affymetrix, Inc.

All other trademarks are the property of their respective owners.

Limited License Notice

Limited License. Subject to the Affymetrix terms and conditions that govern your use of Affymetrix products, Affymetrix grants you a non-exclusive, non-transferable, q license to use this Affymetrix product only in accordance with the manual and written instructions provided by Affymetrix. You understand and agree that except as expressly set forth in the Affymetrix terms and conditions, that no right or license to any patent or other intellectual property owned or licensable by Affymetrix is conveyed or implied by this Affymetrix product. In particular, no right or license is conveyed or implied to use this Affymetrix product in combination with a product not provided, licensed or specifically recommended by Affymetrix for such use.

Patents

Scanner products may be covered by one or more of the following patents: U.S. Patent Nos. 5,578,832; 5,631,734; 5,834,758; 5,936,324; 5,981,956; 6,025,601; 6,141,096; 6,171,793; 6,185,030; 6,201,639; 6,207,960; 6,218,803; 6,225,625; 6,252,236; 6,335,824; 6,403,320; 6,407,858; 6,472,671; 6,490,533; 6,650,411; 6,643,015; 6,813,567; and other U.S. or foreign patents.

Software products may be covered by one or more of the following patents: U.S. Patent Nos. 5,733,729; 5,795,716; 5,974,164; 6,066,454; 6,090,555; 6,185,561; 6,188,783; 6,223,127; 6,228,593; 6,229,911; 6,242,180; 6,308,170; 6,361,937; 6,420,108; 6,484,183; 6,505,125; 6510,391; 6,532,462; 6,546,340; 6,687,692; 6,607,887; 7,062,092 and other U.S. or foreign patents.

Fluidics stations Products may be covered by U.S. Patent No. 6,114,122; 6,287,850; 6,391,623; 6,422,249; and other U.S. or foreign patents.

AutoLoader products may be covered by one or more of the following patents: U.S. Patent Nos. 6,511,277; 6,604,902; 6,705,754; 7,108,472; and other U.S. or foreign patents.

Copyright

Table of Contents

Introduction

The Affymetrix® Genotyping Console™ software (GTC) provides an easy way to create genotype calls for collections of CEL files. Genotyping Console generates Copy Number, Loss of Heterozygosity (LOH), Copy Number Segments data, and copy number variation data, depending on the array type, as listed in the table below (Table 1.1).

Table 1.1 Genotyping Console analyses for different array types

Affymetrix® Array Type	Genotype Calls	Copy Number/LOH Data	Copy Number Segments Data	Copy Number Variation Analysis
Human Mapping 100K Arrays: Mapping50K_Xba240	Yes	Yes	Yes	No
Mapping50K_Hind240	Yes	Yes	Yes	No
Human Mapping 500K Arrays: Mapping 250K_Nsp	Yes	Yes	Yes	No
Mapping 250K_Sty	Yes	Yes	Yes	No
Genome-Wide Human SNP Array 5.0	Yes	No	No	No
Rat and Mouse Arrays	Yes	No	No	No
Genome-Wide Human SNP Array 6.0	Yes	Yes	Yes	Yes
Axiom Genotyping Array plates, including:	Yes	No	No	No

NOTE: The Axiom™ Genome-Wide CEU 1 Array is the same as the Axiom Genome-Wide Human Array.

Genotyping Console displays metrics and annotation information in standard tabular form so you can evaluate the data quality for a given array. Scatter plots, line graphs and the heat map viewer give you the power to quickly identify features of interest in your data set. Numerous data and visualization export features make it easy to share results with other applications and users.

The GTC Browser enables you to survey your Copy Number and Loss of Heterozygosity data.

Genotyping Console is not a secondary analysis package. However, it does create CHP files and tab-delimited text files required for secondary analysis packages available from companies in the Affymetrix GeneChip® Compatible Program.

About this Update

GTC 4.2 includes the following enhancements and new features:

- Edit Calls within the Cluster Graph. (See page 160)
- Add or remove SNPs from SNP lists from within the cluster graph. (See page 166).
- Import SNP lists from APT's SNPolisher application.
- Supports Windows 7 (64 bit) and Windows 8.1 (64-bit). (See Table 2.1)
- Saturation_GC and Saturation_AT metrics are now displayed after an Intensity QC is performed.
- Added Axiom CNV Tool and Viewer applications to the Tools menu. (See Appendix K).

Conventions Used in This Guide

This guide provides a detailed outline for all tasks associated with Affymetrix® GeneChip Command Console. Various conventions are used throughout the guide to help illustrate the procedures described. Explanations of these conventions are provided below.

Steps

Instructions for procedures are written in a step format. Immediately following the step number is the action to be performed. Following the response additional information pertaining to the step may be found and is presented in paragraph format. For example:

1. Click **Yes** to continue.

   The Delete task proceeds.
   In the lower right pane the status is displayed.

   To view more information pertaining to the delete task, right-click **Delete** and select **View Task Log** from the shortcut menu that appears.

Font Styles

Bold fonts indicate names of commands, buttons, options or titles within a dialog box. When asked to enter specific information, such input appears in italics within the procedure being outlined.
For example:

1. Click the **Find** button or select **Edit** > **Find** from the menu bar.

   The Find dialog box appears.
2. Enter *AFFX-BioB-5_at* in the **Find what** box, then click **Find Next** to view the first search result.
3. Continue to click **Find Next** to view each successive search result.

Screen Captures

The steps outlining procedures are frequently supplemented with screen captures to further illustrate the instructions given. The screen captures depicted in this guide may not exactly match the windows displayed on your screen.

Additional Comments

TIP: Information presented in Tips provide helpful advice or shortcuts for completing a task.

NOTE: The Note format presents important information pertaining to the text or procedure being outlined.

IMPORTANT: The Important format presents important information that may affect the accuracy of your results.

CAUTION: Caution notes advise you that the consequence(s) of an action may be irreversible and/or result in lost data.

WARNING: Warnings alert you to situations where physical harm to person or damage to hardware is possible.

Technical Support

Affymetrix provides technical support to all licensed users via phone or E-mail. To contact Affymetrix® Technical Support:

Affymetrix, Inc.
3420 Central Expressway
Santa Clara, CA 95051 USA
E-mail: support@affymetrix.com
Tel: 1-888-362-2447 (1-888-DNA-CHIP)
Fax: 1-408-731-5441

Affymetrix UK Ltd.
Voyager, Mercury Park
Wycombe Lane, Wooburn Green
High Wycombe HP10 0HH
United Kingdom
UK and Others Tel: +44 (0) 1628 552550
France Tel: 0800919505
Germany Tel: 01803001334
E-mail: supporteurope@affymetrix.com
Tel: +44 (0) 1628 552550
Fax: +44 (0) 1628 552585

Affymetrix Japan, K. K.
ORIX Hamamatsucho Bldg, 7F
1-24-8 Hamamatsucho, Minato-ku
Tokyo 105-0013 Japan
Tel: +81-3-6430-4020
Fax: +81-3-6430-4021
salesjapan@affymetrix.com
supportjapan@affymetrix.com

Working with Genotyping Console

Genotyping Console is a stand-alone application. It can be installed on computers that have GeneChip® Operating System (GCOS) software, Affymetrix GeneChip® Command Console™ (AGCC) software, or either.

NOTE: If you are using GCOS files, Affymetrix recommends that you transfer data out of GCOS using the Data Transfer Tool (available at Affymetrix.com) and use the Flat File option in order to retain sample attributes.

The tables below (Table 2.1) show the operating systems that Genotyping Console has been verified on and the recommended minimum requirements. The larger data file size associated with Genome-Wide Human SNP 5.0 and 6.0 Arrays should be taken into account when calculating the necessary available disk space requirement.

Table 2.1 Verified 64 -bit operating systems & recommended requirements for GTC Software

64-bit Operating System	Speed	Memory (RAM)	Available Disk Space*	Web Browser
Windows 7 (with Service Pack 1) 64 bit	2.83 GHz Quad Core Processor	16 GB RAM	150 GB	Internet Explorer 8.0 and above
Windows 8.1 Professional 64 bit	2.83 GHz Quad Core Processor	16 GB RAM	150 GB	Internet Explorer 8.0 and above

The following sections in this chapter describe:

To use Genotyping Console, you must:

1. Install the GTC software *(see Installation Instructions* on page 11.
2. Create a user profile (see Chapter 3, *User Profiles* on page 31).
3. Download or copy the necessary library and annotation files (see Chapter 4, *Library & Annotation Files* on page 34).
4. Set up a workspace and data set(s) (see Chapter 5, *Workspaces & Data Sets* on page 45).

Installation Instructions

1. Download the 64-bit installer software from Affymetrix.com: h**ttp://www.affymetrix.com**
2. Unzip the downloaded software package. It includes the installation program and release notes.
3. Review the release notes and installation instructions before proceeding with the installation.
4. Double-click **GenotypingConsoleSetup.exe** to install the software.
5. Follow the directions provided by the installer.

NOTE: The setup process installs the required Microsoft components, which includes the .NET 3.5 framework and Java components and Visual C++ runtime libraries.

Updates & General Information

New information about Genotyping Console will be made available to customers through the Update Button on the main tool bar in Genotyping Console. There are 3 different options: Updates Available, No New Updates, or Updates (Offline).

When updated information is available, click on the green Updates Available button on the main tool bar and a web browser will be launched indicating what new information is available.

When there are no new updates available, the following button will be displayed on the main tool bar. Clicking on the button will launch a web browser showing the current informational messages.

If the computer is offline, Genotyping Console will be unable to determine if there are any updates available and the Updates button will indicate the offline status.

Notes for Users of Earlier Versions of Genotyping Console

NOTE: GTC 4.2 workspaces cannot be opened in GTC versions before 4.1. Workspaces created in earlier versions of GTC can be opened in GTC 4.2, but no longer can be used in earlier versions of GTC.

NOTE: Custom analysis configurations for earlier versions of GTC are updated to work with GTC 4.2. Once updated, they no longer work with older versions of GTC.

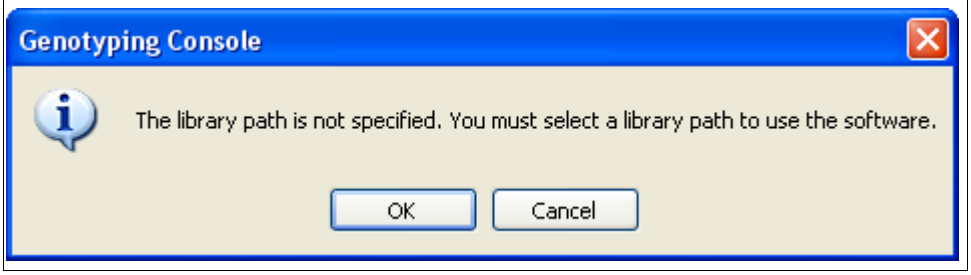
Starting Genotyping Console

1. Double-click the Genotyping Console shortcut  on the desktop. Alternately, from the Windows Start Menu, select **Programs > Affymetrix > Genotyping Console**.

   The Genotyping Console opens with the User Profile window displayed.
2. Select or create a User Profile (see *Creating and Selecting a User Profile* on page 31.

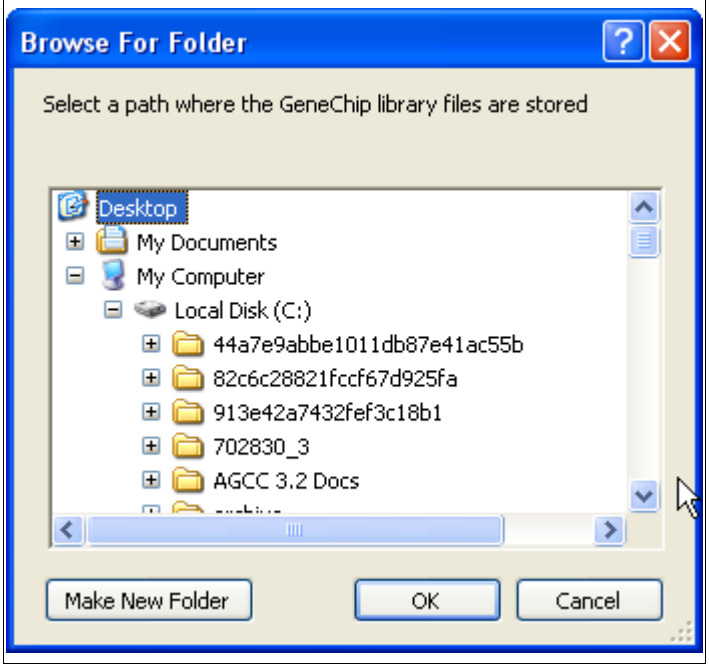
   After creating a User Profile, the library path notice appears (Figure 2.1).

Figure 2.1 Prompt to set the library path



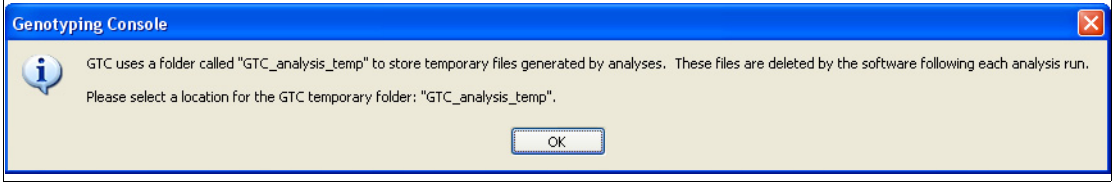
3. Click **OK.**
   The Browse for Folder dialog box opens (Figure 2.2).

Figure 2.2 Select or create a library folder



4. Select or create a location for the GTC 4.2 library folder and click **OK**.
   The Temporary file folder location notice appears (Figure 2.3).

Figure 2.3 Temp Folder location notice

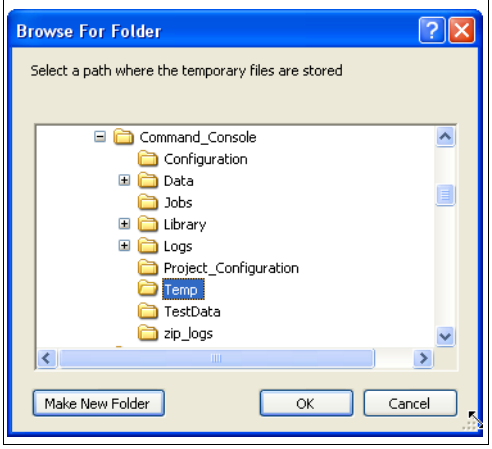


The Affymetrix Power Tools software uses the temporary files folder during data analysis. The temporary files folder must reside on a local hard drive, not a network drive. Users must have write access to the temporary files folder.

See Appendix J, *Hard Disk Requirements* on page 363 for information on local hard drive space requirements.

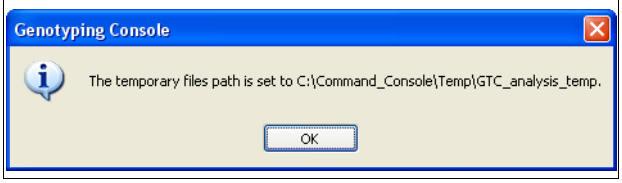
5. Click **OK**.
   The Browse for Temp Folder dialog box opens (Figure 2.4).

Figure 2.4 Browse for Temp Folder dialog box



6. Select or create a location for the GTC 4.2 temp folder and click **OK**.
   The Temp Files path notice appears (Figure 2.5).

Figure 2.5 Temp Folder location notice



Click **OK** and proceed with creating a workgroup (Chapter 5, *Workspaces & Data Sets* on page 45) or other tasks.

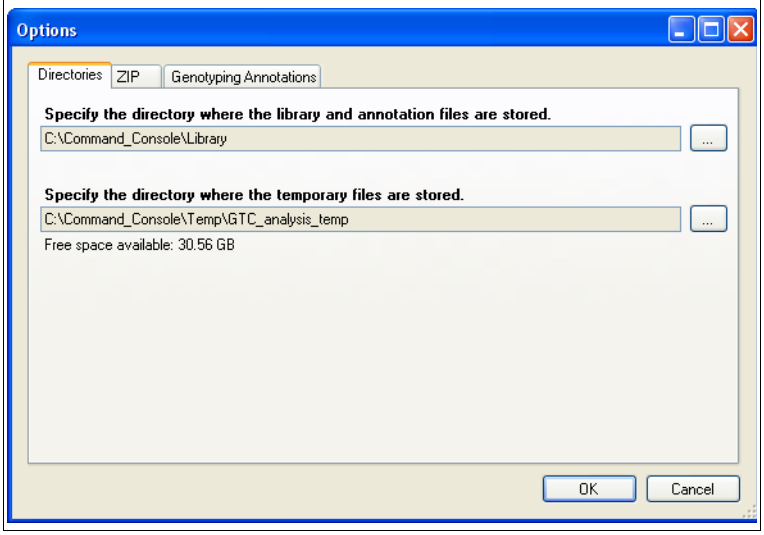
Changing Folder Locations

You can change the location of the library and temp folder in the Options dialog box.

To change the location of a folder:

1. From the Edit menu, select **Options**; or
   Click the **Options** button in the tool bar.
   The Options dialog box opens (Figure 2.6).

Figure 2.6 Options dialog box



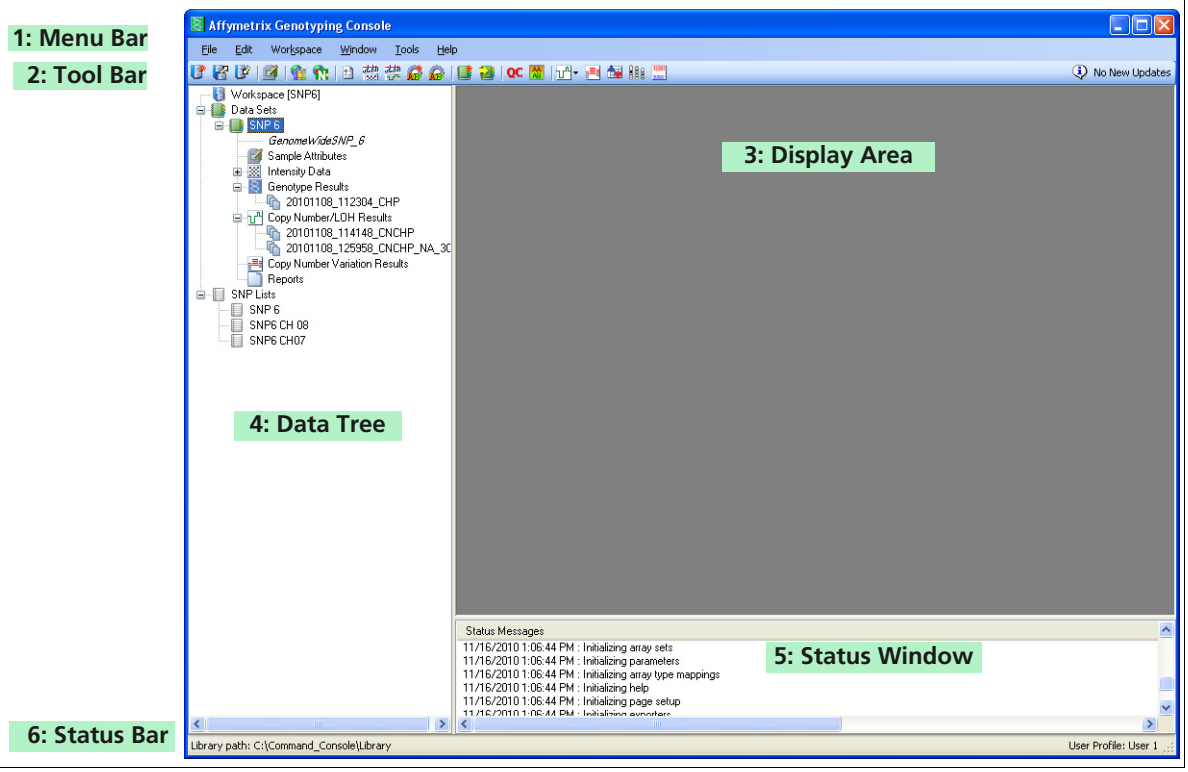
2. Specify a new location for library and temp files by either:
   - Entering a path in the appropriate box
   - Clicking the **Browse** button and browsing to the new location.
3. Click **OK**.

Parts of the Console

After creating or selecting the user profile the Genotyping Console Opens (Figure 2.7).

The components of the GTC interface are introduced below.

Figure 2.7 Genotyping Console with workspace selected



NOTE: See the *Affymetrix GTC Browser 1.2 User Manual* for information about viewing the Copy Number, Loss of Heterozygosity, and Copy Number Segment data in graphical format.

1 and 2: Menu Bar and Tool Bar

The menu bar and tool bar provide quick access to the GTC functions.

3: Display area

Some of the data generated by GTC can be viewed in tables and graphs in the display area, including:

- Intensity file QC data and graphs
- Genotyping Data tables
- SNP Cluster Graph
- Copy Number/Loss of Heterozygosity QC data
- Heat Map for Copy Number and Copy Number Variation data

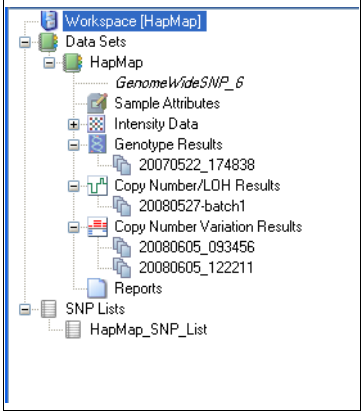
NOTE: The Copy Number, Loss of Heterozygosity, and Copy Number Segment data generated by GTC is displayed in the GTC Browser. See the *Affymetrix GTC Browser 1.2 User Manual* for more information.

4: Data Tree

Genotyping Console displays workspace information in the form of a data tree. The items within the Data Sets section of the data tree are ordered by the typical user workflow (Figure 2.8).

Data sets start as collapsed nodes in the data tree. Double-click a data set to expand the node and show the tree items. By double-clicking on the data tree items, the first item in the right-click menu will automatically open. For example, if you double-click the All Intensity group, the Intensity QC Table will open, showing the QC information for all intensity data files in the data group.

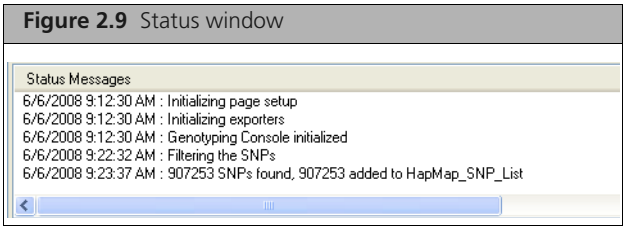
Figure 2.8 GTC data tree showing workspace, data set, and SNP lists



5: Status Window

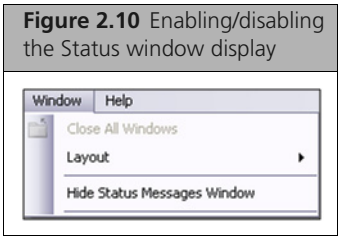
The Status window displays all status and algorithm progress information (Figure 2.9).

Figure 2.9 Status window



To disable this view, go to the Window menu and select **Hide Status Messages Window**.

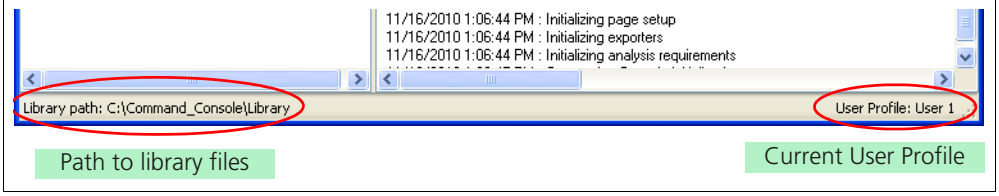
Figure 2.10 Enabling/disabling the Status window display



6: Status Bar

The Status bar at the bottom of the GTC window (Figure 2.11) displays information on the path to library files and the user profile.

Figure 2.11 Status bar shows the library path and current user profile



File Types & Data Organization in GTC

To fully use the capabilities of GTC, you need to understand the file types and data organization used in this software. GTC uses:

- *Data and QC Files* on page 17)

NOTE: QC files (.gqc) are no longer available for AGCC CEL files QCed in GTC 4.0 and higher. The QC information is stored in the CEL file.

- *Support files* on page 18
- *Data Organization in Genotyping Console* on page 18

Data and QC Files

The data and QC files used by GTC are listed below, along with the file extensions used to identify them.

Some data files are generated by other Affymetrix software and used by GTC:

- Sample files (.arr and .xml)
- Intensity data files (.cel)

GTC generates other data files during the analysis of the intensity data files:

- Genotype Data files (.chp)
- Copy Number Data files (.cnchp)
- LOH Data files (.lohchp)
- Copy Number/LOH Data files (.cnchp) for SNP 6.0 analysis
- Copy Number Segment Data (.cn_segments)
- Copy Number Segment Summary (.cn_segments_summary)
- Custom Regions Report (.custom_regions)
- Custom Regions Summary Report (.custom_regions_summary)
- Copy Number Variation Data files (.cnvchp) for SNP 6.0 analysis

GTC generates QC information to help you evaluate your data:

- Intensity QC information for assessing suitability for batch genotyping and/or Copy Number/LOH analysis
- QC data for Copy Number/LOH analysis

NOTE: QC files (.gqc) are no longer available for AGCC CEL files QCed in GTC 4.0 and higher. The QC information is stored in the CEL file.

- Report files for viewing data and record keeping

You access the data in these files through the GTC data tree.

Support files

The support files are necessary to use all of the features of GTC.

- Library file sets, with files for genotyping, copy number/LOH/CN Segment and copy number variation analysis.
- Reference Model files for SNP 6.0 single sample Copy Number/LOH analysis
- Prior and Posterior model files for:
  - BRLMM-P
  - Birdseed V1 and V2
  - Axiom GT1
- Annotation files for the Arrays
- Browser Annotation files
- Optional files, including:
  - SNP lists (both provided by Affymetrix and generated by user)
  - Hints files
  - Inbred sample files
  - Gender files

Data Organization in Genotyping Console

The data used in GTC is organized by:

- *Workspaces* on page 18
- *Data Sets* on page 19
- *SNP Lists* on page 20

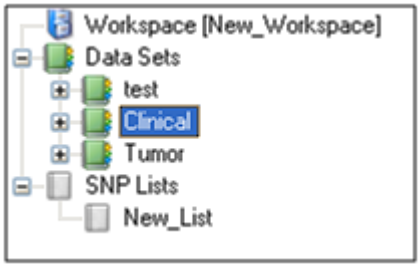
Workspaces

A workspace is a collection of data sets and SNP lists (Figure 2.12).

Only one workspace can be displayed in an open instance of GTC.

NOTE: Once you open a workspace in GTC 4.2, you will no longer be able to use it in earlier versions of GTC.

Figure 2.12 Workspace with data sets and SNP list



A workspace should contain only related data (for example, belonging to one primary investigator or one research study).

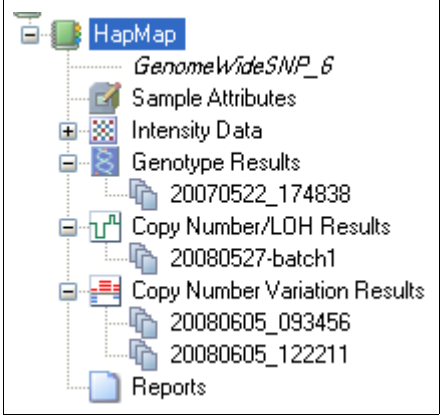
NOTE: Only one user can have the same workspace open at one time. If other users need access to the same data files, they can either make a personal copy of a workspace file that is not in use, or create a new Workspace and add the same data files to the new workspace. Simultaneous genotyping of the same set of CEL files within two workspaces is not recommended.

The workspace file stores the locations of the data files, not a copy of the data files themselves. See Chapter 5, *Workspaces & Data Sets* on page 45 for more information about workspaces.

Data Sets

Each workspace can have multiple *data sets* (Figure 2.13). A data set manages a group of ARR/XML, CEL, CHP, CNCHP (and/or LOHCHP), cn_segments files, and CNVCHP files from a single type of array or array set (e.g. Human Mapping 100K or 500K Arrays, Genome-Wide SNP Array 5.0, Genome-Wide SNP Array 6.0, or Axiom™ Genome-Wide Human Arrays).

Figure 2.13 Data set



A data set manages:

- Sample attributes: ARR or XML files
- Intensity data in CEL files

  During QC the files are grouped into the following categories:
  - All: all CEL files in the data set
  - In Bounds: CEL files that passed intensity QC criteria
  - Out of Bounds: CEL files that failed intensity QC criteria

NOTE: GQC files are not available for AGCC CEL files QCed in GTC 4.0 and higher. The QC information is stored in the CEL file.

You can also assemble custom lists of intensity data. For more information, see:

- *Creating Custom Intensity Data Groups using Intensity QC Data* on page 84
- *Creating a Custom Intensity Group from the CHP File Data* on page 117
- *Creating Custom Intensity Data Groups Using the SNP Cluster Graph* on page 162

- Genotype Results: CHP files. These are grouped into:
  - Batch genotype results, either from direct analysis or import
  - Custom CHP groups assembled by you
- Copy Number/LOH Results: Analysis files for:

  - Copy Number
  - LOH
  - Copy Number Segments and Copy Number Custom Regions
- Copy Number Variation Results in CNVCHP files. These are grouped into:
  - Batch Copy Number Variation results, either from direct analysis or import
- Reports
  - Concordance reports

Within a data set, the following information can be displayed in tables and graphs for viewing and exporting:

- Sample attribute information
- QC metrics
- Signature SNP genotypes
- CHP and SNP summary data
- SNP cluster graphs
- Copy Number/LOH QC information, copy number segment and custom region data (available for Human Mapping 100K/500K Arrays and Genome-Wide Human SNP Array 6.0)
- Copy Number Variation results data

NOTE: Copy Number/LOH data can be displayed in the GTC Browser. See the *Affymetrix GTC Browser 1.2 User Manual* for more information.

NOTE: Copy Number Variation data for SNP 6.0 is also displayed in the Heat Map Viewer together with copy number data. In order to view Copy Number Variation data in the Heat Map Viewer, you must have copy number data that originates from same CEL files. See Chapter 16, *Heat Map Viewer* on page 323 for more information.

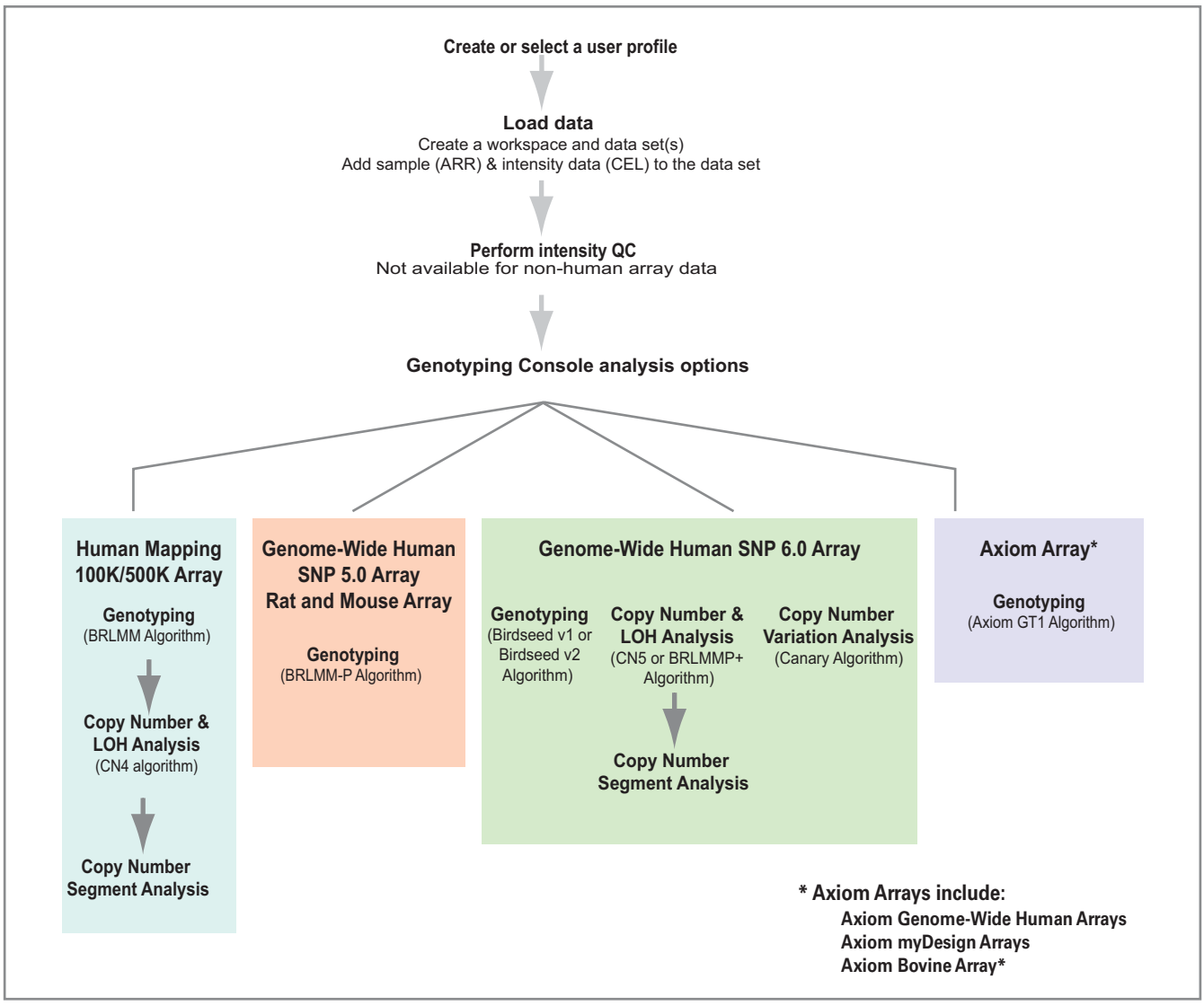
SNP Lists

SNP lists allow you to manage markers of interest. You can generate SNP lists from your genotyping data or import SNP lists from other sources.

Basic Workflows in Genotyping Console

The figure below (Figure 2.14) shows an overview of the GTC workflows.

Figure 2.14 Overview of GTC workflows for different array types



The following sections describe the workflows for the different array types. There are many similarities between the workflows for different array types, but some significant differences, too.

- *GTC Workflow for Axiom Arrays* on page 22
- *GTC Workflow for SNP 6 Arrays* on page 23
- *GTC Workflow for Genome-Wide SNP Array 5* on page 24
- *GTC Workflow for Human Mapping 100K/500K Arrays* on page 25

GTC Workflow for Axiom Arrays

GTC can perform Genotyping analysis using the Axiom GT1 algorithm on the following types of arrays:

- Axiom Genome-Wide Human Arrays and Array Sets
- Axiom myDesign Arrays
- Axiom non-human Arrays

The workflow requires the following sets of steps:

1. Create Workspace and Data Sets
   - **A.** Create a workspace and data set for the data (see *Creating a New Workspace* on page 46
   - **B.** import intensity data (and Sample/Array Data, if available) into the data set (see *Adding Data to a Data Set* on page 49).

     NOTE: QC can also be automatically performed upon import of CEL files to the data set.

2. Perform Intensity QC and break into Reagent Versions.

   NOTE: Axiom Genome-Wide BOS 1 Arrays are not processed with different reagent versions and do not need to be separated into separate intensity data groups.

   NOTE: Axiom CEL files that have been QCed previously in GTC 4.0 or earlier will need to be submitted for intensity QC in GTC 4.2 to provide reagent version information.

   - **A.** Perform intensity QC to determine basic data quality (see Chapter 6, *Intensity Quality Control for Genotyping Analysis* on page 74).

     The intensity quality control check automatically creates the following intensity data groups, based on the Dish QC thresholds:

     - All
     - In Bounds
     - Out of Bounds

     The resulting Dish QC values and other metrics are displayed in tables and graphs, and can be exported.

     Removing poor quality CEL files from the set can improve the quality of the genotypes of the remaining CEL files.
   - **B.** Create custom intensity data file groups for CEL files produced using different reagent sets:
     - Reagent Version 1
     - Reagent Version 2

     Use data files from the In Bounds group to create these custom data file groups to make sure they pass the QC criteria.

     See *Creating a Custom Intensity Group from the CHP File Data* on page 117.
3. Perform Genotyping on samples from different Reagent Versions.
   - **A.** Select an intensity data file (CEL) group with data from Reagent Version 1 or Reagent Version 2.
   - **B.** Perform genotyping analysis on the group of files, as described in Chapter 7, *Genotyping Analysis* on page 89
   - **C.** Review the initial genotyping analysis QC data in the CHP Summary Results table, using Call Rate and other metrics, as described in *CHP Summary Table* on page 110.
   - **D.** Create new Intensity Data file group for samples with good performance in initial genotyping analysis and perform a second genotyping analysis, as described in *Creating a Custom Intensity Group from the CHP File Data* on page 117.

E. View the SNP calls and other metrics in the SNP Summary Results table, as described in Chapter 8, *Review the Genotyping Results* on page 124.

NOTE: You need a SNP list to view genotype result data. See ***Create a SNP List*** on page 125.

F. Review the clustering performance for SNPs of interest in the SNP Cluster Graph.
See Chapter 9, *Using the SNP Cluster Graph* on page 148.

G. Export the genotype calls for downstream analysis.
See Chapter 10, *Exporting Genotype Results* on page 181.

GTC Workflow for SNP 6 Arrays

GTC can perform the following analyses on SNP 6.0 Arrays:

- Genotyping Analysis (using the Birdseed v1 or Birdseed v2 algorithm)
- Copy Number/LOH Analysis (using CN5 BRLMM-P+ algorithm)
- Copy Number Variance (using the Canary algorithm)

The workflow requires the following sets of steps:

1. Create Workspace and Data Sets
   A. Create a workspace and data set for the data (see *Creating a New Workspace* on page 46).
   B. Import intensity data (and Sample/Array Data, if available) into the data set (see *Adding Data to a Data Set* on page 49).
2. Perform Intensity QC
   A. Perform intensity QC to determine basic data quality (see Chapter 6, *Intensity Quality Control for Genotyping Analysis* on page 74).
   The intensity quality control check automatically creates the following intensity data groups, based on the Contrast QC thresholds:
   - All
   - In Bounds
   - Out of Bounds

   Additional custom groupings of CEL files can also be made.
   Removing poor quality CEL files from the data set can improve the quality of the genotypes of the remaining CEL files.
3. Perform Genotyping
   A. Select a group or set of intensity data files (CEL) in a data set.
   B. Perform genotyping analysis on the group of files, as described in Chapter 7, *Genotyping Analysis* on page 89.
   C. Review the initial genotyping analysis QC data in the CHP Summary Results table, using Call Rate and other metrics, as described in *CHP Summary Table* on page 110.
   D. Create new Intensity Data file group for samples with good performance in initial genotyping analysis and perform a second genotyping analysis, as described in *Creating a Custom Intensity Group from the CHP File Data* on page 117.
   E. View the SNP results in the SNP Summary Results table, as described in Chapter 8, *Review the Genotyping Results* on page 124

NOTE: You need a SNP list to view genotype result data. See ***Create a SNP List*** on page 125.

F. Review the clustering performance for SNPs of interest in the SNP Cluster Graph.
See Chapter 9, *Using the SNP Cluster Graph* on page 148.

G. Export the genotype calls for downstream analysis.
See Chapter 10, *Exporting Genotype Results* on page 181.

4. Perform Copy Number/LOH Analysis for SNP 6.0 Arrays

A. Perform Copy Number and/or LOH analysis in GTC to generate Copy Number/LOH data files. See Chapter 13, *Copy Number & LOH Analysis for Genome-Wide Human SNP 6.0 Arrays* on page 242.

B. Run the Segment Reporting Tool on the CNCHP files to generate:
- Segment Data files
- Segment Summary file
- Custom Region Data files
- Custom Region Summary file

See *Using the Segment Reporting Tool* on page 285.

C. Review the data in the GTC Browser.
See *Loading Data into the GTC Browser* on page 305

D. View the log2ratio values in the Heat Map Viewer.
See Chapter 16, *Heat Map Viewer* on page 323.

E. Export the data for further analysis.

5. Perform Copy Number Variation Analysis

NOTE: Copy Number Variation (CNV) analysis can be performed only on Genome-Wide Human SNP Array 6.0 data.

For CNV analysis, the Canary algorithm makes CN state calls (0, 1, 2, 3, 4) for regions with known copy number variants (CNV) or copy number polymorphisms (CNP). The region within known copy number variants can contain one or more CN/SNP probe sets.

A. Perform the Copy Number Variation analysis. See Chapter 15, *Copy Number Variation Analysis* on page 315.

B. View the results in the Heat Map viewer with copy number results. See Chapter 16, *Heat Map Viewer* on page 323.

GTC Workflow for Genome-Wide SNP Array 5

GTC can perform Genotyping analysis using the BRLMM-P algorithm on the following types of arrays:

- Genome-Wide SNP Array 5.0
- Rat Array
- Mouse Array

NOTE: Intensity QC and Signature SNPs are not available for Rat and Mouse Arrays.

The workflow requires the following sets of steps:

1. Create Workspace and Data Sets

A. Create a workspace and data set for the data (see *Creating a New Workspace* on page 46.

B. Import intensity data (and Sample/Array Data, if available) into the data set (see *Adding Data to a Data Set* on page 49).

2. Perform Intensity QC
   A. Perform intensity QC to determine basic data quality (see Chapter 6, *Intensity Quality Control for Genotyping Analysis* on page 74.

      The intensity quality control check automatically creates the following intensity data groups, based on the Contrast QC thresholds:

      - All
      - In Bounds
      - Out of Bounds

      Additional custom groupings of CEL files can also be made.

      Removing poor quality CEL files from the data set can improve the quality of the genotypes of the remaining CEL files.
3. Perform Genotyping and Review Data
   A. Select an intensity data files (CEL) group.
   B. Perform genotyping analysis on the group of files, as described in Chapter 7, *Genotyping Analysis* on page 89.
   C. Review the initial genotyping analysis QC data in the CHP Summary Results table, using Call Rate and other metrics, as described in *CHP Summary Table* on page 110.
   D. Create new Intensity Data file group for samples with good performance in initial genotyping analysis and perform a second genotyping analysis, as described in *Creating a Custom Intensity Group from the CHP File Data* on page 117.
   E. View the SNP results in the SNP Summary Results table, as described in Chapter 8, *Review the Genotyping Results* on page 124

      NOTE: You need a SNP list to view genotype result data. See *Create a SNP List* on page 125.

   F. Review the clustering performance for SNPs of interest in the SNP Cluster Graph.

      See Chapter 9, *Using the SNP Cluster Graph* on page 148.
   G. Export the genotype calls for downstream analysis.

      See Chapter 10, *Exporting Genotype Results* on page 181

GTC Workflow for Human Mapping 100K/500K Arrays

You can perform the following types of analyses on Human Mapping 100K/500K Array data:

- Genotyping
- Copy Number/Loss of Heterozygosity

1. Create Workspace and Data Sets
   A. Create a workspace and data set for the data (see *Creating a New Workspace* on page 46).
   B. Import intensity data (CEL) (and Sample/Array Data) into the data set (see *Adding Data to a Data Set* on page 49).
2. Perform Intensity QC
   A. Perform intensity QC to determine basic data quality (see Chapter 6, *Intensity Quality Control for Genotyping Analysis* on page 74).

      The intensity quality control check automatically creates the following intensity data groups, based on the Contrast QC thresholds:

      - All
      - In Bounds
      - Out of Bounds

Additional custom groupings of CEL files can also be made.

Removing poor quality CEL files from the data set can improve the quality of the genotypes of the remaining CEL files.

3. Perform Genotyping

   A. Select a group of intensity data files (CEL) in a data set.

   B. Perform genotyping analysis on the group of files, as described in Chapter 7, *Genotyping Analysis* on page 89.

      For mapping arrays, one intensity data group can contain CEL files from two different array types. But during genotyping, GTC will automatically separate them and genotyping results will be grouped by array type. Users can make a custom genotype result batch and manually add CHP files with different array types.

   C. Review the initial genotyping analysis QC data in the CHP Summary Results table, using Call Rate and other metrics, as described in *CHP Summary Table* on page 110.

   D. Create new Intensity Data file group for samples with good performance in initial genotyping analysis and perform a second genotyping analysis, as described in *Creating a Custom Intensity Group from the CHP File Data* on page 117.

   E. View the SNP results in the SNP Summary Results table, as described in Chapter 8, *Review the Genotyping Results* on page 124

      NOTE: You need a SNP list to view genotype result data. See *Create a SNP List* on page 125.

   F. Review the clustering performance for SNPs of interest in the SNP Cluster Graph.

      See Chapter 9, *Using the SNP Cluster Graph* on page 148.

   G. Export the genotype calls for downstream analysis.

      See Chapter 10, *Exporting Genotype Results* on page 181.

4. Copy Number/LOH Workflow for 100K/500K Arrays

   To perform a CN/LOH analysis for 100K/500K arrays, you must have both the CEL intensity data files and the genotyping CHP files for the arrays you wish to analyze.

   A. Perform Copy Number and/or LOH analysis in GTC, producing:

      - Copy Number Data Files
      - LOH Data Files

      See Chapter 12, *Copy Number & LOH Analysis for Human Mapping 100K/500K Arrays* on page 205.

   B. Run the Segment Reporting Tool on the CN files to generate:

      - Segment Data Files
      - Segment Summary file
      - Custom Region Data files
      - Custom Region Summary File

      See *Using the Segment Reporting Tool* on page 285.

5. Review data in the GTC Browser

   See *Loading Data into the GTC Browser* on page 305.

6. Export data for further analysis.

Two-Step Genotyping Workflow

The two-step genotyping workflow enables you to get optimal call rates when working with genotyping data.

In the two-step workflow, you evaluate the performance of the array data using both intensity QC metrics and initial genotyping call rate in the following steps:

1. Perform intensity QC and remove samples that do not meet the QC thresholds.
2. Perform a first round of genotyping on the remaining samples.
3. Remove samples based on outlier call rates (for Axiom arrays, use a call rate < 97% as the cutoff).
4. Perform a second round of genotyping to get optimal call rates.

   This workflow is described in more detail in *Creating a Custom Intensity Group from the CHP File Data* on page 117.

Working with Commands in Genotyping Console

Commands in Genotyping Console can be accessed from:

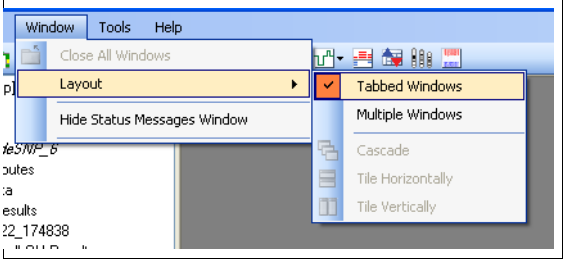
- Main menus
- Tool bar shortcuts
- Right-clicks on tree items
- Right clicks on table rows
- Right-clicks on graphs or from the graph tool bar

The tree items serve dual functions, organizing the data and results as well as guiding you through the workflow. The file menus are context sensitive, which means that some commands will be hidden until you've selected the items in the tree or table to which the command applies.

Window Layout Options

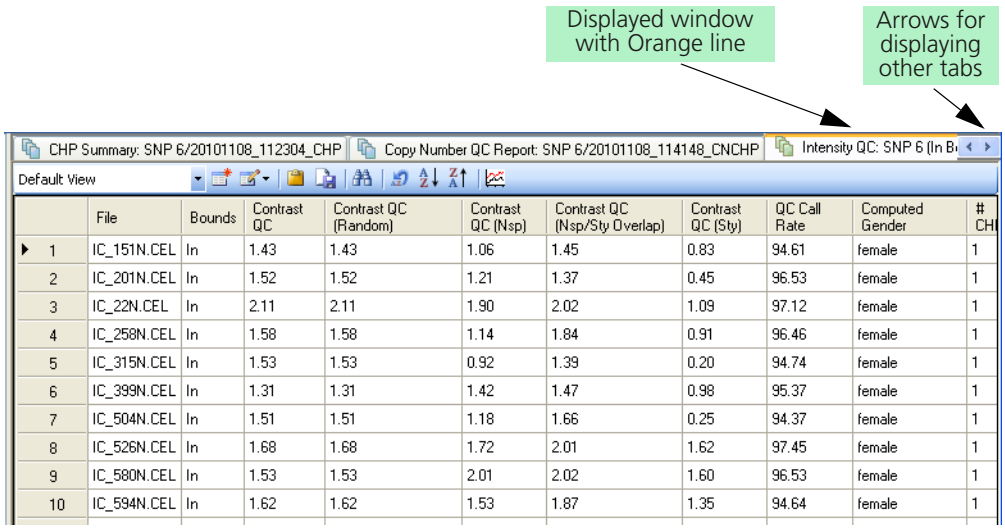
Genotyping Console windows can be arranged either as tabbed windows or multiple windows. To select a layout option, choose Tabbed Windows or Multiple Windows from the Window/Layout menu (Figure 2.15).

Figure 2.15 Window layout options



In the tabbed window layout, each open table or graph fills the entire available space and switching between active windows can be accomplished by clicking the tabs at the top of the window. The active window is highlighted with a white background and an orange line on the top (Figure 2.16).

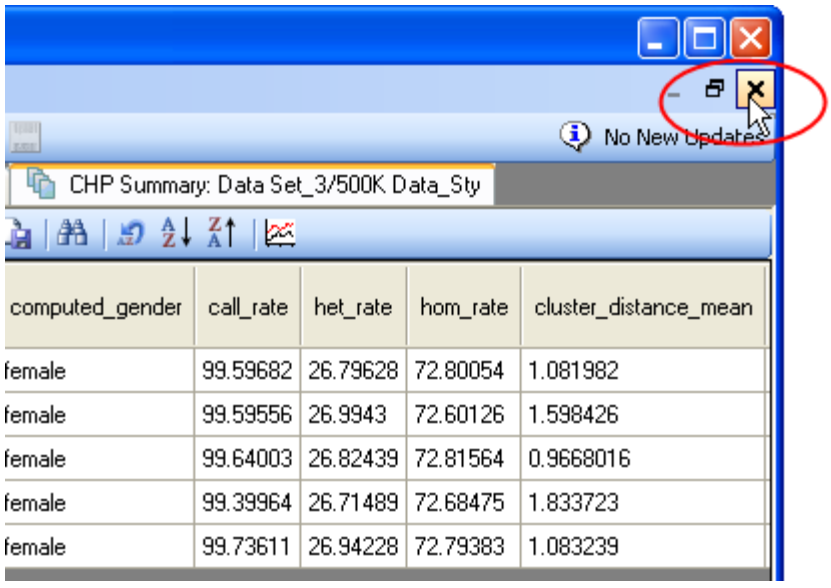
Figure 2.16 Tabbed window layout



	File	Bounds	Contrast QC	Contrast QC (Random)	Contrast QC (Nsp)	Contrast QC (Nsp/Sty Overlap)	Contrast QC (Sty)	QC Call Rate	Computed Gender	# CH
1	IC_151N.CEL	In	1.43	1.43	1.06	1.45	0.83	94.61	female	1
2	IC_201N.CEL	In	1.52	1.52	1.21	1.37	0.45	96.53	female	1
3	IC_22N.CEL	In	2.11	2.11	1.90	2.02	1.09	97.12	female	1
4	IC_258N.CEL	In	1.58	1.58	1.14	1.84	0.91	96.46	female	1
5	IC_315N.CEL	In	1.53	1.53	0.92	1.39	0.20	94.74	female	1
6	IC_399N.CEL	In	1.31	1.31	1.42	1.47	0.98	95.37	female	1
7	IC_504N.CEL	In	1.51	1.51	1.18	1.66	0.25	94.37	female	1
8	IC_526N.CEL	In	1.68	1.68	1.72	2.01	1.62	97.45	female	1
9	IC_580N.CEL	In	1.53	1.53	2.01	2.02	1.60	96.53	female	1
10	IC_594N.CEL	In	1.62	1.62	1.53	1.87	1.35	94.64	female	1

To close a tabbed window, use the button at the top right of the tab (Figure 2.17).

Figure 2.17 Close a tabbed window

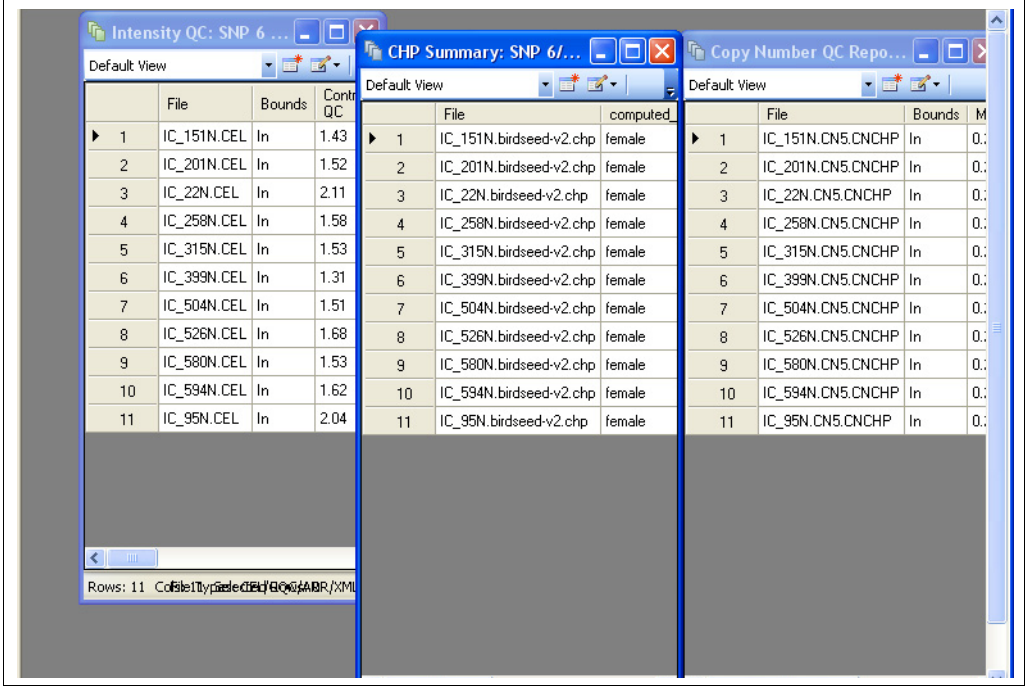


computed_gender	call_rate	het_rate	hom_rate	cluster_distance_mean
female	99.59682	26.79628	72.80054	1.081982
female	99.59556	26.9943	72.60126	1.598426
female	99.64003	26.82439	72.81564	0.9668016
female	99.39964	26.71489	72.68475	1.833723
female	99.73611	26.94228	72.79383	1.083239

In the Multiple Window layout (Figure 2.18), each open table or graph can be:

- Individually sized
- Expanded to the maximum size
- Minimized

Figure 2.18 Multiple layout



- Displayed in a cascade, tiled horizontally, or tiled vertically (see below)

To select the Cascade, Tile Horizontally, or Tile Vertically layout:

From the Window Menu, select **Layout > [display option]**:

- Cascade (Figure 2.19)
- Tile Horizontally (Figure 2.20)
- Tile Vertically (Figure 2.21)

Figure 2.19 Cascade layout

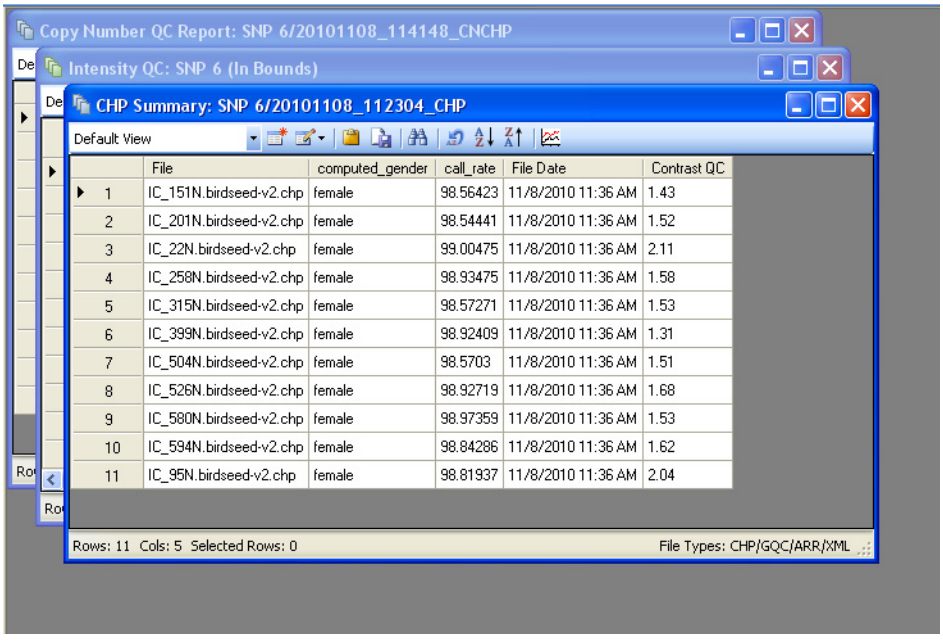


Figure 2.20 Horizontal Tiled layout

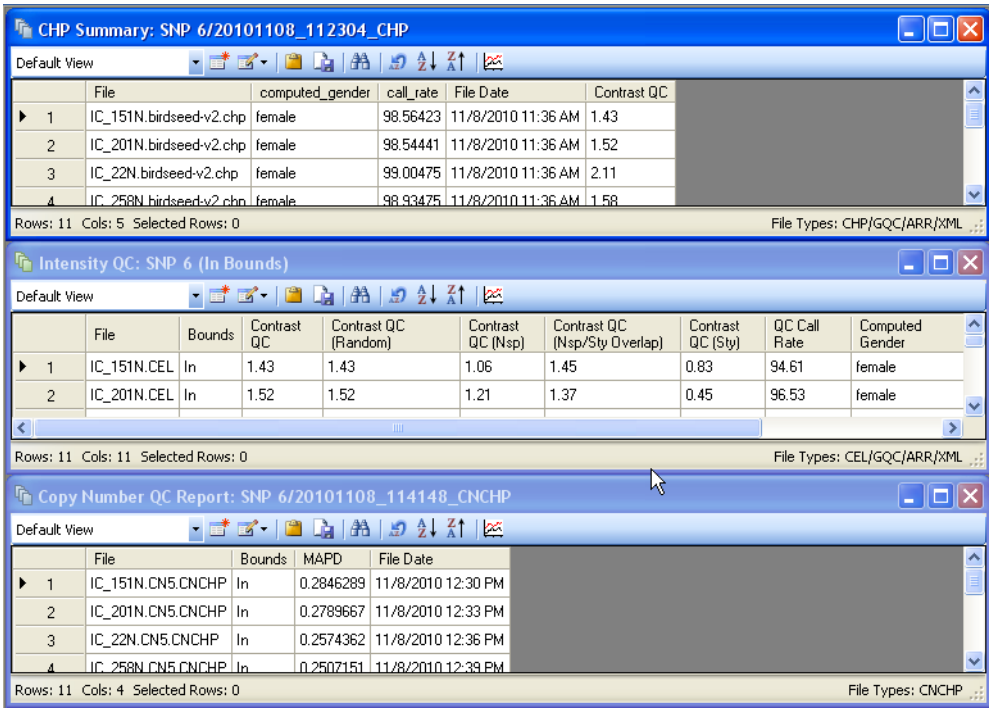
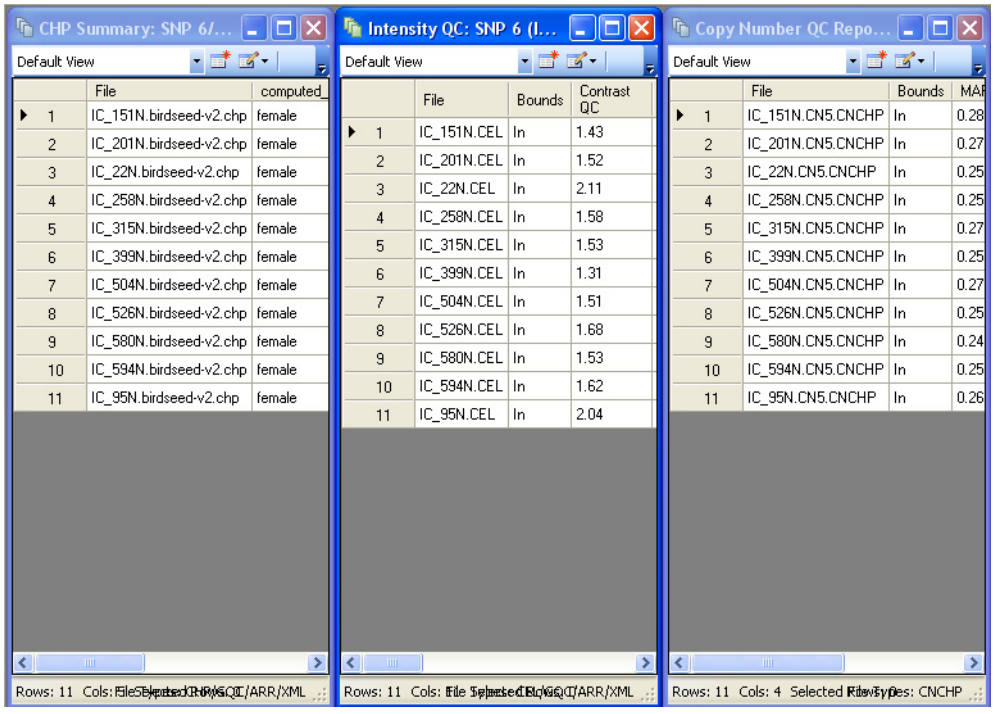


Figure 2.21 Vertical Tiles layout



User Profiles

A user profile stores a user's preferences for custom analysis settings, table and graph viewing options, and other application settings.Security by profiles is not provided by the application; it is simply a means of storing application parameters.

This chapter describes:

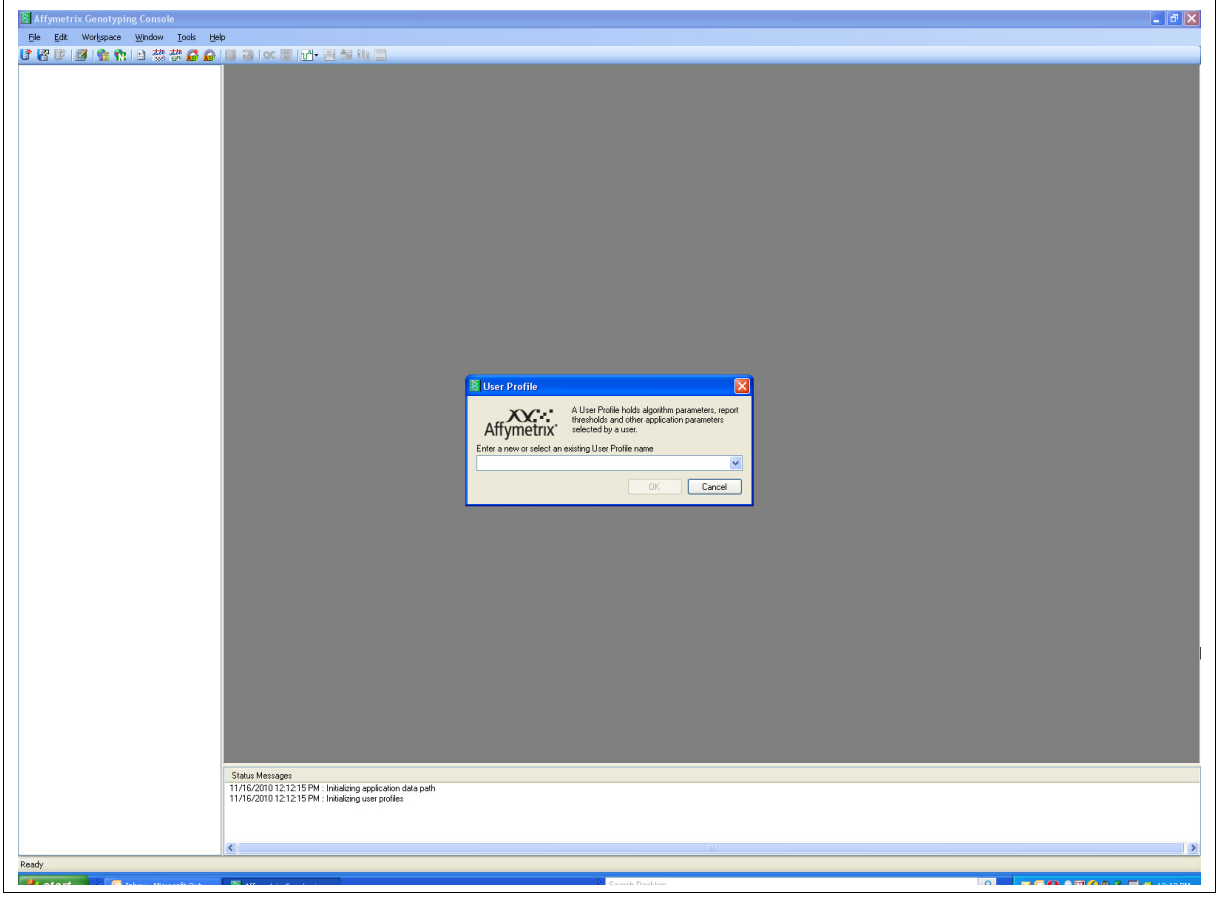
Creating and Selecting a User Profile

You can create a new user profile or create a previously selected one when you start Genotyping Console.

To create a new User Profile:

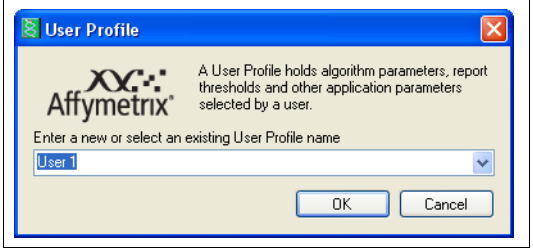
1. Start Genotyping Console by double-clicking on its shortcut on the Desktop; or
   From the Windows Start Menu select **Programs** > **Affymetrix** > **Genotyping Console**.
   The Genotyping Console opens with the User Profile dialog box displayed (Figure 3.1).

Figure 3.1 Genotyping Console main window and User Profile dialog box



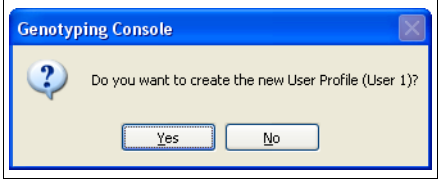
2. Type in a name for the new profile in the User Profile dialog box (Figure 3.2).

Figure 3.2 User Profile dialog box



3. Click **OK**.
   The software will prompt you to create the new profile (Figure 3.3).

Figure 3.3 Confirmation dialog box



After setting up a user profile, the software will either prompt you to select:

- a library file path (if Affymetrix Command Console is not installed on the workstation or the library folder has not already been specified during a prior session). See *Setting the Library Path* on page 34
- a workspace to open. See *Creating a New Workspace* on page 46.

To select an existing Profile:

- Use the drop-down menu on the User Profile window (Figure 3.4).

Figure 3.4 User Profile dialog box

NOTE: You can select a different profile without terminating the program, but the Workspace must be closed.

To change profiles:

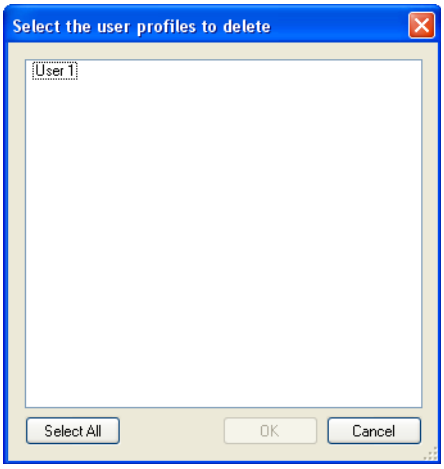
1. From the Edit menu, select **Change User Profile**.
   The User Profile dialog box appears (Figure 3.4).
2. Enter a new profile name or select a previously generated profile from the drop-down box (see the User Profile dialog box (Figure 3.4)).

Deleting a User Profile

To remove profiles no longer needed:

1. From the Edit menu, select **Delete User Profile**.
   The Select the user profiles to delete dialog box opens (Figure 3.5).

Figure 3.5 Delete a user profile



2. Select the User Profile to be deleted and select **OK**.
   The selected User Profile, and all parameter files associated with the profile, will be removed. To add a new User Profile, see *Creating and Selecting a User Profile* on page 31.

Library & Annotation Files

Genotyping Console requires information stored in library files to analyze the CEL files generated by GCOS or Affymetrix GeneChip® Command Console™ (AGCC) software. These library files are available from NetAffx and can be downloaded within Genotyping Console. Genotyping Console downloads only those library files it requires from NetAffx for analysis; these files are not registered with GCOS or Command Console and are not sufficient to scan arrays.

Genotyping Console uses SQLite annotation files (*.annot.db) to display and export additional information about the SNP and CN probe sets (such as Chromosome, chr start and chr stop, dbSNP RS ID, etc.) as well as for certain analysis and filtering steps. You can use custom annotation files in GTC 4.2, but the files must be in SQLite format.

You can use the Annotation Converter (AC) to generate SQLite annotation files for Axiom myDesign arrays. Users will be able to customize NetAffx annotation files by using the AC with text (.csv) files as input.

See the documentation on the Annotation Converter for more information.

GTC 4.2 updates genotyping config files from GTC 4.0; depends on the file types updated, a subfolder will be created within the library folder to host different corrupted or outdated config files.

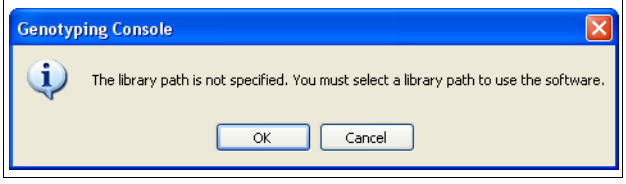
The following sections in this chapter include:

Setting the Library Path

If Genotyping Console software is installed on a workstation with Command Console, the library path is automatically set to the library path used by Command Console. If Command Console is not installed and a path is not specified, Genotyping Console prompts you to select a location for the library path (Figure 4.1). You can set the library path without terminating the program, but any open workspace(s) must be closed.

NOTE: Users must have write access to the library folder. Make sure that all of the library files for use in Genotyping Console are copied to only one library folder. You can select any location for the library files folder; however it is recommended that the library folder not be located within the GTC application folder.

Figure 4.1 Library path notification

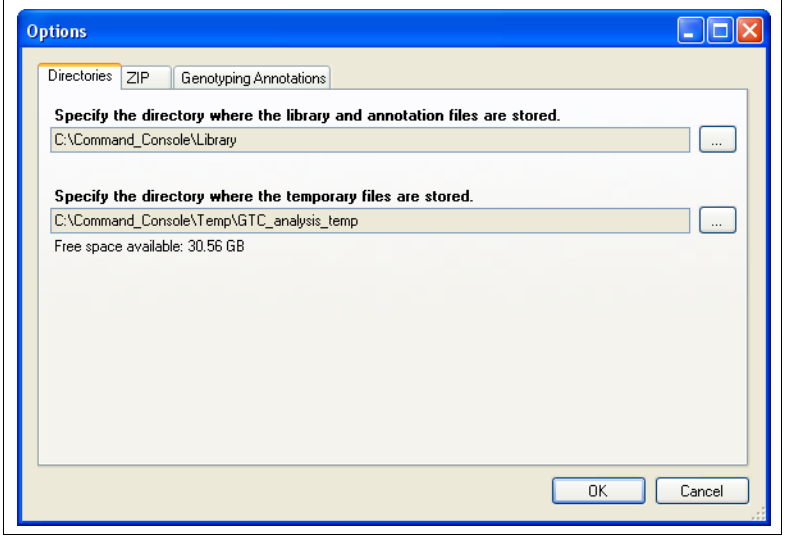


To change an existing library path:

1. Close any open workspaces.
2. Click the **Options** tool bar shortcut; or from the File menu, select **Option**.

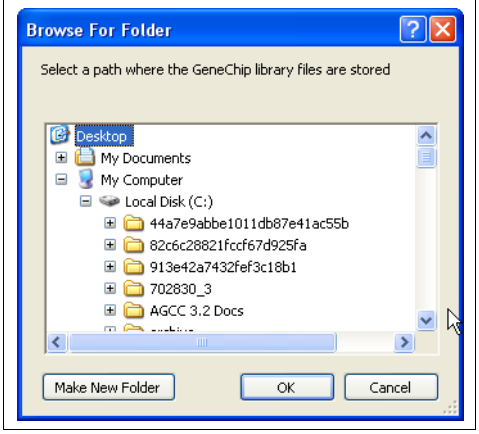
The Options dialog box appears (Figure 4.2).

Figure 4.2 Options dialog box, Directories tab



3. In the Directories tab, enter the path to the new directory or click the **Browse** button ... . The Browse For Folder dialog box opens (Figure 4.3).

Figure 4.3 Browsing for library folder

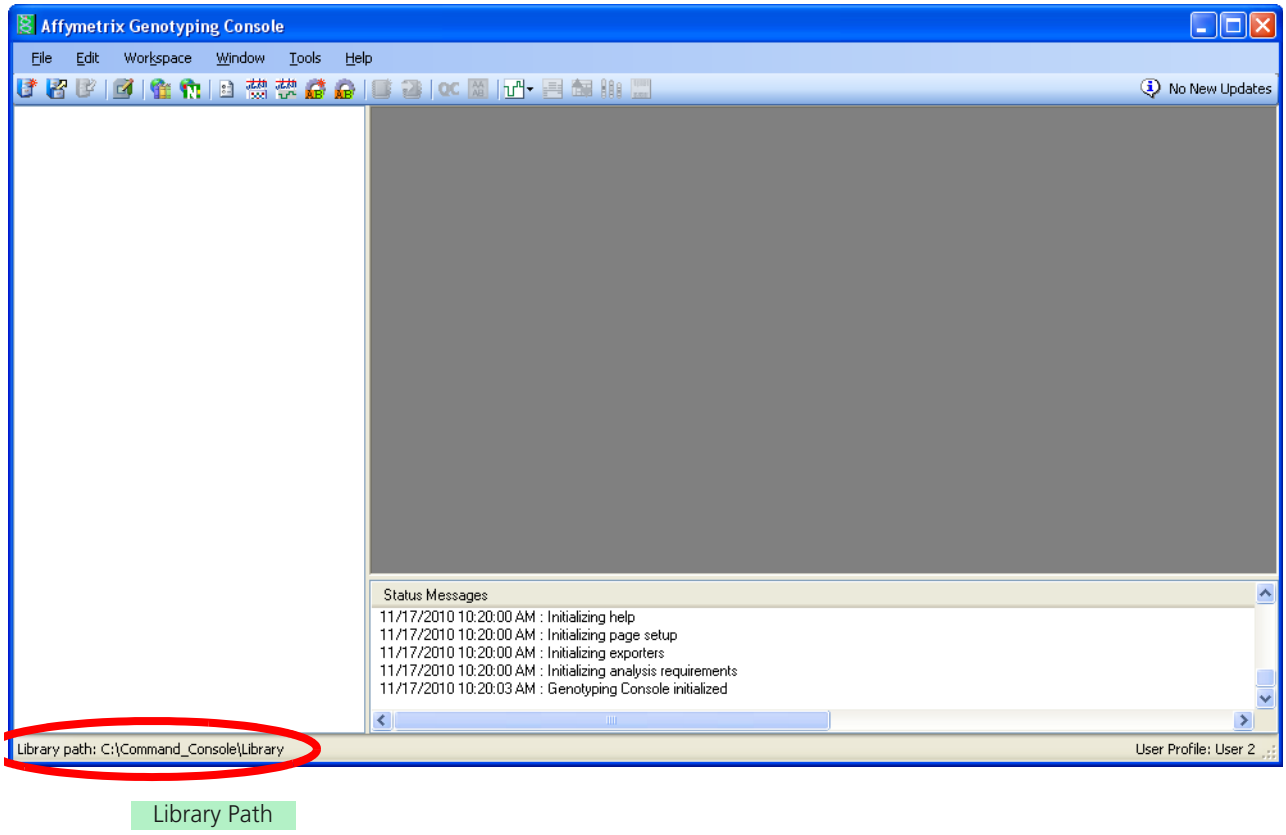


NOTE: You can select any location for the library files folder. If the Affymetrix GeneChip® Operating System software (GCOS) is installed on your system, Affymetrix recommends that you do NOT select the GCOS library file directory as the library file directory for Genotyping Console, to avoid confusion. Do not place any library files in a subfolder. Genotyping Console cannot find library files in a subfolder!

   A. Browse to the folder which contains the library files or create a new folder for your library files. Make sure all library files for use in Genotyping Console are copied to this folder or are downloaded to this folder through NetAffx using the GTC download functions from the File menu.
   B. Click **OK** in the Browse to Folder dialog box.

4. Click **OK** in the Options dialog box.
   The selected library path is displayed in the bottom left corner of the application window (Figure 4.4).

Figure 4.4 Genotyping Console main window displaying the library path



NOTE: **GCOS users must use the Data Transfer Tool (DTT) using the Flat File option to transfer files to be analyzed by Genotyping Console software from the GCOS database to an independent folder, in order to retain all sample attributes. More detailed instructions can be found at www.affymetrix.com.**

Obtaining Library & Annotation Files

The library and SQLite annotation files can be downloaded from the Affymetrix website, NetAffx, or from within GTC.

There are several ways to obtain library and annotation files.

Table 4.1 Obtaining Library and Annotation Files

To Obtain…	Computer With Internet Access	Computers Without Internet Access
Library files	*Download Library Files* on page 37 from within GTC (click the tool bar button)	*Manually Copy Library Files* on page 38
Annotation files	*Download Annotation Files* on page 38 from within GTC (click the tool bar button)	*Manually Copy & Optimize Annotation Files* on page 39

Downloading the GTC_Analysis_Files Zip Package

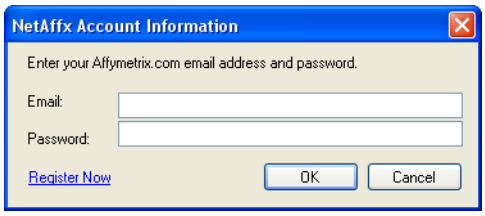
The zip package contains library files for Axiom™ Genome-Wide Human Arrays (CEU and ASI) and the Axiom Genome-Wide BOS 1 Array. The zip package can be downloaded from the Affymetrix website. It includes the files required for processing samples processed with Reagent Version 1 or Reagent Version 2.

1. Go to the Affymetrix web site and download the zip package GTC_Analysis_Files.
2. Unzip this file and then copy the files from the GTC_Analysis_Files folder to the Genotyping Console library folder.

Download Library Files

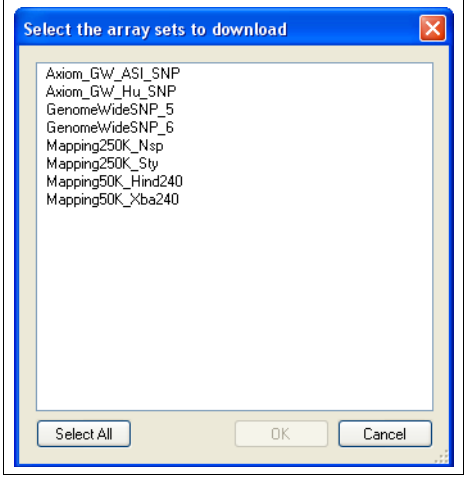
1. Click the **Download Library Files button**  : or
   From the File menu, select **Download Library Files**.
   The NetAffx Account Information dialog box opens (Figure 4.5).

Figure 4.5 NetAffx Account Information dialog box



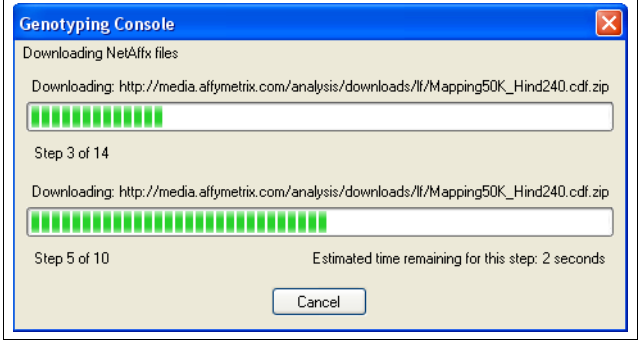
2. Enter your Affymetrix account information and click **OK**.
   If you do not have a NetAffx account, click **Register Now** which launches *www.affymetrix.com*. Follow the instructions to set up an account.
   The Select the array sets to download dialog box opens (Figure 4.6).
3. Select the array set library files to download and click **OK**.

Figure 4.6 Select the array sets to download dialog box



The Downloading NetAffx files box opens and displays the download progress (Figure 4.7).

Figure 4.7 Download progress



NOTE: The download may take several minutes or more, depending on the connection speed, as the library files are large. Please be patient.

Manually Copy Library Files

If the workstation with Genotyping Console does not have an Internet connection and cannot download the library files, manually copy the necessary files to the library folder.

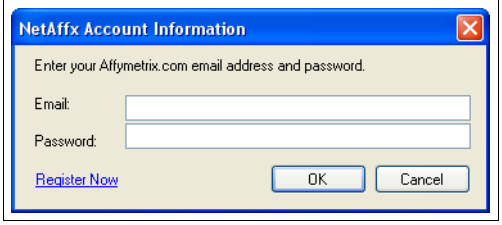
NOTE: Do not create subdirectories within the library file folder. Genotyping Console does not look at subdirectories!

Download Annotation Files

1. Click the **Download Annotation Files** button on the tool bar. Alternately, select **File > Download Annotation Files** on the menu bar.

   The NetAffx Account Information dialog box opens (Figure 4.8).

Figure 4.8 NetAffx Account Information

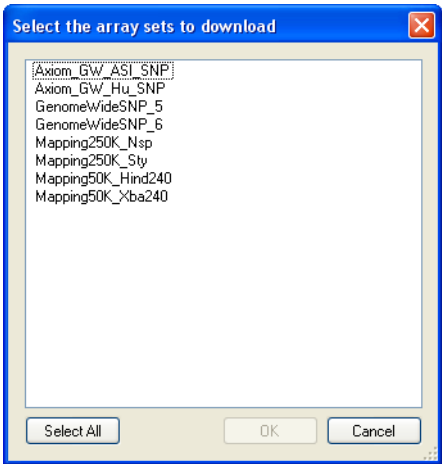


2. Enter your NetAffx account information and click **OK**.

   If you do not have a NetAffx account, click Register Now which launches *www.affymetrix.com*. Follow the instructions to set up an account.

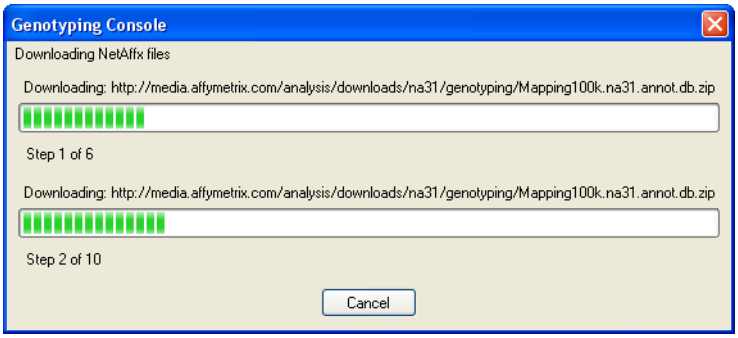
   The Select the array sets to download dialog box opens (Figure 4.9).

Figure 4.9 Selecting annotation files to download



3. Select the Array set annotation files to download and click **OK**.
   The download progress is displayed in the Downloading NetAffx files box (Figure 4.10).

Figure 4.10 Download progress



NOTE: Please be patient. The download may take several minutes or more, depending on the connection speed, as these files are large.

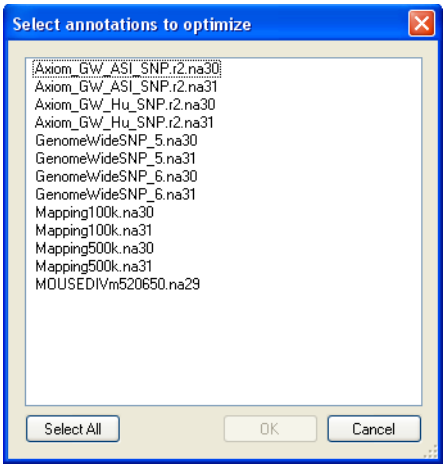
After Genotyping Console downloads the selected *.annot.db file from NetAffx, it optimizes the file for application use. This may take several minutes. We recommend that you not cancel this operation. If you cancel this operation, you can manually optimize the annotation file (select **File > Optimize Annotation Files** on the menu bar).

Manually Copy & Optimize Annotation Files

If the workstation with Genotyping Console does not have an Internet connection and cannot download the annotation files, manually copy the necessary files to the library folder. After the annotation files are copied to the library folder, they must be optimized to improve application performance.

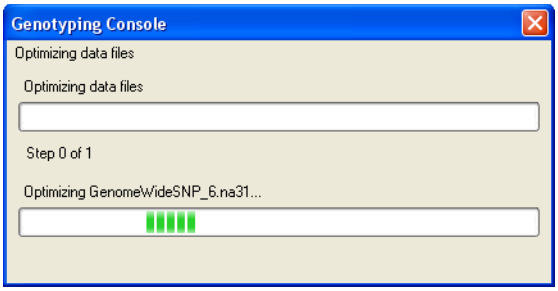
1. Copy the required annotation files (.annot.db) to the library folder.
2. From the File menu, select **Optimize Annotation Files**.
   The Select annotations to optimize dialog box opens (Figure 4.11).

Figure 4.11 Select annotations to optimize dialog box



3. Select the annotations file(s) to optimize, and click **OK**.
   The Optimizing data files box displays the progress of optimization (Figure 4.12).
   File optimization may take several minutes or more, depending on your computer configuration.

Figure 4.12 Optimizing annotation files progress



NOTE: If you do not manually optimize the annotation files, GTC automatically optimizes the file the first time it is used.

NOTE: To export Human Mapping 100K/500K analysis results or to further process samples in segment reporting tool (SRT), na24 annotation files (*.annot.db) are required.

NOTE: For Genome-Wide Human SNP Array 6.0, to export CN/LOH analysis results or to further process samples in SRT, na25 or higher annotation files (*.annot.db) are required, depending on which annotation version was used to generated the CNCHP files.

Annotation Options

You can select a particular annotation version for use with an array type in GTC using the Genotyping Annotations tab of the Options dialog box. Pre-selecting an annotation version will let you avoid having a prompt window appear to select the version during later operations.

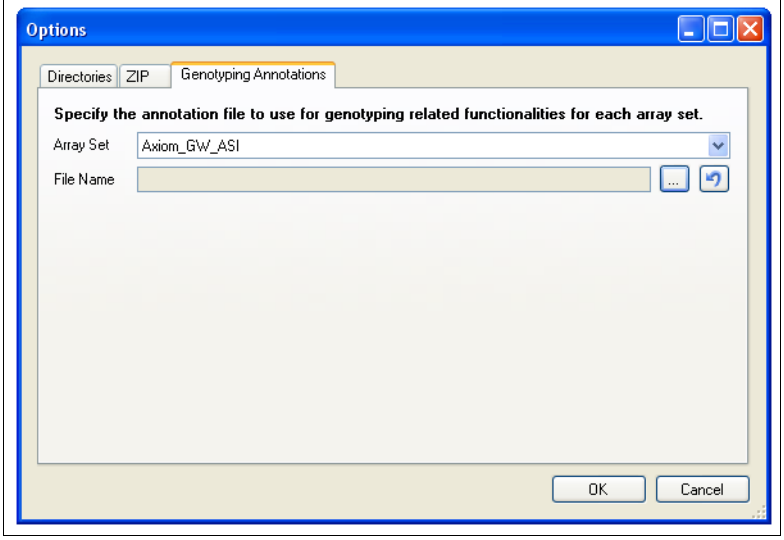
If you download a newer version of the annotation file, the selected annotation version will be updated to the newer version.

To select an annotation version:

1. Close any open workspaces.
2. Click the **Options** tool bar shortcut; or from the File menu, select **Option**.

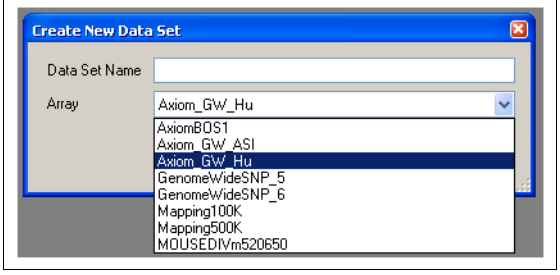
   The Options dialog box appears.
3. Click the Genotyping Annotations tab (Figure 4.13).

Figure 4.13 Options dialog box, Genotyping Annotations tab



4. Select the array set from the drop-down list (Figure 4.14).

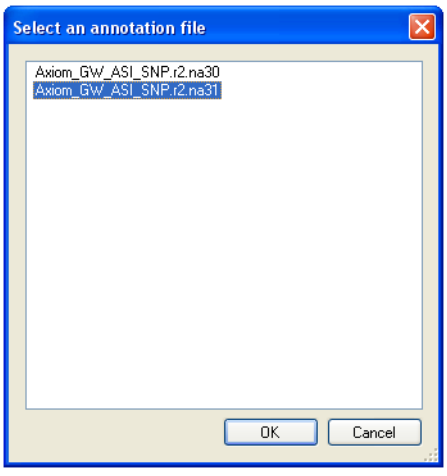
Figure 4.14 Selecting Array Set



5. Click the **Browse** button ... .

   The Select an Annotation file dialog box opens (Figure 4.15).

Figure 4.15 Select an annotation file dialog box



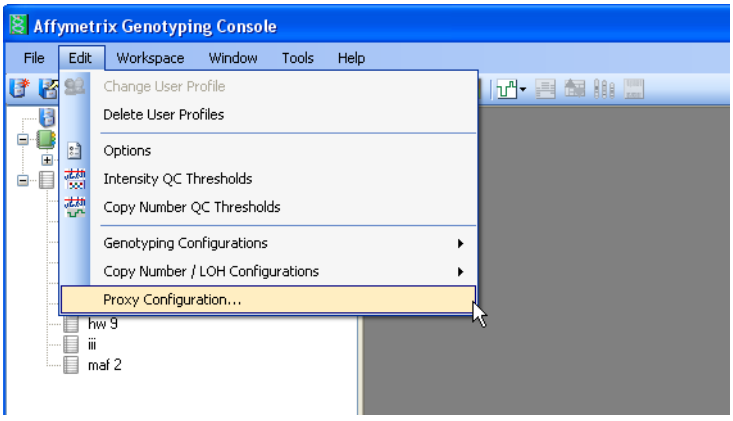
6. Select the desired annotation file in the list and click **OK** in the Select an annotation file dialog box.
7. Click **OK** in the Options dialog box.
   The selected annotation file is used as the default file.

Setting Proxy Server Access

This configuration should only be done if the user's system has to go through a proxy server to access Affymetrix NetAffx server. **Please contact your IT department if you do not know or are not sure about the answer to this question.**

1. From the Edit menu, select **Proxy Configuration…** (Figure 4.16).

Figure 4.16 Proxy Configuration menu item



The **Proxy Server Settings** dialog appears (Figure 4.17). By default, it has '**Use System Proxy**' option selected.

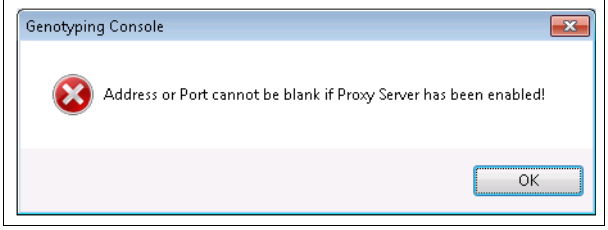
Figure 4.17 Proxy Server Settings

2. Select '**Use Custom Proxy**' and enters the proxy server address and port for their proxy server.
3. Click **OK**.

   GTC software validates the entries for the proxy server address and port.

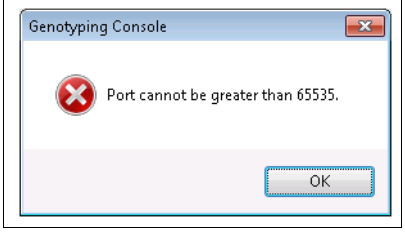
   If either the proxy server address or the port is left blank, the following dialog box pops up (Figure 4.18).

Figure 4.18 Blank Address or port notice



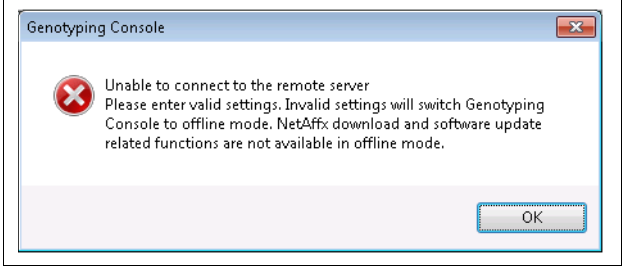
Proxy port cannot be greater than 65535. Otherwise, the following dialog box will pop up (Figure 4.19).

Figure 4.19 Incorrect port value notice



If the proxy server address or port is incorrect, the following dialog box pops up (Figure 4.20).

Figure 4.20 Unable to connect to the remote server notice

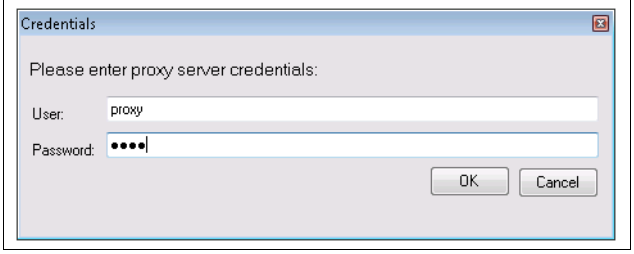


Click **OK** to return to the **Proxy Server Settings** dialog box ().

If you then click **Cancel** on the '**Proxy Server Settings**' dialog box, GTC software exits proxy server configuration and defaults to the previous successful setting.

Once the proxy server address and port validation is successful and the server requires user authentication, the following '**Credentials**' dialog pops up. (Figure 4.21)

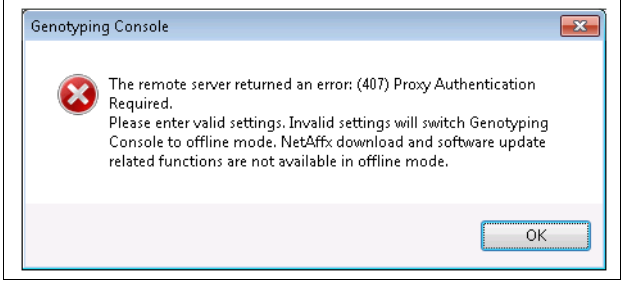
Figure 4.21 Proxy Server Credentials dialog box



NOTE: Please note that this user id and password is not the same ID and password used to connect to the Affymetrix NetAffx server

If validation fails, the following dialog box pops up (Figure 4.22).

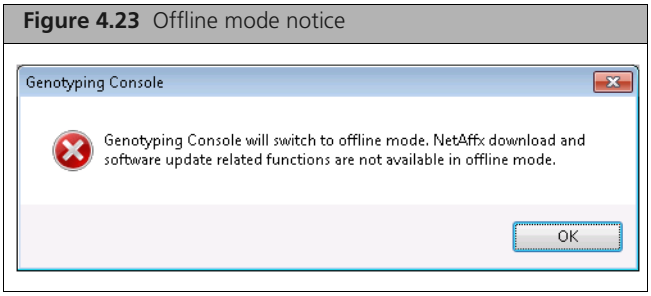
Figure 4.22 Proxy Authentication Error notice



Click **OK** to return to the '**Credentials**' dialog box ().

If the user clicks '**Cancel**' on the '**Credentials**' dialog box, the following dialog box pops up (Figure 4.23).

Figure 4.23 Offline mode notice



GTC software cannot download library and annotation files from the Affymetrix NetAffx server.

Once the proxy server userid/password validation succeeds, GTC software can download library and annotation files from the Affymetrix NetAffx server for the rest of the user session. The Proxy password must be entered at the next start of GTC software.

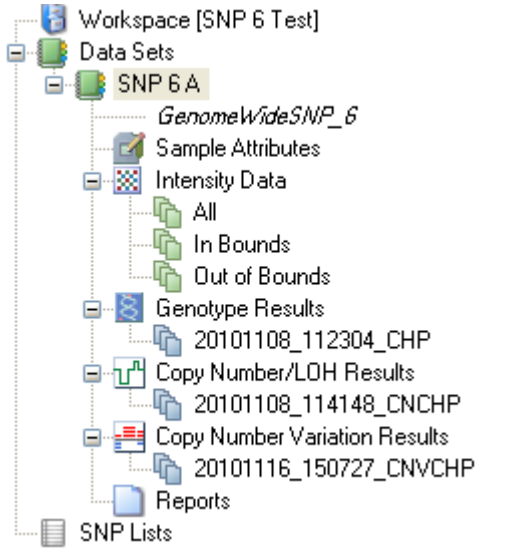
Workspaces & Data Sets

To get started using Genotyping Console, you will create a workspace and add a data set(s) consisting of a collection of the following types of files for analysis and examination:

- Sample files (ARR/XML)
- Intensity files (CEL)
- Genotyping files (CHP)
- Copy number (CNCHP), LOH (LOHCHP), and/or copy number segment files (cn_segments)/copy number custom region files (custom_regions)
- Copy number variation files (CNVCHP)

The files in the workspace are organized in data sets and SNP lists (Figure 5.1).

Figure 5.1 Data tree with workspace and data set with data



Data sets contain:

- Sample Attributes
- Intensity Data
- Genotype Results
- Copy Number/LOH results (if available)
- Copy Number Variation results (if available)
- Reports

The workspace file stores the locations of the data files, not a copy of the data files themselves. Only one user can have a workspace open at one time. If other users need to have access to the same data files, they can either make a personal copy of a Workspace file that is not in use, or create a new workspace and add the same data files to the new workspace. Simultaneous genotyping of the same set of CEL files within two workspaces is not recommended.

Creating a new workspace and loading it with data files requires the following sets of steps:

1. *Creating a New Workspace* on page 46
2. *Creating a Data Set* on page 48

3. *Adding Data to a Data Set* on page 49

This chapter also describes:

- *Opening a Created Workspace File* on page 56
- *Viewing the Location of Data Files* on page 58
- *Removing Data from a Data Set* on page 60
- *Viewing attributes in the Sample Attributes Table* on page 63
- *Editing Sample Attributes* on page 64
- *Locating Missing Data* on page 67
- *Sharing Data* on page 69

NOTE: GCOS users must use DTT v1.1, using the Flat File option, to transfer files to be analyzed by Genotyping Console from the GCOS database to an independent folder, in order to retain all sample attributes. More detailed instructions can be found at www.affymetrix.com.

NOTE: Affymetrix recommends that you do not use long file names for the .CEL and .CHP files, since these long names can cause display problems in the Heat Map Viewer. The status bar in the Heat Map Viewer will not be able to display all the information if the CNCHP and CNVCHP file names (derived from the .CEL file names) are too long.

NOTE: GTC 4.2 workspaces cannot be opened in versions of GTC before 4.1. Workspaces created in earlier versions of GTC can be opened in GTC 4.2, but then cannot be used in earlier versions of GTC.

Creating a New Workspace

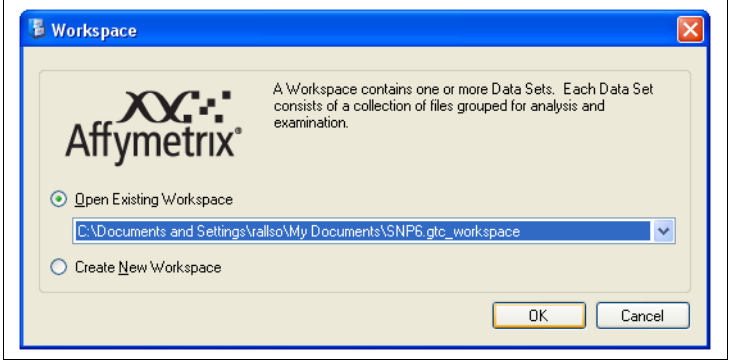
If you create a new workspace, Genotyping Console will also prompt you to:

1. *Create a new data set* (page 48).
2. *Select data to add to the data set* (page 49).

To create a new workspace:

1. Do one of the following:
   A. Launch GTC and create a user profile, if necessary.
      See Chapter 3, *User Profiles* on page 31.
      The Workspace dialog box opens (Figure 5.2).

Figure 5.2 Workspace dialog box



B. Select the **Create New Workspace** radio button and select **OK**.
The Save As dialog box opens (Figure 5.3).

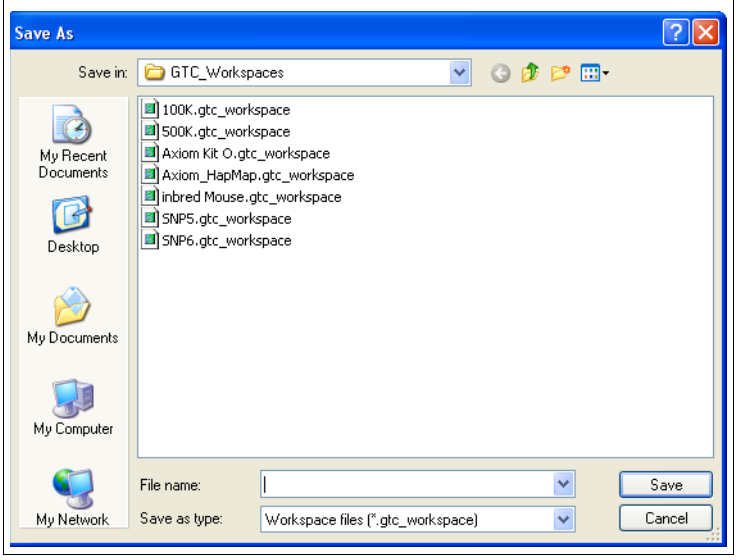
Or:

A. Close all workspaces in GTC.

NOTE: Only one workspace can be opened at a time.

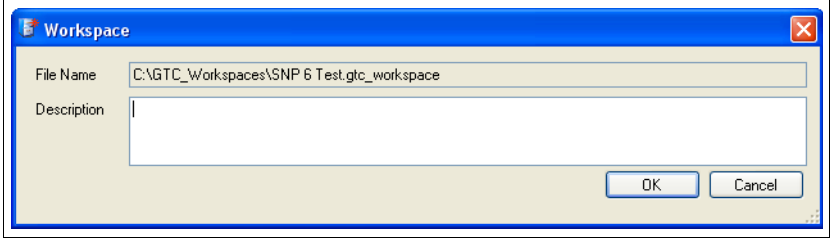
B. From the File menu, select **New Workspace**; or
Click the New Workspace button in the tool bar.
The Save As dialog box opens (Figure 5.3).

Figure 5.3 Save As dialog box



2. Use the Save As dialog box navigation tools to find or create a folder for the workspace.
3. Enter the Workspace name in the File name box.
4. Click **Save**.
The Workspace description dialog box opens (Figure 5.4).

Figure 5.4 Workspace description dialog box



5. Enter a description of the Workspace by typing in the Description window (optional).
6. Click **OK**.
   GTC prompts you to create a Data Set (see *Creating a Data Set* on page 48).

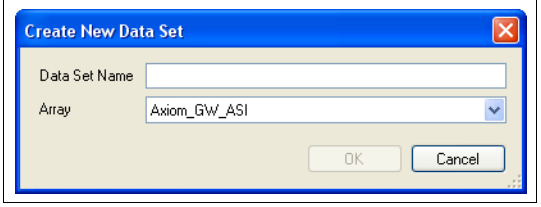
Creating a Data Set

To create a new data set in a workspace:

1. Do one of the following:
   - Click the **Create Data Set** shortcut  on the main tool bar; or
   - Right-click the **Data Sets** node in the tree and select **Create Data Set**; or
   - From the Workspace menu, select **Data Sets** > **Create Data Set**.

   The Create New Data Set dialog box opens (Figure 5.5).

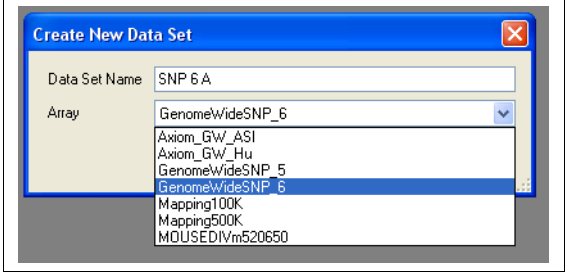
Figure 5.5 Create New Data Set dialog box



NOTE: This dialog box opens automatically when you have finished creating a new workspace.

2. Enter a name for the data set.
3. Select the array type for the new Data Set from the Array drop-down list (Figure 5.6).

Figure 5.6 Create New Data Set dialog box, Array drop-down list



4. Click **OK**.

NOTE: Data Sets can only contain files which belong to the same array type. For example, a GenomeWideSNP_5 Data Set cannot contain data from the GenomeWideSNP_6 array. If you wish to have data from multiple arrays in one Workspace, you need to create at least one Data Set for each array type.

NOTE: For Human Mapping 100K/500K, you can include arrays from both enzyme sets (for example, Mapping 250K_Nsp and Mapping250K_Sty for a set of 500K arrays) in the same data set. If you select a CEL intensity group that contains both types of arrays, the resulting genotyping data will be divided into two results sets, one for each enzyme set.

After you create a data set, the software will automatically prompt you to add data to this data set. See *Adding Data to a Data Set* on page 49 for more information.

Adding Data to a Data Set

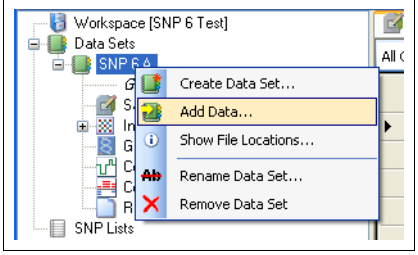
NOTE: Only data files (ARR/XML, CEL, or CHP) generated by Affymetrix® software or GeneChip® compatible software partners can be imported into Genotyping Console. Any supported data files that are edited outside of these software packages may cause import to fail or Genotyping Console software to crash.

NOTE: Affymetrix recommends using data files in AGCC format, as there is only limited support for GCOS files. For example, editing of XML sample attributes is not supported. Also, CHP files that are generated by Genotype Console and then imported into another workspace will not include sample attribute information if the CHP files were generated from GCOS-format CEL files. Affymetrix recommends using the Data Transfer Tool (DTT v1.1.1, provided with GCOS) Flat File transfer out option to create a copy of the XML and CEL files for use by Genotyping Console. For more information, go to: ***http://www.affymetrix.com/support/downloads/manuals/data_transfer_tool_user_guide.pdf*** or *www.affymetrix.com*; then to *Support/Technical/Tutorial/GCOS*.

To add ARR/XML, CEL, and/or CHP files to a existing data set:

1. Do one of the following:
   - Right click the data set *in* the tree and select **Add Data** on the shortcut menu (Figure 5.7).

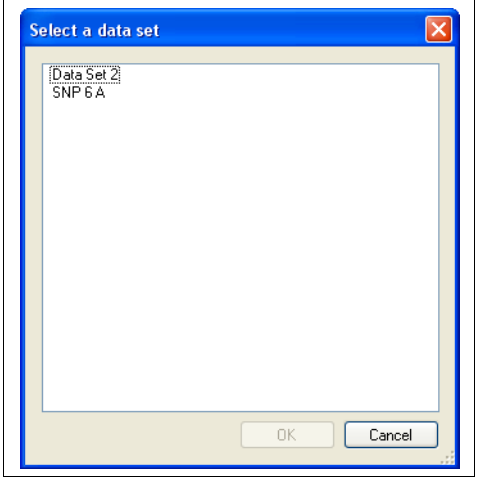
Figure 5.7 Data set shortcut menu



   - Click on the **Add Data** shortcut on the main tool bar.
   - From the Data Sets menu, select **Add Data.**
   - Use the CTRL-A shortcut.

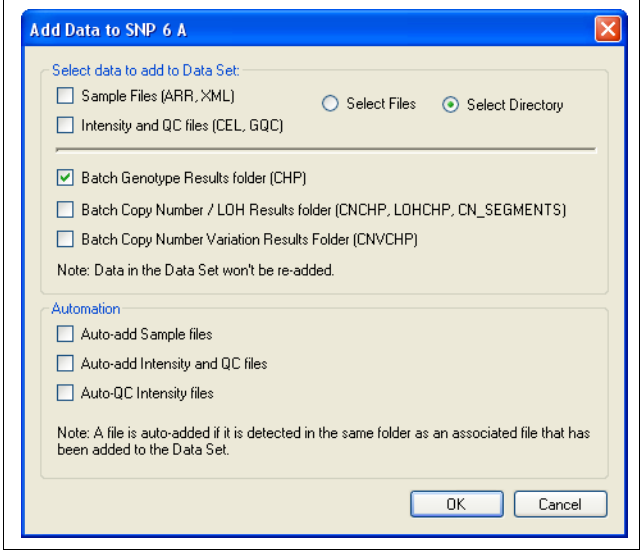
   If more than one data set is available, the Select a data set dialog box opens (Figure 5.8).

Figure 5.8 Select a data set dialog box



2. Select the data set from the list and click **OK**.
   The Add Data to [data set name] dialog box opens (Figure 5.9).

Figure 5.9 Add Data to [data set name] dialog box



NOTE: This dialog box appears automatically when you have finished creating a new data set.

The Add Data dialog box provides a set of options for adding data to a data set.

3. Select the data type (ARR/XML, CEL and GQC, and/or CHP) to add to the newly created Data Set using the options described in the table below (Table 5.1).

Table 5.1 Add data options

Select data to add to Data Set	Description
Select Files radio buttons	Add files selected from a directory to the data set.
Select Directory radio button	Add all files in a selected directory to the data set.

Table 5.1 Add data options

Select data to add to Data Set	Description
Sample Files (ARR, XML)	If selected, Genotyping Console will add user-selected sample files to the Data Set. These files can be in either AGCC format (ARR, preferred) or GCOS format (XML).
Intensity and QC Files (CEL,GQC)	If selected, Genotyping Console will add user-selected Intensity (CEL) and associated Genotyping Console QC files (GQC) to the Data Set.
Batch Genotype Results folder (CHP)	If selected, Genotyping Console software will add CHP files in the user-selected folder. If the CHP files are not from the same batch genotyping operation, they will be separated into multiple Genotype Result groups.
Batch Copy Number/LOH Results folder (CNCHP, LOHCHP, CN_SEGMENTS, CUSTOM_REGIONS)	If selected, Genotyping Console software will add CNCHP and/or LOHCHP and CN_SEGMENTS and CUSTOM_REGIONS files in the user-selected folder.
Batch Copy Number Variation Results folder (CNVCHP)	If selected, Genotyping Console software will add CNVCHP files in the user-selected folder.

If you want to select an entire directory, click the Select Directory radio button.

NOTE: When loading a large set of files, it is recommended that you use the "Select Directory" option, load all contained files, and then optionally remove undesired files after import. Windows has a fixed buffer that limits how many files can be returned to the application using the "Select Files" option. It is possible to select more files than the Windows buffer causing only a subset of the files to be returned. The maximum number of files varies. As an example, when trying to add 800 ARR and CEL files to the Data Set at one time, although all files could be selected only a subset are actually added to the Workspace.

4. Check-mark any automated steps that should also occur, such as auto-add data or auto-QC intensity files using the options described in the table below (Table 5.2).

Table 5.2 Automation options

Automation	Description
Auto-add Sample Files	Some CEL files in the Data Set may be missing the associated Sample files. If this option is selected, Genotyping Console software will look for these Sample files in the same folder as the associated CEL files, and add them to the Data Set.
Auto-add Intensity and QC Files	Some sample files in the data set may be missing the associated CEL and QC files. If this option is selected, Genotyping Console will look for these CEL files in the same folder as the associated sample files, and add them to the data set. When a CEL file is added to the data set, Genotyping Console software will also load the associated QC file (.gqc), if it exists in the same folder as the CEL file.
Auto-QC Intensity Files	Genotyping Console software will automatically initiate QC analysis of imported CEL files that do not include QC information or are not associated with a QC file (.gqc), provided the necessary library files are present in the library folder.

5. Click **OK**.

NOTE: You must have write access to the folder in which the CEL files are located for GTC to be able to write QC information. If you only have read access, you must first copy the data to a folder where you have write access.

NOTE: Genotyping Console will only add files to the data set that use the same array type as the data set.

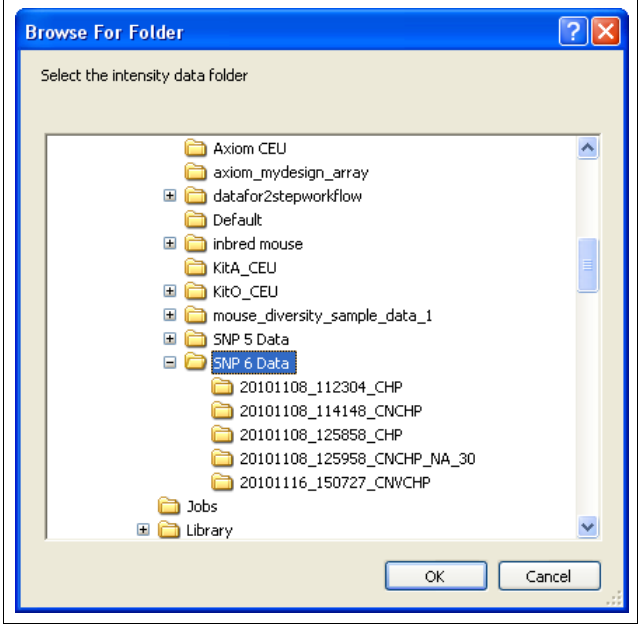
The next steps depend upon the types of data you wish to import and the options you have selected for the import:

- *Importing XML/ARR/CEL/GQC files by Selecting the Directory* on page 52
- *Importing XML/ARR/CEL/GQC files by Selecting Individual Files* on page 53
- *Selecting CHP Data* on page 53

Importing XML/ARR/CEL/GQC files by Selecting the Directory

If you have chosen to select the directory containing the files you wish to import, the Select the intensity data folder dialog box opens (Figure 5.10).

Figure 5.10 Select the intensity data folder dialog box



- Browse to the folder with the data you wish to import and click **OK**.

  If you are importing CHP files, another Browse for Folder dialog box opens, asking you to select the appropriate folder for the results. See *Selecting CHP Data* on page 53.

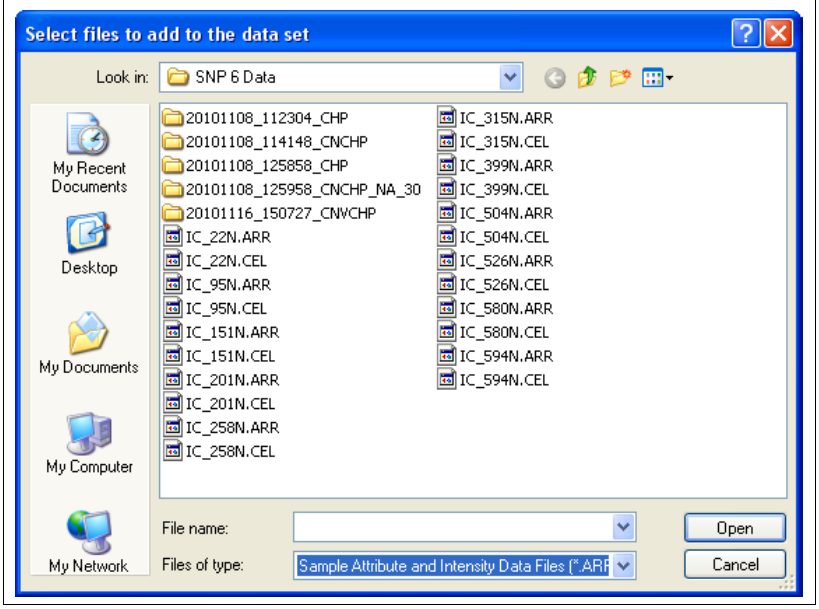
  If you are not importing CHP files, the loading progress bar displays the progress of the import. See *Viewing the Loading Progress* on page 55.

NOTE: When loading a large set of files, it is recommended that you use the "Select Directory" option, load all contained files, and then optionally remove undesired files after import. Windows has a fixed buffer that limits how many files can be returned to the application using the "Select Files" option. It is possible to select more files than the Windows buffer causing only a subset of the files to be returned. The maximum number of files varies. As an example, when trying to add 800 ARR and CEL files to the Data Set at one time, although all files could be selected only a subset are actually added to the Workspace.

Importing XML/ARR/CEL/GQC files by Selecting Individual Files

If you choose to select individual Sample files and/or Intensity and QC files, the Select files to add to the data set dialog box opens (Figure 5.11).

Figure 5.11 Select files to add to the data set



- Select the files to be imported and click **Open**.

TIP: You can quickly select all files in a folder with the CTRL-A shortcut.

If you are importing CHP files, another Browse for Folder dialog box opens, asking you to select the appropriate folder for the results. See *Selecting CHP Data* on page 53.

If you are not importing CHP files, the loading progress bar displays the progress of the import. See *Viewing the Loading Progress* on page 55.

The selected ARR/XML/CEL/GQC files will be added to the data set only if they are:

- From the same array type as is used by the data set
- Not already in the data set

NOTE: When loading a large set of files, it is recommended that you use the "Select Directory" option, load all contained files, and then optionally remove undesired files after import.

NOTE: If you selected "Auto-QC Intensity Files" and the required library files are not found, a warning message will appear and all import actions will be aborted. See Chapter 4, *Library & Annotation Files* on page 34 on page for information on downloading and setting up the library path.

Selecting CHP Data

After selecting the intensity data directory or files for import, you will be prompted to select batch results folders for the following results files:

- Genotype analysis results files (.CHP)

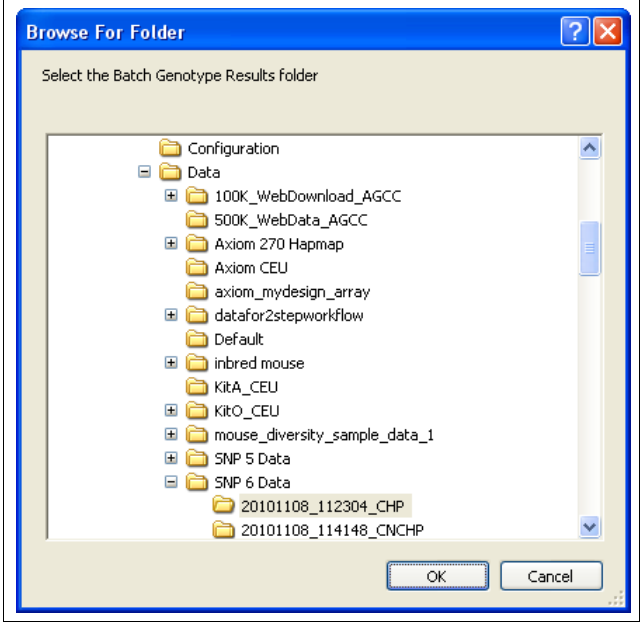
- Copy Number/Loss of Heterozygosity (CN/LOH) analysis results files (CNCHP and LOHCHP)
- Copy Number Variation (CNV) analysis results files (.CNVCHP)

Depending upon the array type, not all of these file types may be available.

After selecting intensity data for loading:

An appropriate Browse For Folder dialog box opens (Figure 5.12).

Figure 5.12 Browse For Folder dialog box

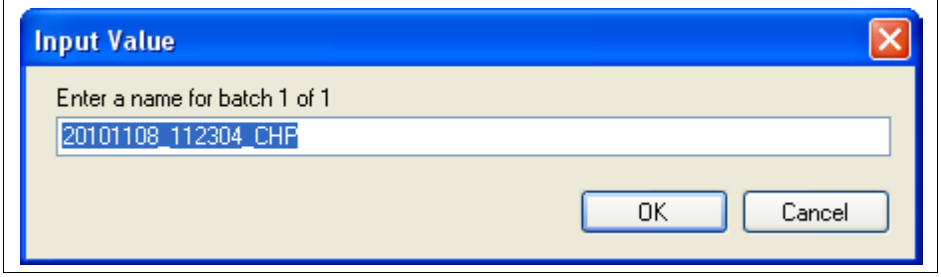


1. Browse to the folder containing the CHP files you wish to load and click **OK**.

   You do not have the option of selecting individual CHP files.

   Genotyping Console scans the set of CHP files in the selected folder (subfolders are ignored). If all the CHP files belong to the same batch analysis operation, and they belong to the same array used by the Data Set, then you will be asked to provide a name for the added Results Group (Figure 5.13).

Figure 5.13 Enter name



If the CHPs belong to multiple batch operations, Genotyping Console will import them as multiple Groups. You will be asked to provide a name for each Group.

NOTE: By default, the Genotype Results, Copy Number/LOH Results and Copy Number Variation Results Group names are based on the folder name. If you later rename a Results Group name, you will need to use the Windows files system to rename the actual folder if you wish them to continue to have the same name. You can view the actual folder names by using the file location features. See *Viewing the Location of Data Files* on page 58.

2. Click **OK**.

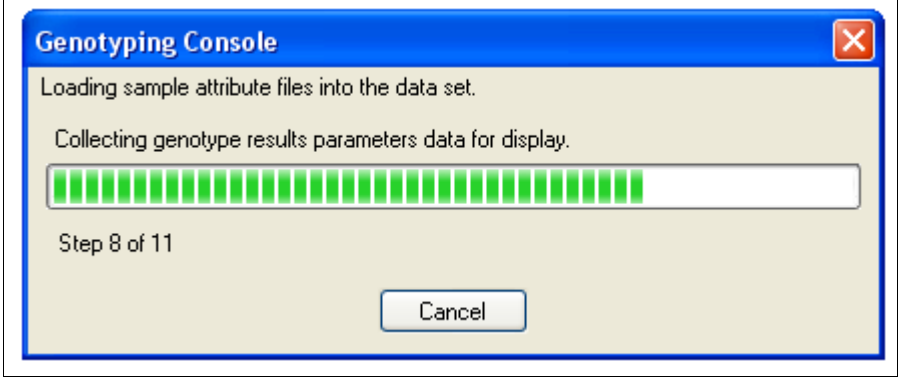
   If there are other types of results files available, the appropriate browse to window opens and you will be prompted for a batch results name after selecting the directory.

   If there are no other types of results files available, the data import starts and the Loading Progress dialog box is displayed (below).

Viewing the Loading Progress

The progress of loading the files into the data set is displayed in a dialog box with a progress bar (Figure 5.14).

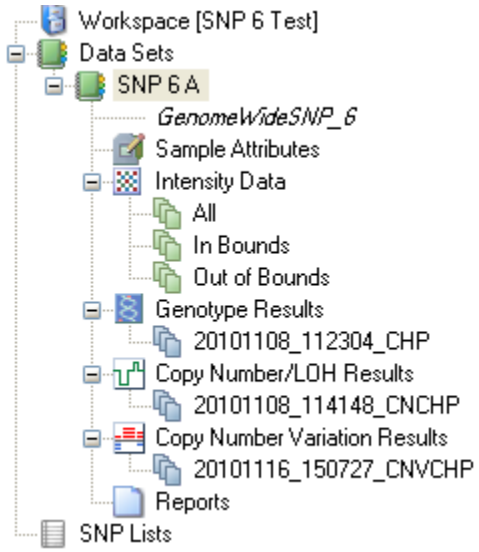
Figure 5.14 Loading files progress bar



When the loading is complete, the new data set is displayed in the data tree (Figure 5.15).

NOTE: Based on the type(s) of data added, the Sample Attribute Table, the Intensity QC Table, and/or CHP Summary Table will automatically open, displaying information about the existing and added files. The Status Message Pane will report any problems with the Add Data step.

Figure 5.15 Data tree with workspace and data set with data



Opening a Created Workspace File

There are two ways to open a workspace file that has been previously created:

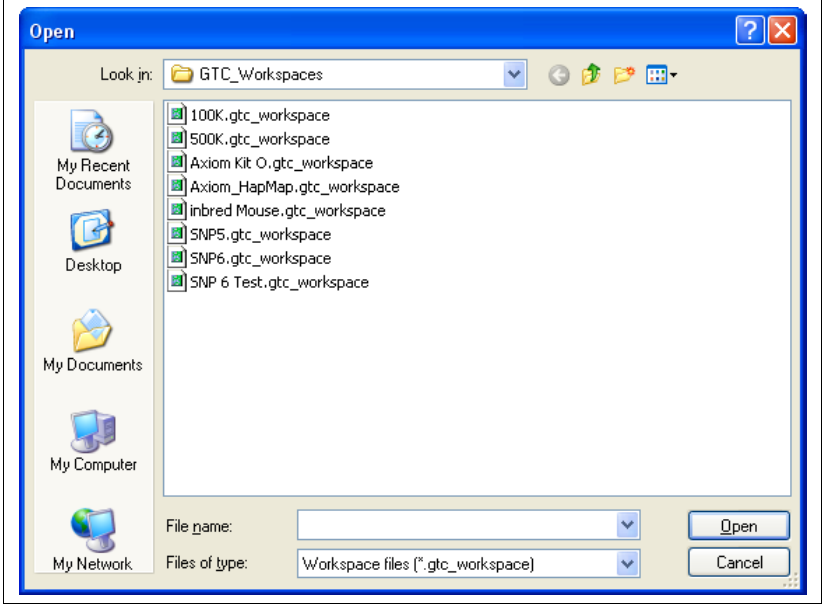
- In Windows Explorer, double-click the workspace file (.gtc_workspace). This will open the workspace in a new session of Genotyping Console.
- You can also open an existing workspace in Genotyping Console, if no workspace is currently open.

To open a workspace In Genotyping Console:

1. Do one of the following:
   - Select **File/Open Workspace**
   - Use the shortcut CTRL-O
   - Click the **Open Workspace** shortcut on the main tool bar

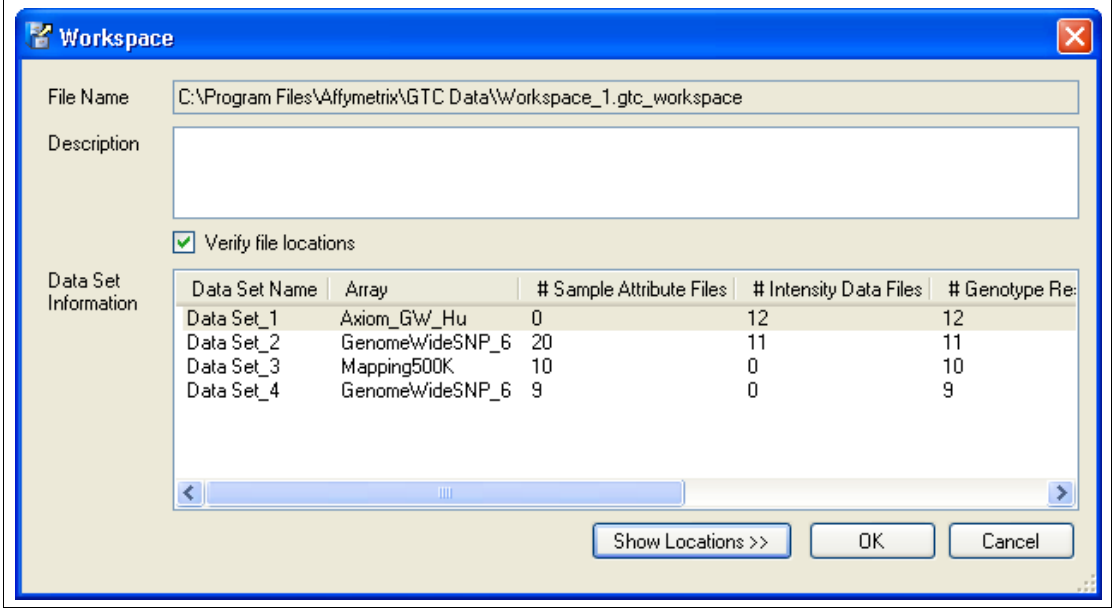
   The Open dialog box opens (Figure 5.16).

Figure 5.16 Open dialog box



2. Browse to the location of the workspace file, select the file, and click **Open**.
   The Workspace dialog box opens and displays the description and data set information (Figure 5.17).

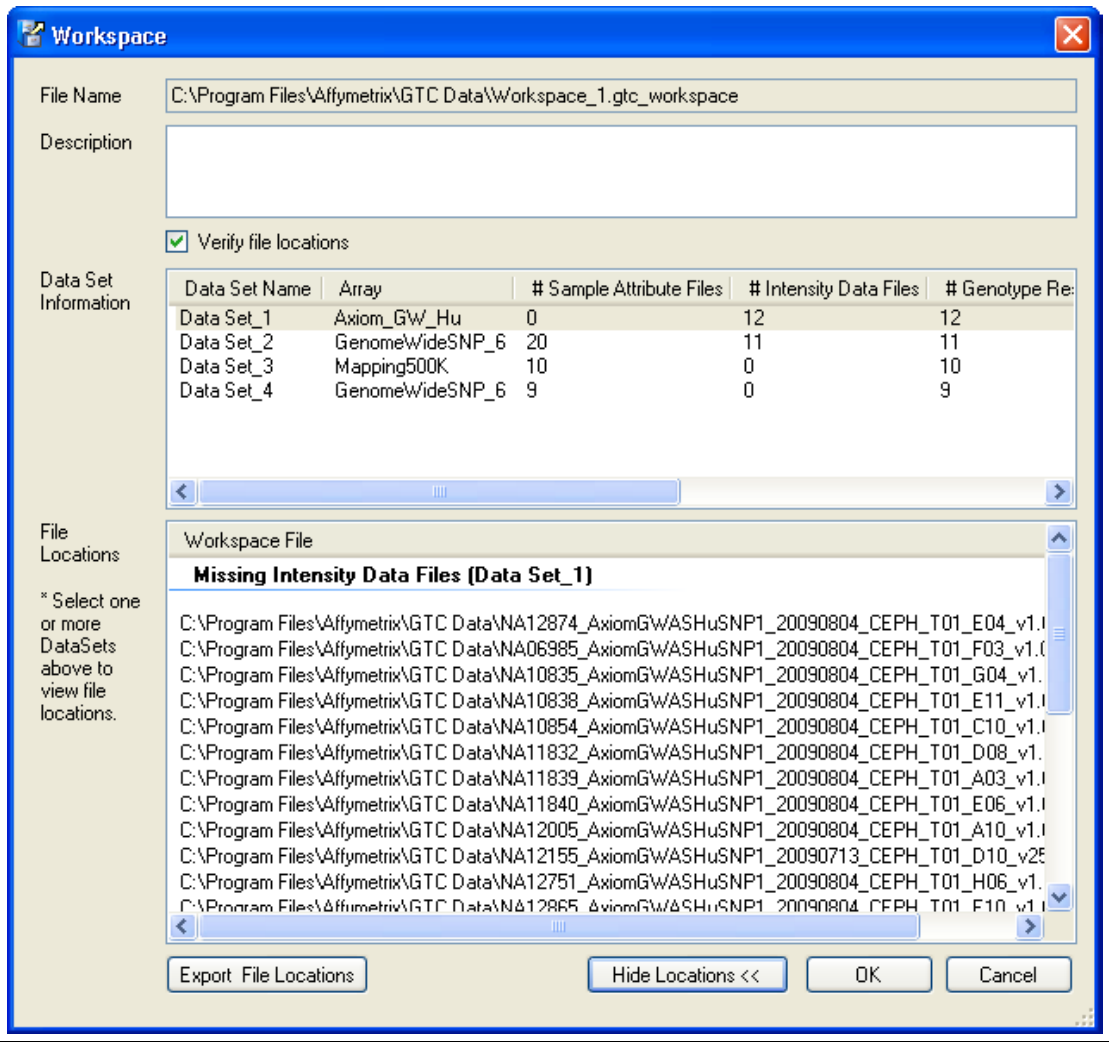
Figure 5.17 Workspace dialog box



   The Verify file locations option will confirm all data file locations upon opening the Workspace. If any files are missing or have been deleted, you will be prompted to either update the file paths or ignore the missing files. See *Locating Missing Data* on page 67 for more information.

   Click **Show Locations** to display the full path names of the data files (Figure 5.18).

Figure 5.18 Workspace file locations displayed

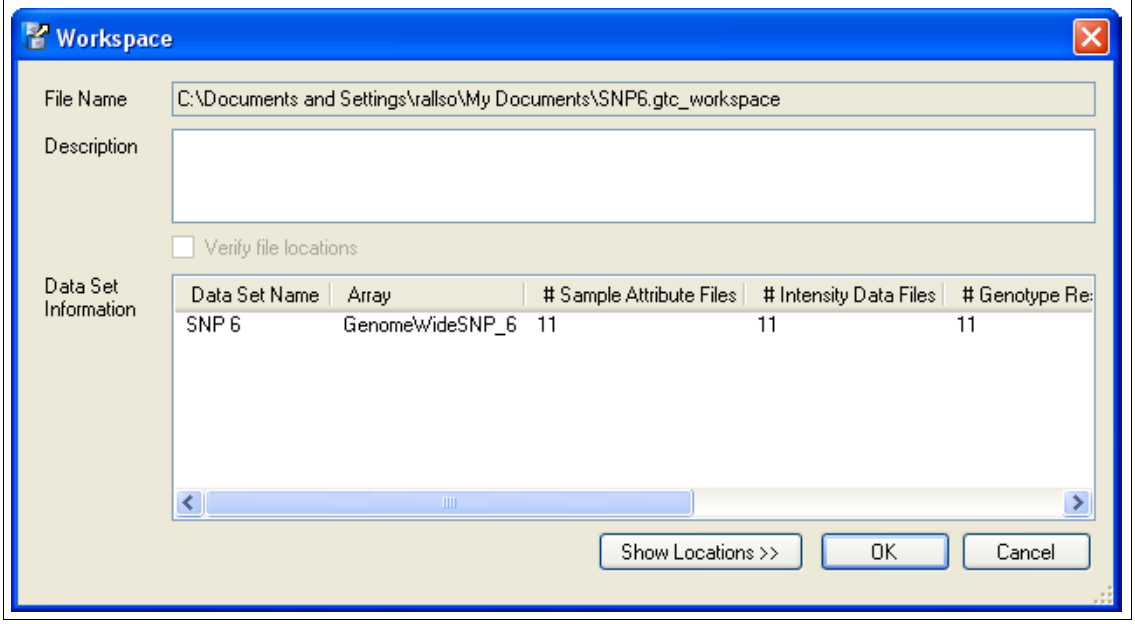


Viewing the Location of Data Files

You can view the location of the data files using one of the following methods:

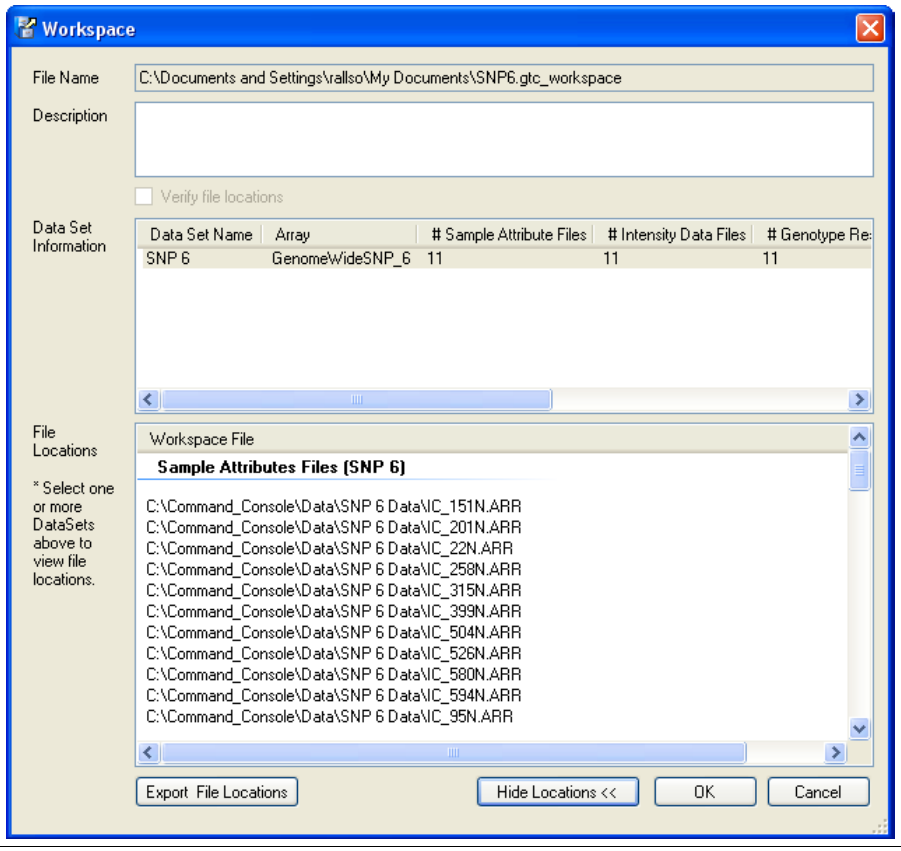
1. From the Workspace Menu, select **Properties > Show Information;** or
   Press Control + I.
   The Workspace dialog box opens (Figure 5.19).

Figure 5.19 Workspace dialog box



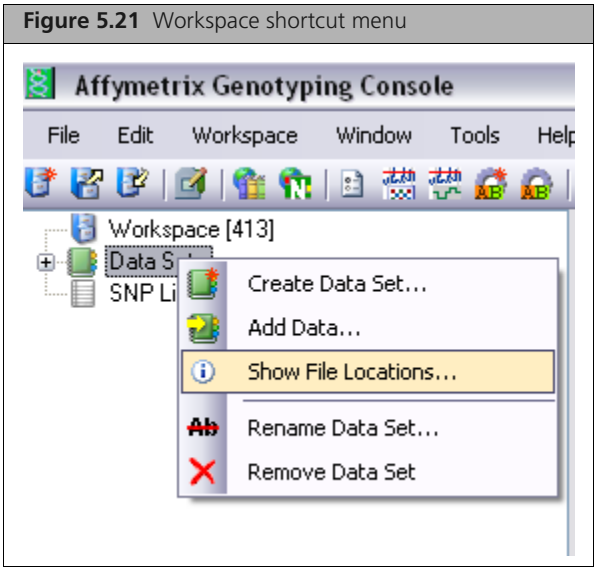
2. Click **Show Location**.
   The File Locations are displayed in the File Locations box (Figure 5.20).

Figure 5.20 File Locations displayed



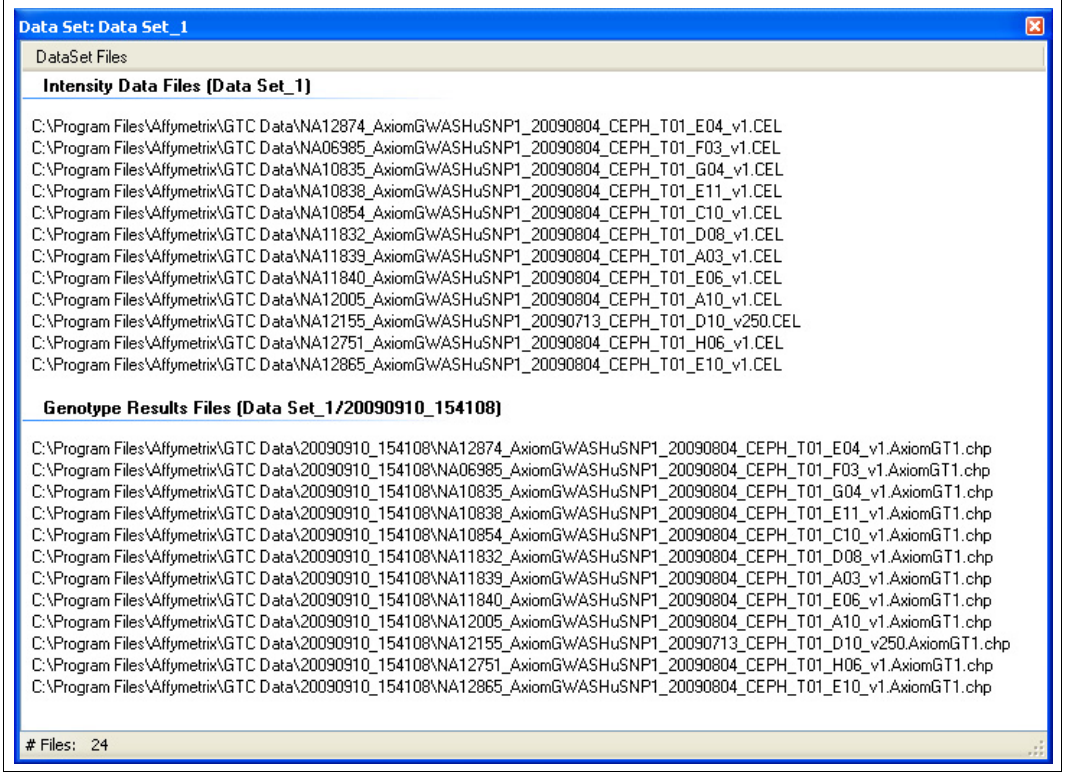
You can also right-click a data set in the directory tree and select **Show File Locations** on the shortcut menu (Figure 5.21).

Figure 5.21 Workspace shortcut menu



The Data Set window displays the file locations for the files in the workspace (Figure 5.21).

Figure 5.22 Locations of data set files



Removing Data from a Data Set

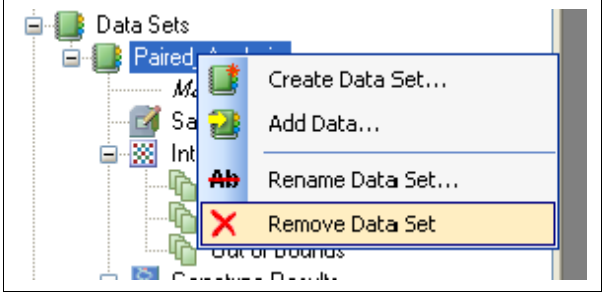
In Genotyping Console, data can be removed by either removing the entire Data Set or by removing subsets of files of a particular type of data (e.g. attribute (ARR/XML) files only, CEL intensity files only or CHP batch results).

To remove the entire Data Set:

- Right-click on a Data Set and select **Remove Data Set** (Figure 5.23).

  This will remove all data files for that Data Set from the Workspace.

Figure 5.23 Shortcut menu, Remove Data Set



NOTE: Removing all data or sub-sets of data from a workspace or data set does not delete the files from the file system, just the pointers to the data used by GTC.

Both individual as well as sets of data files can be removed from the Workspace in Genotyping Console. The following sections explain how to:

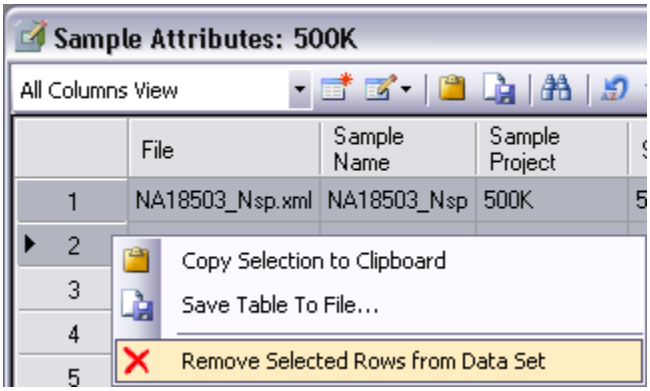
- *Remove Sample Files from a Data Set* on page 61
- *Remove Intensity Files from a Data Set* on page 62
- *Remove Genotyping, Copy Number/LOH or Copy Number Variation Results from a Data Set* on page 62

Remove Sample Files from a Data Set

To remove Sample (ARR/XML) files:

1. Open the Sample Attribute Table and highlight the rows (or ARR/XML files) to be removed.
2. Right-click and select **Remove Selected Data from Data Set** (Figure 5.24).

Figure 5.24 Remove sample files from a data set



The software prompts you to confirm the deletion. The highlighted rows (ARR/XML files) will be removed from the Data Set.

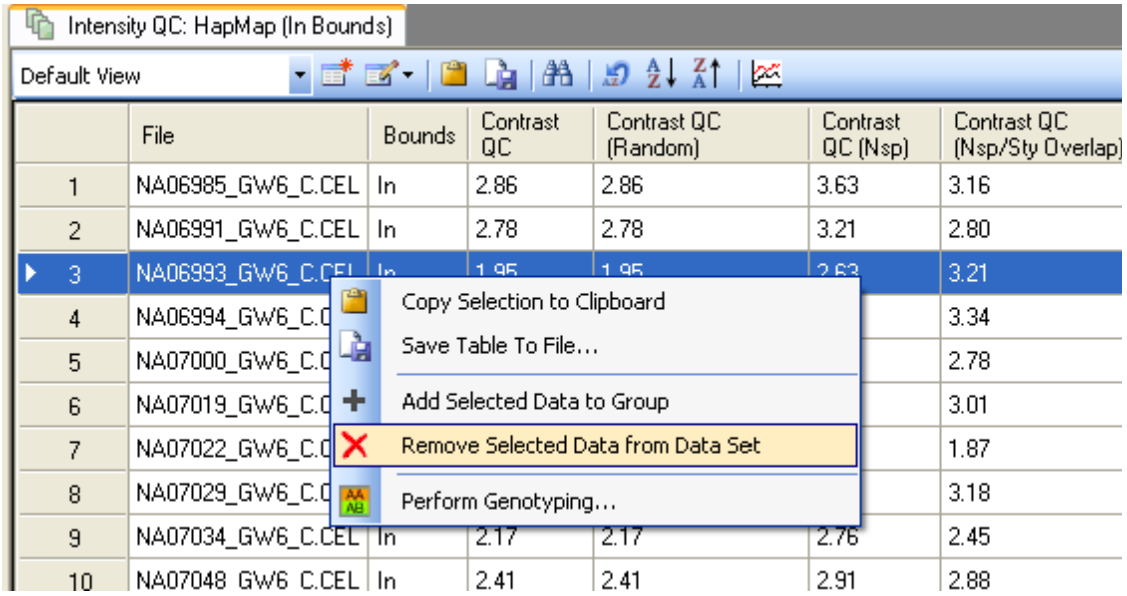
NOTE: If there are associated CEL and/or CHP files with these ARR files, they will not be removed from the Data Set.

Remove Intensity Files from a Data Set

To remove intensity (CEL) files:

1. Open the Intensity QC Table and highlight the rows (or CEL files) to be removed.
2. Right-click and select **Remove Selected Data from Data Set** (Figure 5.25).

Figure 5.25 Remove sample files from a data set



The software prompts you to confirm the deletion. The highlighted rows (CEL files) will be removed from the Data Set.

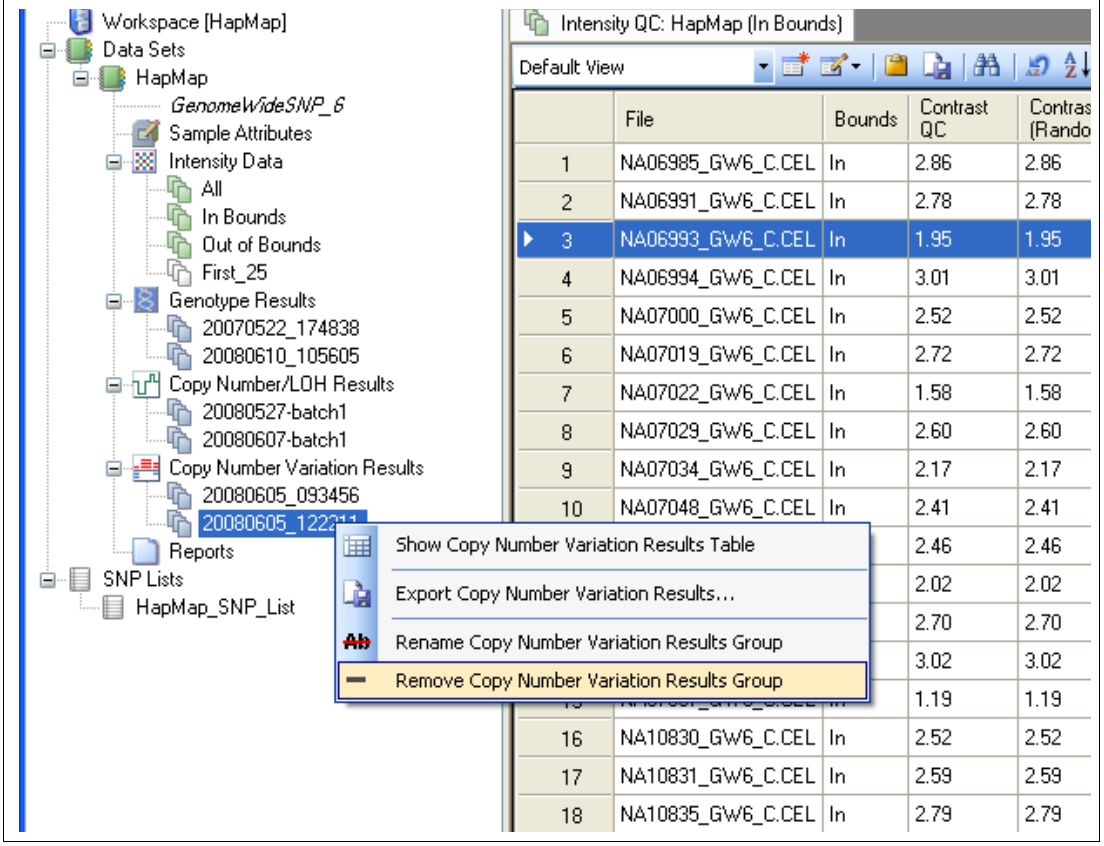
NOTE: If there are associated ARR/XML and/or CHP files with these CEL files, they will not be removed from the Data Set.

Remove Genotyping, Copy Number/LOH or Copy Number Variation Results from a Data Set

To remove Genotyping, Copy Number/LOH (CHP/CNCHP/LOHCHP) or Copy Number Variation (CNVCHP) files:

- Right-click on the batch of results and select **Remove Batch/Results** (Figure 5.26).

Figure 5.26 Remove Copy Number/LOH results group from a data set



The software prompts you to confirm the deletion.

NOTE: In Genotyping Console, individual CHP files cannot be removed; only entire batch results can be removed. If there are associated ARR/XML and/or CEL files with these CHP files, they will not be removed from the data set.

Sample Attributes Table

The Sample Attributes Table contains attribute information from the ARR/XML file. See *Table Features on page 198* for more information on customizing the table view.

The columns displayed will vary depending on whether this data was:

- Generated by AGCC
- Generated by GCOS
- Converted from GCOS to AGCC format

The attributes displayed in the table also depend upon the templates used in creating the sample file.

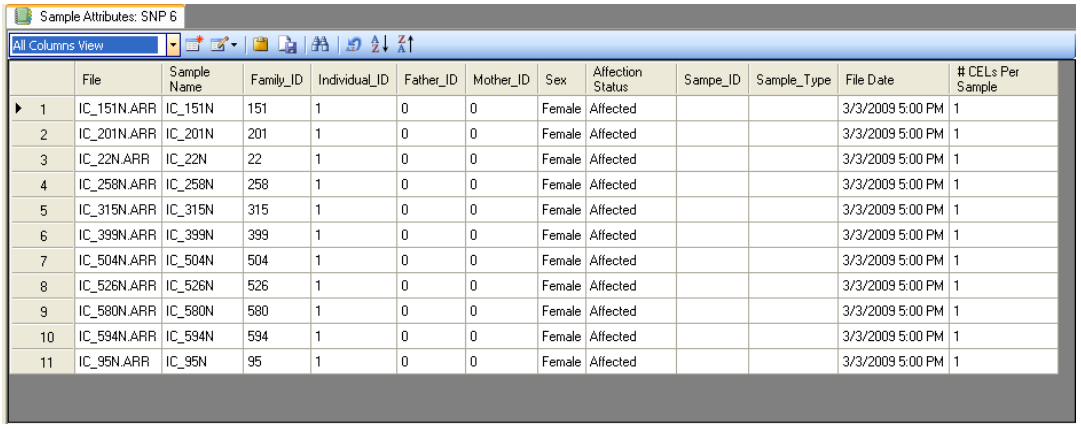
See the *Affymetrix® GeneChip® Command Console® User Manual* for more information on ARR files and attributes.

To open the Sample Attributes table:

- Double-click the Sample Attributes icon in the data tree. Alternately, from the Workspace Menu, select **Sample Attributes > Show Sample Attributes**.

  The Sample Attributes table displays the ARR/XML file information for the files in the Workspace (Figure 5.27).

Figure 5.27 Sample Attributes table for the SNP 6 data set



	File	Sample Name	Family_ID	Individual_ID	Father_ID	Mother_ID	Sex	Affection Status	Sampe_ID	Sample_Type	File Date	# CELs Per Sample
1	IC_151N.ARR	IC_151N	151	1	0	0	Female	Affected			3/3/2009 5:00 PM	1
2	IC_201N.ARR	IC_201N	201	1	0	0	Female	Affected			3/3/2009 5:00 PM	1
3	IC_22N.ARR	IC_22N	22	1	0	0	Female	Affected			3/3/2009 5:00 PM	1
4	IC_258N.ARR	IC_258N	258	1	0	0	Female	Affected			3/3/2009 5:00 PM	1
5	IC_315N.ARR	IC_315N	315	1	0	0	Female	Affected			3/3/2009 5:00 PM	1
6	IC_399N.ARR	IC_399N	399	1	0	0	Female	Affected			3/3/2009 5:00 PM	1
7	IC_504N.ARR	IC_504N	504	1	0	0	Female	Affected			3/3/2009 5:00 PM	1
8	IC_526N.ARR	IC_526N	526	1	0	0	Female	Affected			3/3/2009 5:00 PM	1
9	IC_580N.ARR	IC_580N	580	1	0	0	Female	Affected			3/3/2009 5:00 PM	1
10	IC_594N.ARR	IC_594N	594	1	0	0	Female	Affected			3/3/2009 5:00 PM	1
11	IC_95N.ARR	IC_95N	95	1	0	0	Female	Affected			3/3/2009 5:00 PM	1

By default, columns are displayed for every available attribute type in the ARR file.

See *Table Features* on page 198 for more information about customizing the displayed columns.

The file attributes listed in the table below (Table 5.3) are displayed in the table, as well as the attributes in the file.

Table 5.3 Sample Attribute Table columns

Column Name	Description
File	ARR/XML file name
# CELs Per Sample	Number of CEL files in this data set for the ARR/XML file
File Date	The date and time the ARR/XML files was last modified.

You can *edit sample attributes for AGCC sample files* (see below).

Editing Sample Attributes

Only AGCC sample files (ARR) can be edited in Genotyping Console software.

To make full use of the features in Genotype Console, data files should be in the Command Console format. Affymetrix provides the Data Exchange Console software (DEC) to convert your GCOS formatted data to Command Console format. The conversion to the new format will embed a unique file identifier that is used to track the relationship between ARR, CEL, and CHP files, removing the dependence on the file names to track relationships.

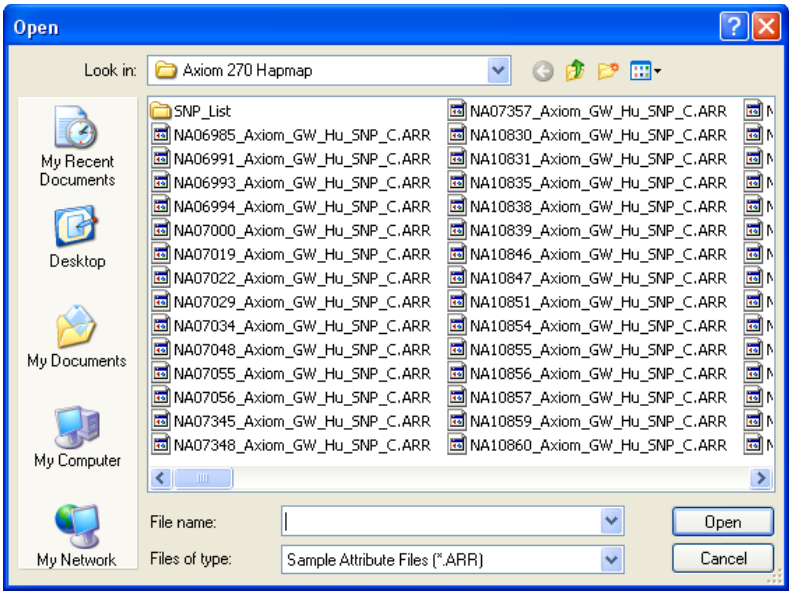
NOTE: The sample attributes contained in the XML files created by the Data Transfer Tool cannot be edited within Genotyping Console. If edits are needed, please edit the information in GCOS or GTYPE prior to using the Data Transfer Tool.

See the *Affymetrix® GeneChip® Command Console® User Manual* for more information on ARR files and attributes.

To edit an ARR file in Genotyping Console:

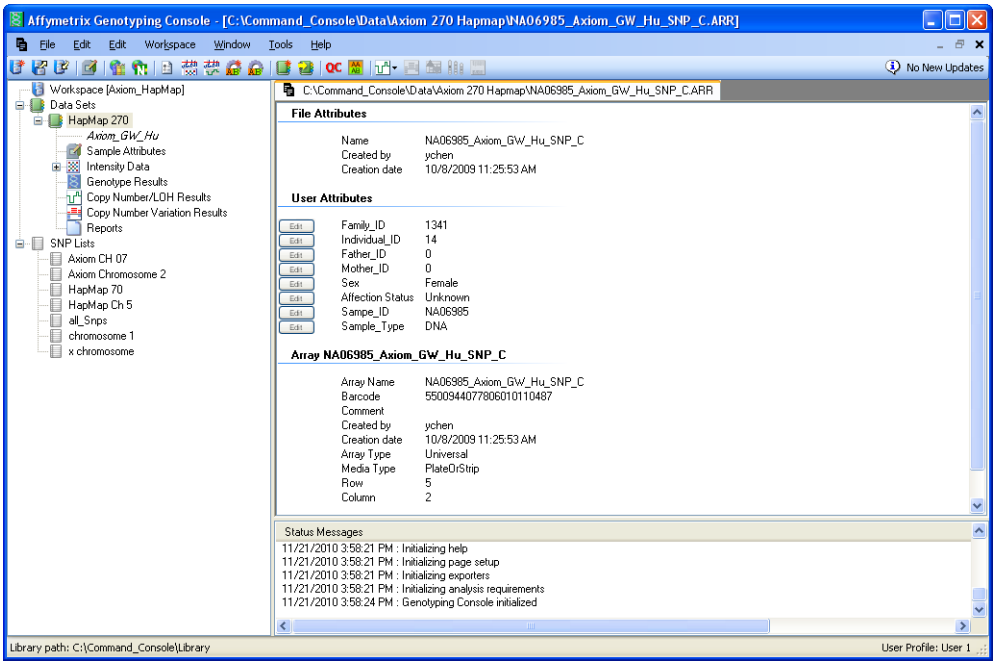
1. From the File menu, select **Open/Edit Sample File**.
   The Open dialog box opens (Figure 5.28).

Figure 5.28 Open dialog box



2. Browse to the directory that contains the ARR file to be edited.
3. Select the file and select **Open**.
   The Attribute Editor opens in Genotyping Console (Figure 5.29).

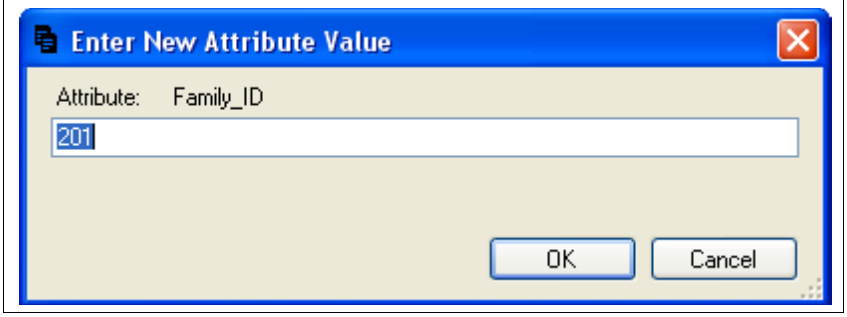
Figure 5.29 Attribute Editor



4. Select the attribute to edit (e.g. edit the gender).
   The appropriate Enter New Attribute Value dialog box opens (Figure 5.30, Figure 5.31, Figure 5.32).
5. Enter a new value for the attribute and click **OK**.

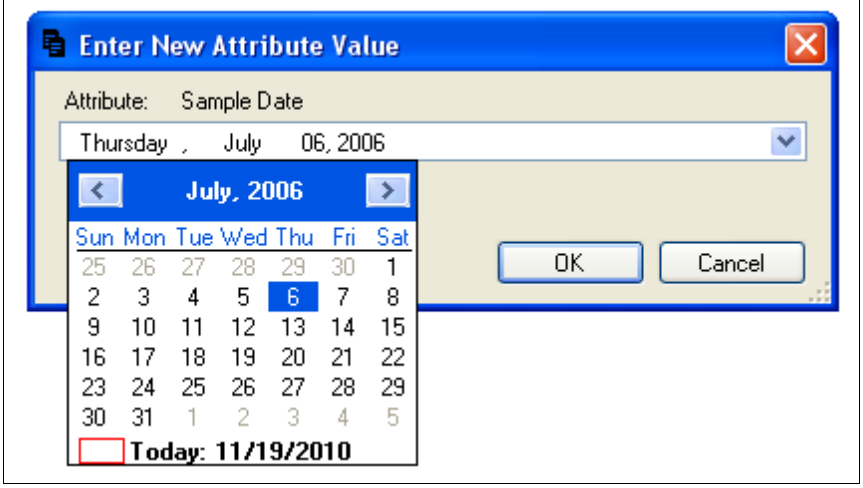
- If the attribute is a text attribute (Figure 5.30), type the value of the attribute in the Enter New Attribute Value window.

Figure 5.30 Enter New Attribute Value dialog box for single-select attribute



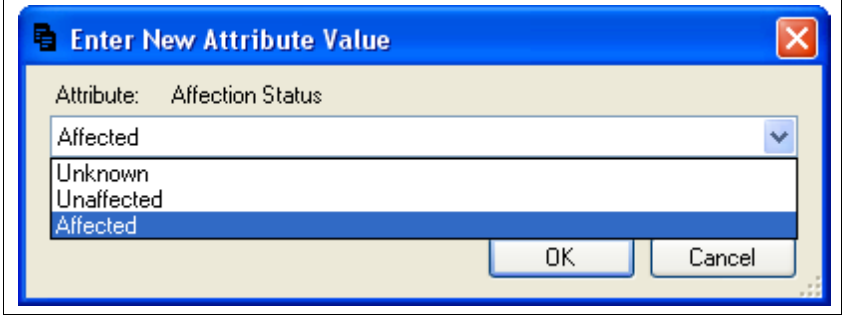
- If the attribute is a date (Figure 5.31), then select the correct date from the calendar.

Figure 5.31 Enter New Attribute Value dialog box for single-select attribute



- If the attribute is a single select attribute (enables you to select a single value from a controlled vocabulary list) (Figure 5.32), select the correct value from the pull-down menu.

Figure 5.32 Enter New Attribute Value dialog box for single-select attribute



Genotyping Console will prompt you to save the changes.

NOTE: Only one ARR file can be edited at a time. To batch edit ARR files, use the AGCC Portal.

NOTE: ARR files are updated by the attribute editor. If the ARR file is in a directory that is monitored by AGCC then changes made in Genotyping Console will also be reflected in AGCC.

Locating Missing Data

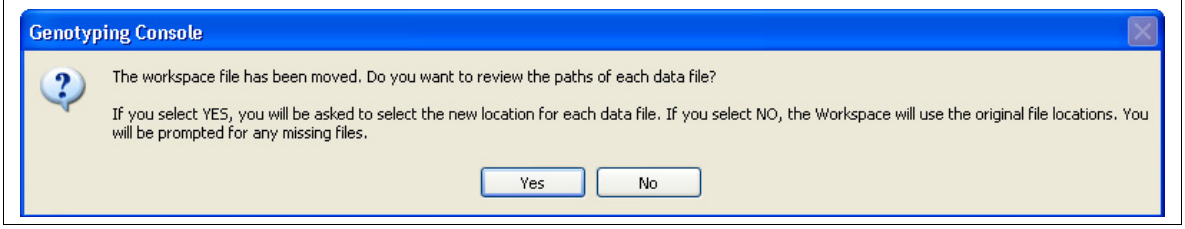
When opening a workspace, Genotyping Console software will confirm all of the locations for all files in the specified workspace as well as the workspace file itself.

If any file has been moved or deleted, including the workspace file, Genotyping Console software will prompt you to update the file locations or ignore the missing file(s).

If the workspace file has been moved:

1. The Workspace file has been moved dialog box appears (Figure 5.33).

Figure 5.33 Workspace file has been moved dialog box

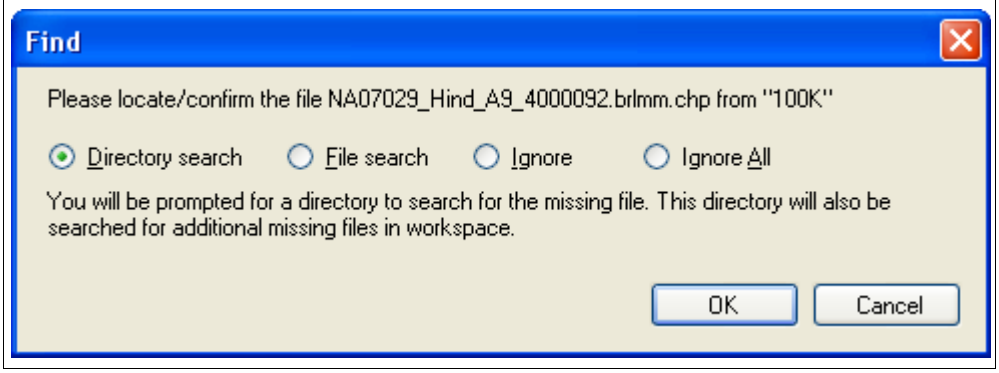


- If you click **Yes**, you will be asked to select the new location for each data file in the workspace. The Find File dialog box (Figure 5.34) opens for each data file in the workspace.
- If you click **No**, you will only be asked to select the location of missing files. If the data files haven't been moved, you won't be asked for locations.

If you are asked for the location of a missing data file:

1. The Find dialog box opens (Figure 5.34).

Figure 5.34 Find file dialog box



The Find dialog box options include:

- **Directory Search**: Locate the directory which contains this file
- **File search**: Locate the file itself
- **Ignore**: Ignore this file and open the workspace without it
- **Ignore All**: Ignore all missing files

2. Select the desired option and click **OK**.

   If the Ignore or Ignore All option is selected, the file(s) will be flagged as missing in the software until they are either deleted from the workspace or the path is corrected. This may result in data being missing from data tables.

   If you selected Directory Search, see *Directory Search* on page 68.

   If you selected File Search, see *File Search* on page 68.

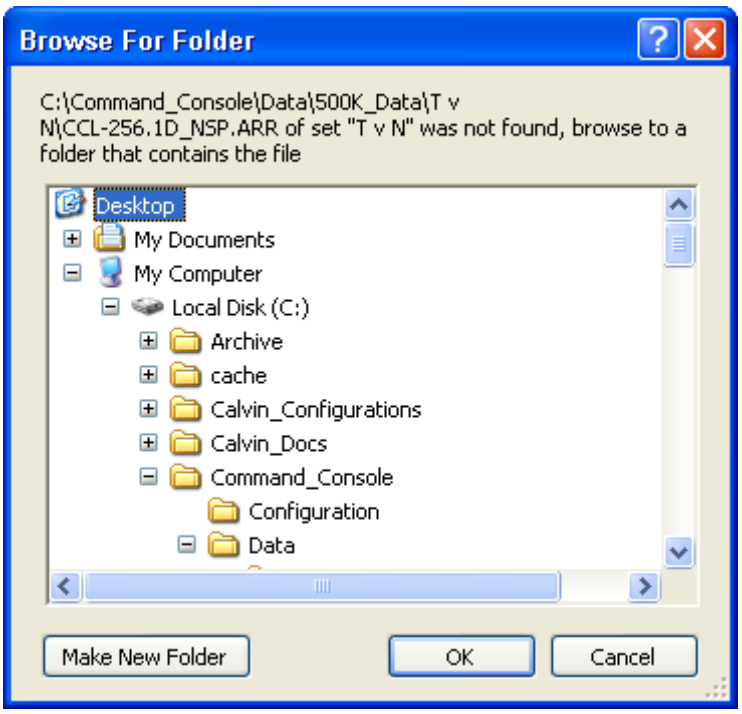
NOTE: If a workspace is already opened, go to Workspace/Verify File Locations to perform this check.

Directory Search

If the directory search option is chosen:

1. The Browse for Folder dialog box opens (Figure 5.35).

Figure 5.35 Browse For Folder dialog box



2. Browse to the folder containing the specified missing file and click **OK**.

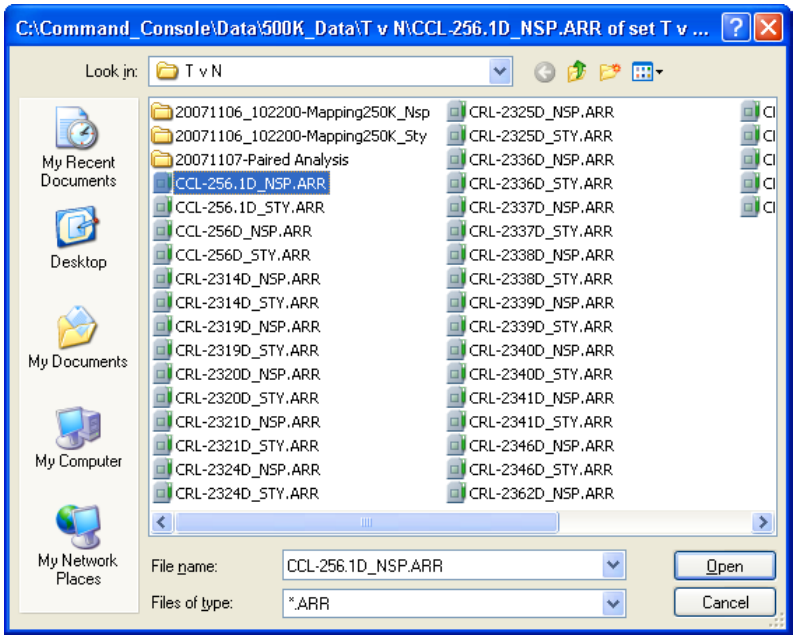
NOTE: Genotyping Console will look for the missing file in that directory. If there are additional files from the specified Workspace in this new directory, their paths will also be updated.

File Search

If the file search option is chosen:

1. The File Search dialog box opens (Figure 5.36).

Figure 5.36 File search dialog box



2. Browse to the correct folder and select the missing file.
3. Click **OK**.

NOTE: In the file search option, Genotyping Console will add the specified file to the Workspace. You will be prompted to locate each missing file.

Sharing Data

If multiple users in the same organization want to share the same workspace from different computers, you may decide to place the Workspace file in a shared folder. However, only one user can have the same workspace file open at a time. Also note that processing data and viewing some tables will be significantly faster if the data files are on the same computer as the Genotyping Console.

The Zip Workspace feature in GTC gathers all of the files in a selected workspace (as well as the workspace file) into a single package file. The package file can then be used to easily move the entire workspace from one location to another. The Zip Workspace feature will modify the data file locations in the workspace file when unpacking the file.

NOTE: Files not part of the workspace, such as Segment Summary reports and Custom Region Summary reports are not packaged as part of the zipped workspace. GTC 4.0 and higher versions cannot unzip workspace zip files > 4 GB that were created in earlier versions of GTC. However, GTC 4.2 can zip and unzip workspaces created within GTC 4.0 and higher versions with a zip file size > 4 GB.

Alternately, individual data files can be shared by simply copying the files to a new location and generating a new Workspace file.

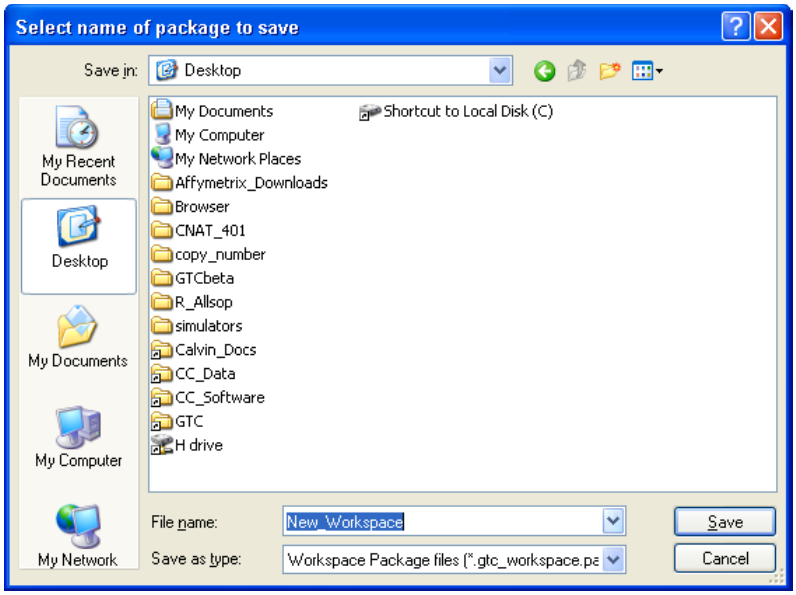
If you decide to simply move the data files and/or the Workspace file, Genotyping Console will ask you locate the missing files. See *Locating Missing Data* on page 67 for more information.

Using Zip Workspace

To zip a workspace:

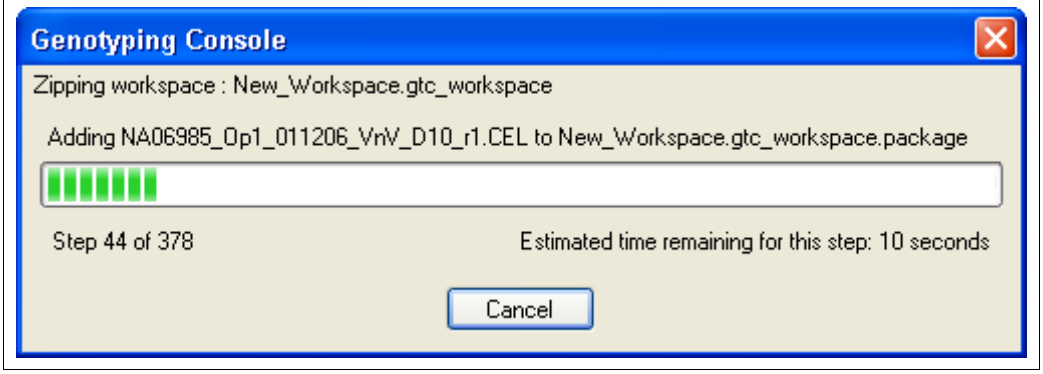
1. From the File menu, select **Zip Workspace**.
   The Select name of package to save dialog box opens (Figure 5.37).

Figure 5.37 Select name of package to save dialog box



2. Enter a name for the workspace you wish to save in the File name box.
3. Use the navigation tools in the dialog box tool bar to select a location for the packed workspace.
4. Click **Save**.
   The Workspace Zip progress indicator appears (Figure 5.38).

Figure 5.38 Workspace Zip Progress bar

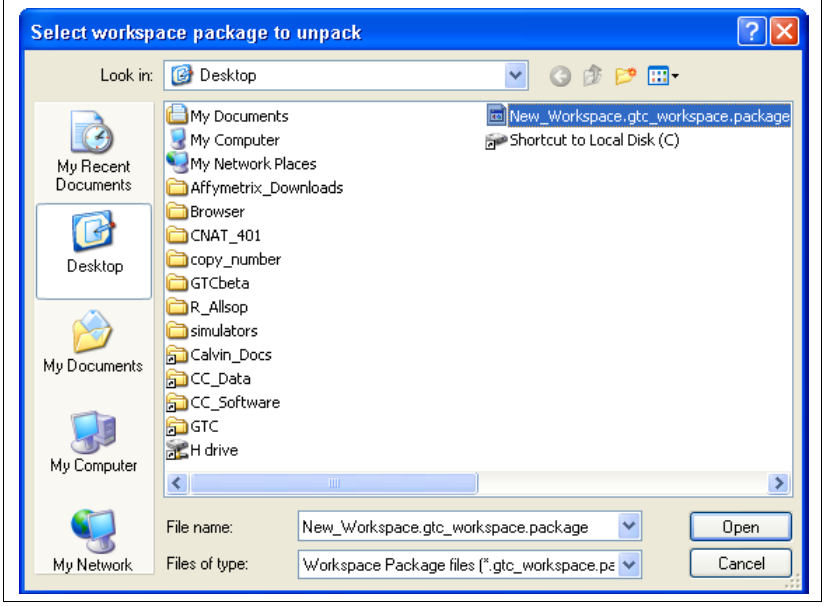


The progress indicator provides an estimate of the time needed to finish the packing. When packing is finished, the package appears in the location specified and can be archived or shared with another user.

To unzip a workspace package:

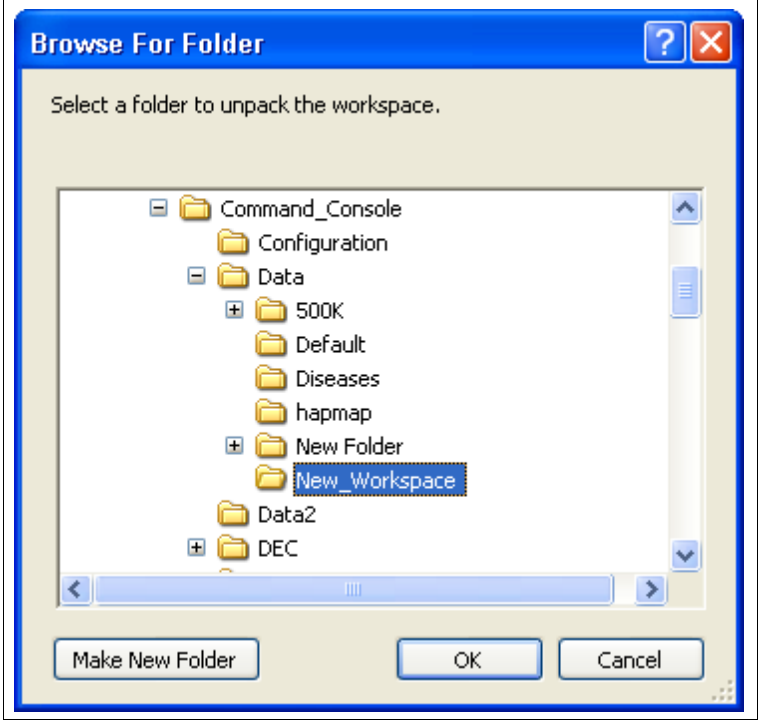
1. From the File menu, select **Unzip Workspace**.
   The Select Workspace Package to unpack dialog box opens (Figure 5.39).

Figure 5.39 Selecting a workspace package to unpack dialog box



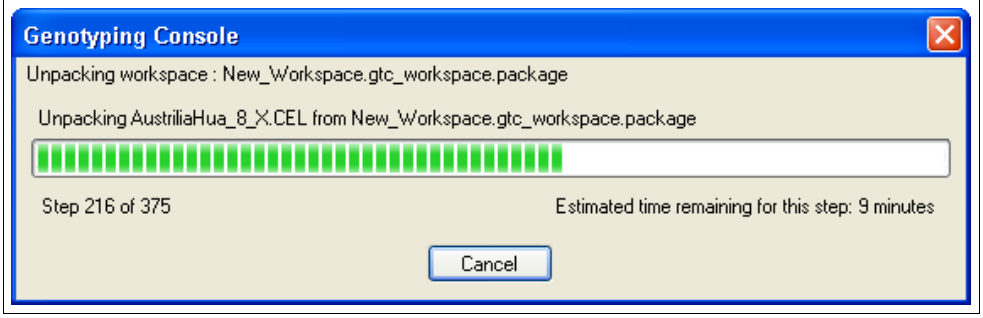
2. Select the workspace package you wish to unzip and click **Open**. The Unpack Location dialog box opens (Figure 5.40).

Figure 5.40 Select a folder to unpack the workspace



3. Browse to the folder where you wish to unzip the files in the workspace package and click **OK**. The Unpacking Progress indicator appears (Figure 5.41).

Figure 5.41 Unpacking progress indicator



When the unpacking operation is finished, the progress indicator disappears. You can now open the workspace in GTC.

Changing Zip Compression Level

You can change the settings for the zip operation to balance the time it takes to create a zip file and the size of the file.

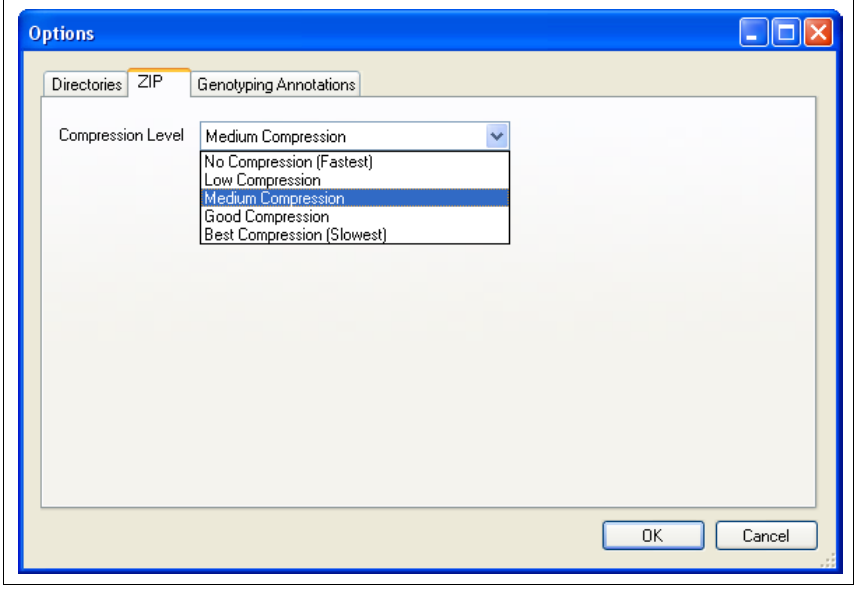
To change the zip operation settings:

1. Click the **Options** tool bar shortcut; or from the File menu, select **Option**.

   The Options dialog box appears.

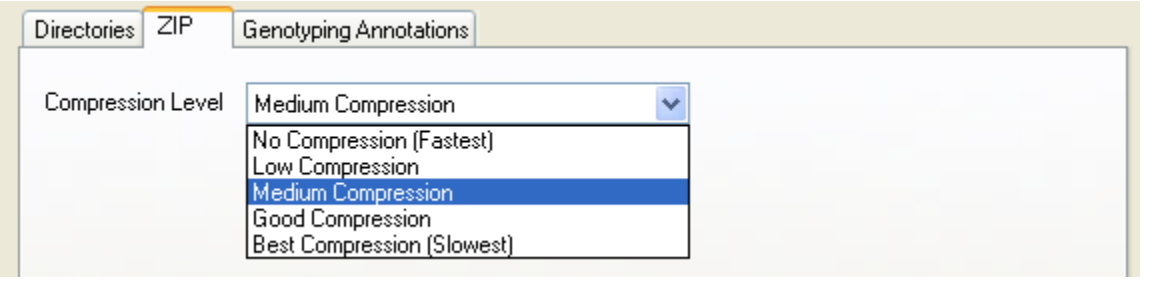
2. Click the ZIP tab (Figure 5.42).

Figure 5.42 Zip options



3. Select the Compression Level setting from the drop-down list (Figure 5.43).

Figure 5.43 Compression Level options



4. Click **OK** in the Options dialog box.
   The selected setting will be used for creating ZIP files.

Intensity Quality Control for Genotyping Analysis

Affymetrix has developed several control features to help researchers establish quality control processes for genotyping analyses. Researchers are encouraged to monitor these controls on a regular basis to assess assay data quality. These features include:

- Intensity QC Metrics
- Signature SNPs genotype calls

This chapter provides a description of the intensity QC features in the following sections:

NOTE: Intensity QC and Signature SNPs are not available for all array types.

The overall QC operations when performing Genotyping are described in *Genotyping QC Steps* on page 124.

Performing Intensity QC

The QC analysis provides an estimate of the overall quality for a sample based on the QC algorithm shown in the table below (Table 6.1). This analysis provides a quick preview of data quality prior to performing a full clustering analysis.

Table 6.1 Intensity QC information

Array	Number of SNPs used for QC	QC Algorithm
Human Mapping 100K Array: Mapping50K_Hind240 Mapping 50K_Xba240	 All All	Dynamic Model (DM) algorithm with QC Call Rate
Human Mapping 500K Array: Mapping250K_Nsp Mapping250K_Sty	 All All	Dynamic Model (DM) algorithm with QC Call Rate
Genome-Wide Human SNP Array 5.0	3022	Dynamic Model (DM) algorithm with QC Call rate
Genome-Wide Human SNP Array 6.0	3022	Contrast QC (CQC) is the primary QC method, as Dynamic Model (DM) algorithm was also used for QC
Axiom™ Human Arrays:	4070 non-polymorphic probes from 22 autosomal chromosomes	Dish QC (DQC) followed by measuring the genotype cluster call rate as generated during 2nd pass genotyping with the Axiom GT1 algorithm
Axiom™ Genome-Wide BOS 1 Array	5115 non-polymorphic probes from 29 autosomal chromosomes	Dish QC (DQC) followed by measuring the genotype cluster call rate as generated during 2nd pass genotyping with the Axiom GT1 algorithm

NOTE: Intensity QC is not available for Rat and Mouse arrays.

NOTE: GTC looks for existing QC information in the CEL file first, then a QC file (.gqc). If available, GTC uses this QC information and does not execute the QC algorithm. If the information is not available, GTC performs intensity QC and stores the information in the CEL file if it is an AGCC CEL file or in the gqc file, if it is a GCOS CEL file. However, it is required to perform intensity QC again for SNP 6.0 arrays with QC information generated in GTC 2.0 due to a QC algorithm update since GTC 2.1.

Only samples that meet QC thresholds should be genotyped.

NOTE: It is recommended that samples not meeting the QC thresholds be re-hybridized or rescanned.

NOTE: The intensity QC metric is well-correlated with clustering performance and is an effective single-sample metric for deciding what samples should be used in downstream clustering. However the correlation between the metric and genotyping performance is not perfect and there will occasionally be a sample that passes the metric but which has sub-optimal genotyping performance. See the following sections for recommendations on additional per-sample QC to perform after the clustering analysis.

- ***Two-Step Genotyping Workflow*** **on page 123**
- **Chapter 9, *Using the SNP Cluster Graph* on page 148**

NOTE: The majority of the time Genome-Wide Human SNP Array 5.0 samples that meet the default QC Call Rate criteria will have a BRLMM-P genotyping call rate of at least 96% and an accuracy of at least 99% (with average performance significantly higher) when analyzed with Genotyping Console at default settings.

NOTE: The majority of the time Genome-Wide Human SNP Array 6.0 samples that meet the default Contrast QC criteria will have a Birdseed genotyping call rate of at least 97% and an accuracy of at least 99% (with average performance significantly higher) when analyzed with Genotyping Console at default settings.

NOTE: You will need to run QC on Axiom CEL files even if they have already been QCed in GTC 4.0. GTC 4.2 provides information on the reagent version used to process the arrays, and this information is required to perform Genotyping analysis on Axiom data in GTC 4.2.

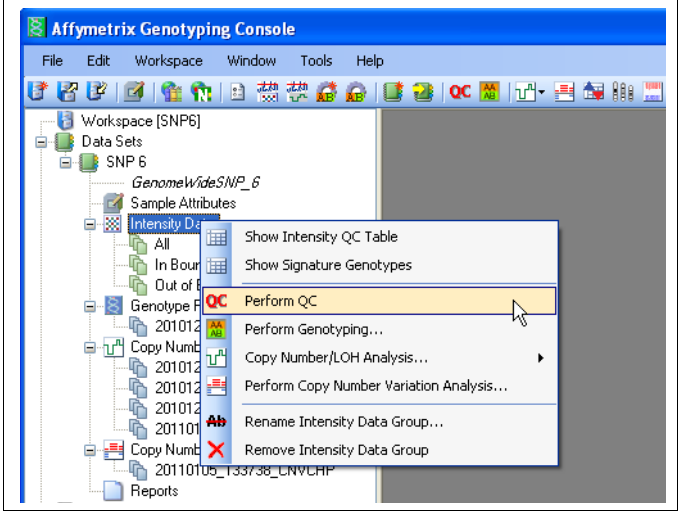
QC can be automatically initiated upon import of CEL files by selecting the Auto-QC Intensity Files option. See *Adding Data to a Data Set* on page 49.

Gender analysis is also performed during the QC step. It provides a gender call that will be used to select models for the X and Y chromosomes during genotyping. Different processes are used for the gender call, depending upon the type of array being analyzed. See Appendix E, *Gender Calling in GTC* on page 353 for more details.

To initiate QC on CEL files already in the workspace:

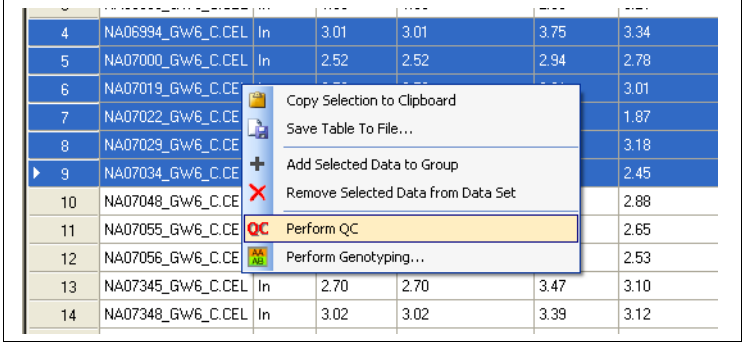
1. Do one of the following:
   - Select an intensity group from the tree (e.g. All) (Figure 6.1).

Figure 6.1 Starting QC from Data Tree right-click menu



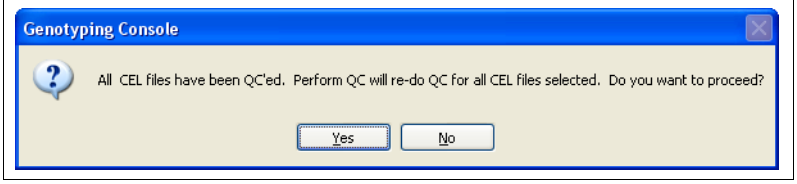
   - Select row(s) from an open Intensity QC table (Figure 6.2).

Figure 6.2 Starting QC from Intensity QC table



2. Right-click and select **Perform QC**.
   If you have already performed QC on the selected data, the following notice appears (Figure 6.3):

Figure 6.3 Redo QC notice



   Click **Yes** to proceed with the QC.
   When the QC is completed the results will automatically be displayed in the Intensity QC table (see *Intensity QC Tables* on page 82).

The Results will automatically be parsed into 3 groups:

- "All" group contains results for all Intensity files in the Data Set (both newly added and existing files).
- "In Bounds" group contains the results for Intensity files which pass the QC Threshold(s).
- "Out of Bounds" group contains the results for all Intensity files which do not meet the QC Threshold(s).

NOTE: After performing QC on Axiom CEL files, you will need to create intensity data groups to group the data from arrays processed with Reagent Version 1 and Reagent Version 2 into separate groups. See *Creating Custom Intensity Data Groups using Intensity QC Data* on page 84 for more information.

By default, the In and Out of Bounds grouping is based upon the default QC parameter for the array type (Table 6.2):

Table 6.2 QC Parameters for different Array types

Array	QC Parameter
Human Mapping 100K Array: Human Mapping 500K Array: Genome-Wide Human SNP Array 5.0	*QC Call Rate* on page 81
Genome-Wide Human SNP Array 6.0	*Contrast QC* on page 81
Axiom Genotyping Array plates, including:	*DISH QC* on page 81

To modify, see *Modifying QC Thresholds* on page 79.

To view the QC results for all data in the data set, open the Intensity QC table for all data. Out of bounds samples will be flagged with a red highlight in the QC Call Rate column (Figure 6.4).

Figure 6.4 Intensity QC table showing out-of-bounds data (Dish QC threshold has been changed from default)

Intensity QC: HapMap 270 (Kit A)

kit View

	File	Bounds	AT Channel FLD	GC Channel FLD	Dish QC	Log Difference QC	Computed
1	NA06985_Axiom_GW_Hu_SNP_C.CEL	In	4.21382	4.33465	0.90659	3.39584	female
2	NA06994_Axiom_GW_Hu_SNP_C.CEL	In	4.72934	4.80840	0.94297	3.69570	male
3	NA07022_Axiom_GW_Hu_SNP_C.CEL	In	4.88362	4.71473	0.94051	3.66321	male
4	NA07048_Axiom_GW_Hu_SNP_C.CEL	In	5.17923	5.07099	0.93904	3.66582	male
5	NA07345_Axiom_GW_Hu_SNP_C.CEL	In	4.67252	4.25828	0.90708	3.38619	female
6	NA10830_Axiom_GW_Hu_SNP_C.CEL	In	3.83475	4.43015	0.90610	3.45658	male
7	NA10838_Axiom_GW_Hu_SNP_C.CEL	In	4.73959	4.93989	0.94690	3.72962	male
8	NA10847_Axiom_GW_Hu_SNP_C.CEL	Out	4.65656	4.03955	0.88987	3.26868	female
9	NA10855_Axiom_GW_Hu_SNP_C.CEL	In	4.68222	4.61678	0.92625	3.55216	female
10	NA10859_Axiom_GW_Hu_SNP_C.CEL	In	4.76944	4.55072	0.91691	3.47630	female
11	NA10863_Axiom_GW_Hu_SNP_C.CEL	In	4.44491	4.23842	0.91200	3.43266	female
12	NA11831_Axiom_GW_Hu_SNP_C.CEL	In	5.10838	4.65681	0.93854	3.64386	male
13	NA11840_Axiom_GW_Hu_SNP_C.CEL	Out	4.28273	4.09522	0.89381	3.35484	female
14	NA11992_Axiom_GW_Hu_SNP_C.CEL	In	5.28089	4.81950	0.94838	3.76597	male
15	NA11995_Axiom_GW_Hu_SNP_C.CEL	In	4.41564	4.37435	0.92035	3.47630	female
16	NA12005_Axiom_GW_Hu_SNP_C.CEL	In	5.15016	5.23840	0.94985	3.79306	male
17	NA12044_Axiom_GW_Hu_SNP_C.CEL	In	4.88465	4.83420	0.94395	3.71430	female
18	NA12144_Axiom_GW_Hu_SNP_C.CEL	In	4.67015	4.99229	0.93904	3.63324	male
19	NA12154_Axiom_GW_Hu_SNP_C.CEL	In	5.29459	5.16934	0.94936	3.74280	male
20	NA12234_Axiom_GW_Hu_SNP_C.CEL	Out	4.49268	3.58314	0.87365	3.19153	female
21	NA12248_Axiom_GW_Hu_SNP_C.CEL	In	4.30922	4.43060	0.93215	3.58412	male

For more information, see:

- *Intensity QC Table for Axiom™ Data (default view)* on page 82
- *Intensity QC Table for SNP 6.0 Data (default view)* on page 83
- *Intensity QC Table for Human Mapping 100K/500K & SNP 5.0 Data (default view)* on page 84

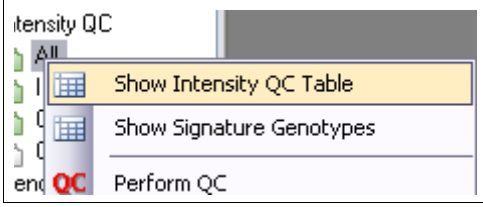
For more information on displaying data in the Intensity QC Table see Chapter 11, *Table & Graph Features* on page 198.

NOTE: For faster performance, Affymetrix recommends performing QC analysis with all files stored locally.

To review the QC Results at any time:

- Right-click on an Intensity Group and select Show Intensity QC Table (Figure 6.5) or double click on an Intensity group in the data tree.

Figure 6.5 Opening the Intensity QC Table



The *Intensity QC Tables* on page 82 contains the QC results. If the QC step is skipped, some or all of the Intensity files may have no QC results (the GQC file is missing or not updated with Contrast QC values, or the QC information is missing from the CEL file). If no intensity files in the data set have been QCed, the QC metrics columns will not appear in the Intensity QC table.

NOTE: The Contrast QC metric, the default metric for the Genome-Wide Human SNP Array 6.0, is not present in GQC files generated in GTC 2.0 software. SNP Array 6.0 data generated in GTC 2.0 will need to be re-QCed to generate the Contrast QC data. See *Genotyping QC Steps* on page 124 for more information on running the QC step.

QC Call Rate data will also be (re)generated during the QC step and available in the All Columns View, or by making a custom view. See *Table Features* on page 198 for more information on customizing the table view. Choosing All Columns View displays all data columns.

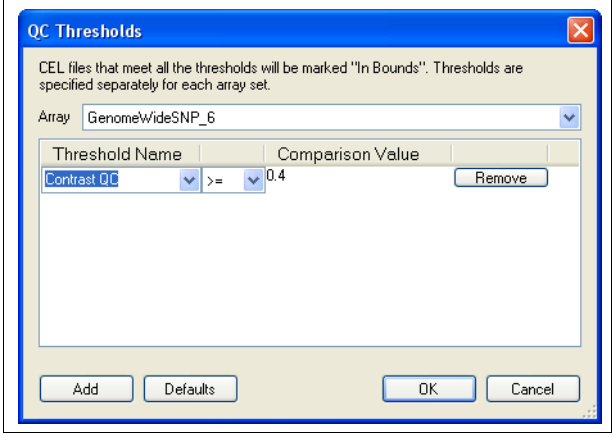
Modifying QC Thresholds

Genotyping Console maintains default thresholds for QC metrics, and will highlight in the Intensity QC tables the metrics which are outside of the threshold values. You can modify the QC thresholds as needed.

To modify the QC threshold options:

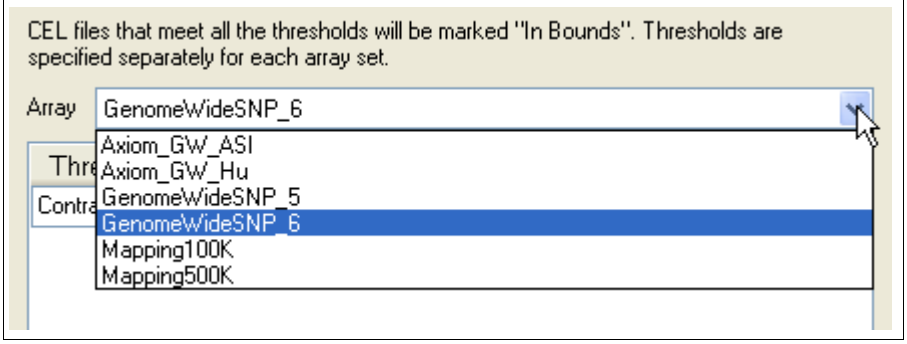
1. Click on the **QC Thresholds** button on the main tool bar: or
   From the Edit menu, select **QC Thresholds**.
   The QC Thresholds dialog box opens (Figure 6.6).

Figure 6.6 QC Thresholds dialog box



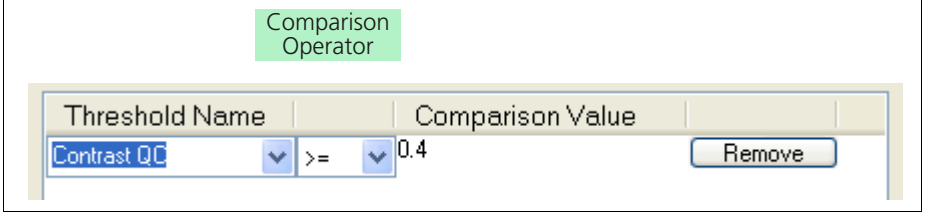
2. Select the array type to be modified (Figure 6.7).

Figure 6.7 Selecting array type



3. Select the metric, the comparison operator (less than (<), less than or equal to (≤), greater than (>), greater than or equal to (≥), equal to (=), or not equal to (!=)), and the value (Figure 6.8).

Figure 6.8 Changing the threshold



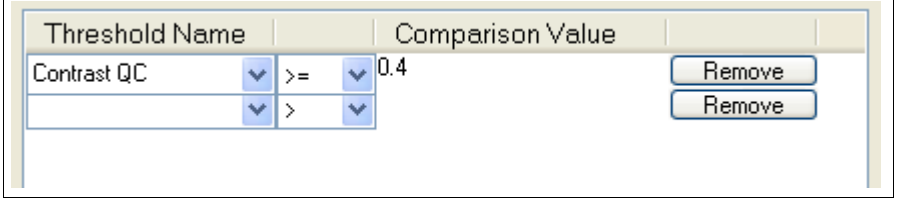
To use a different metric, select the text in the "Threshold Name" cell and type the exact name, case-sensitively, of the new metric in this field. For metrics to be applied, they must exist in the *Intensity QC Tables* on page 82 (All Columns view).

4. Enter a new Threshold Name if desired:

   A. Click **Add**.

   A new row appears in the dialog box (Figure 6.9).

Figure 6.9 New Row added



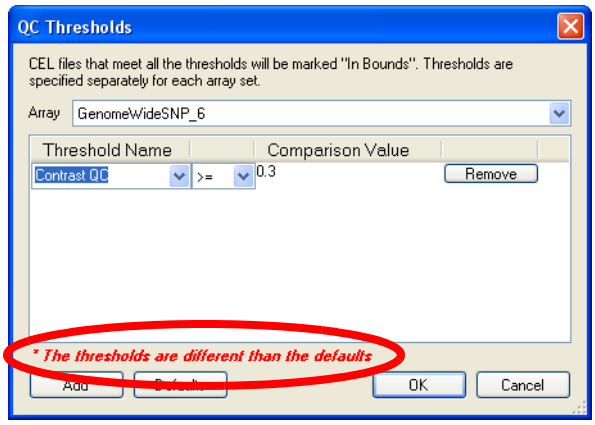
   B. Enter a new threshold name. The metric must exist in the *Intensity QC Tables* on page 82 (All Columns view).

   C. Select a comparison operator.

   D. Enter the comparison value.

5. To delete a threshold item, click **Remove**.
6. Click **OK** in the QC Thresholds dialog box when you have finished editing the thresholds (Figure 6.10).

Figure 6.10 Flag in the QC Thresholds dialog box indicating that the QC thresholds have been changed from the default values



QC Call Rate

QC call rate is the recommended QC metric for:

- Genome-wide SNP Array 5.0
- Human Mapping 100K Array
- Human Mapping 500K Array

NOTE: The QC Call Rate threshold has a default value for each array type. If you adjust this value or add additional metrics to threshold by, a flag will indicate that the thresholds are different from the defaults (Figure 6.10).

Contrast QC

NOTE: Contrast QC is the recommended QC metric for the Genome-Wide Human SNP 6.0 array. The default threshold is >= 0.4 for each sample. If you adjust this value or change the SNP 6.0 QC threshold settings to another metric such as QC Call Rate, or add additional metrics to threshold by, a flag will indicate that the thresholds are different from the defaults.

Contrast QC is a metric that captures the ability of an experiment to resolve SNP signals into three genotype clusters. It uses 10,000 random SNP 6.0 SNPs. See Appendix F, *Contrast QC for SNP 6.0 Intensity Data* on page 358 and Appendix G, *Best Practices SNP 6.0 Analysis Workflow* on page 360 for more details.

DISH QC

Dish QC (DQC) is the recommended Genotyping Console QC metric for the Axiom™ Genome-Wide Array Plates and Axiom myDesign Array Plates in Genotyping Console. The default threshold is greater than or equal to 0.82 for each sample. For bovine samples, the threshold is 0.95. It operates by measuring signal at a collection of sites in the genome that are known not to vary from one individual to the next. Because it monitors non-polymorphic locations, it is known at each position which of the two channels in the assay should contain signal and which channel should be just background. DQC is a measure of the extent to which the distribution of signal values separate from background values, with 0 indicating no separation and 1 indicating perfect separation.

DQC is a useful single-sample metric of performance and, under normal circumstances, it correlates well with genotyping performance. One exception is the case of sample mixing. A sample consisting of different individuals mixed together can still have a good DQC score, since the signals at non-polymorphic locations will remain the same in a mixture. Such samples can generally be identified by having abnormally low genotyping call rates, though they may still have good DQC values.

Intensity QC Tables

The intensity QC table displays different data for different array types:

- *Intensity QC Table for Axiom™ Data (default view)* on page 82
- *Intensity QC Table for SNP 6.0 Data (default view)* on page 83
- *Intensity QC Table for Human Mapping 100K/500K & SNP 5.0 Data (default view)* on page 84

Intensity QC Table for Axiom™ Data (default view)

The following information can be displayed for Axiom data after running QC in Genotyping Console (Table 6.3).

Table 6.3 Intensity QC metrics for Axiom data

Column Name	Description
File	CEL file name.
Bounds	In Bounds/Out of Bounds indicates whether the CEL file met the specified QC threshold(s).
Reagent Version	The reagent version used for processing the arrays, based on data intensity values. You can only perform batch genotyping analysis on CEL files processed using the same reagent version. You can create *custom intensity data groups* (page 84) to group CEL files processed using the same reagent version before genotyping analysis.
Dish QC	A QC metric that evaluates the overlap between the two homozygous peaks (AT versus GC) using normalized intensities of control non-polymorphic probes from both channels. It is defined as the fraction of AT probes not within two standard deviations of the GC probes in the contrast space.
Log Difference QC	A cross channel QC metric, defined as mean(log(AT_SBR))/std(log(AT_SBR)) + mean(log(GC_SBR))/std(log(GC_SBR)), where signal and background are calculated for control non-polymorphic probes after intensity normalization.
saturation_GC	Fraction of features in the GC channel with intensity greater than or equal to 3800. **IMPORTANT:** This metric/column does not appear for CEL files QC'd with GTC v.4.1 (or earlier).
saturation_AT	Fraction of features in the AT channel with intensity greater than or equal to 3800. **IMPORTANT:** This metric/column does not appear for CEL files QC'd with GTC v.4.1 (or earlier).
AT Channel FLD	Linear Discriminant for signal and background in the AT channel, defined as (median_of_GC_probe_intensities – median_of_AT_probe_intensities)^2 / [0.5 * (Axiom_signal_contrast_AT_B_IQR^2 + Axiom_signal_contrast_AT_S_IQR^2)].
GC Channel FLD	Linear Discriminant for signal and background in the GC channel, defined as (median_of_GC_probe_intensities – median_of_AT_probe_intensities)^2 / [0.5 * (Axiom_signal_contrast_GC_B_IQR^2 + Axiom_signal_contrast_GC_S_IQR^2)].
Computed Gender	Computed gender of organism sample was taken from (see Appendix H, *Best Practices Axiom Analysis Workflow* on page 361).
#CHP/CEL	Number of CHP files in this data set for the specified CEL file.
File Date	The date and time the CEL file was last modified.

NOTE: See *Table Features* on page 198 for more information on customizing the table view.

NOTE: The ligation nucleotide is the nucleotide at the 3' end of a solution probe which is the nucleotide that is ligated to the array probe. The AT channel is the optical channel in which signal from ligated A or T nucleotides are detected. The GC channel is the optical channel in which signal from ligated G or C nucleotides are detected. The AT probes are those control probes that correspond to non-polymorphic genomic positions for which the expected ligation nucleotide is A or T. The GC probes are those control probes that correspond to non-polymorphic genomic positions for which the expected ligation nucleotide is G or C.

Intensity QC Table for SNP 6.0 Data (default view)

The following information can be displayed for SNP 6.0 data after running QC in Genotyping Console (Table 6.4):

Table 6.4 Intensity QC Table metrics for SNP 6.0 data

Column Name	Description
File	CEL file name.
Bounds	In Bounds/Out of Bounds indicates whether the CEL file met the specified QC threshold(s).
Contrast QC	Computed Contrast QC for all QC SNPs.
Contrast QC (Random)	Contrast QC for 10K random autosomal SNPs.
QC Call Rate	Computed QC Call Rate for all QC SNPs.
Computed Gender	Computed gender. For more details, see Appendix G, *Best Practices SNP 6.0 Analysis Workflow* on page 360.
# CHP/CEL	Number of CHP files in this data set for the specified CEL file.
File Date	The date and time the CEL file was last modified.

NOTE: See *Table Features* on page 198 for more information on customizing the table view.

NOTE: The Genome-Wide Human SNP Array 6.0 contains SNPs and CN probe sets from two enzyme sets (Nsp and Sty). Some SNPs and CN probe sets are only present on fragments generated by one of the enzymes, while other SNPs and CN probe sets are present on fragments generated from both of the enzymes.
There are situations where a sample may work properly with one enzyme set, but not with the other.
Contrast QC is broken down by enzyme set to help you evaluate the data for this issue.

Intensity QC Table for Human Mapping 100K/500K & SNP 5.0 Data (default view)

The following information can be displayed for Human Mapping 100K/500K and SNP 5.0 data after running QC in Genotyping Console (Table 6.5):

Table 6.5 Intensity QC Table metrics for 100K/500K and SNP 5.0 data

Column Name	Description
File	CEL file name.
Bounds	In Bounds/Out of Bounds indicates whether the CEL file met the specified QC threshold(s).
QC Call Rate	Computed QC Call Rate for all QC SNPs.
QC Call Rate (Nsp)	See note below.
QC Call Rate (Nsp/Sty Overlap)	See note below.
QC Call Rate (Sty)	See note below.
Computed Gender	Computed gender. For more details, see Appendix E, *Gender Calling in GTC* on page 353.
# CHP/CEL	Number of CHP files in this Data Set for the specified CEL file.
File Date	The date and time the CEL file was last modified.

NOTE: The Genome-Wide Human SNP Array 5.0 contains SNPs and CN probe sets from two enzyme sets (Nsp and Sty). Some SNPs and CN probe sets are present on fragmented from one of the enzyme sets, while other SNPs and CN probe sets are present on fragments generated from both of the enzymes.
There are situations where a sample may work properly with one enzyme set, but not with the other.
The QC Call rate is broken down by enzyme set to help you evaluate the data for this issue.

NOTE: See *Table Features* on page 198 for more information on customizing the table view.

Creating Custom Intensity Data Groups using Intensity QC Data

IMPORTANT: Axiom array plates may be processed with more than one reagent version. You can only perform batch genotyping analysis on CEL files processed using the same reagent version. You can create custom intensity data groups with CEL files processed using the same reagent version.

Genotyping Console allows for custom grouping of intensity data (CEL) Files based on Intensity QC performance.

You can also create custom groups of intensity data files by selecting sample files using:

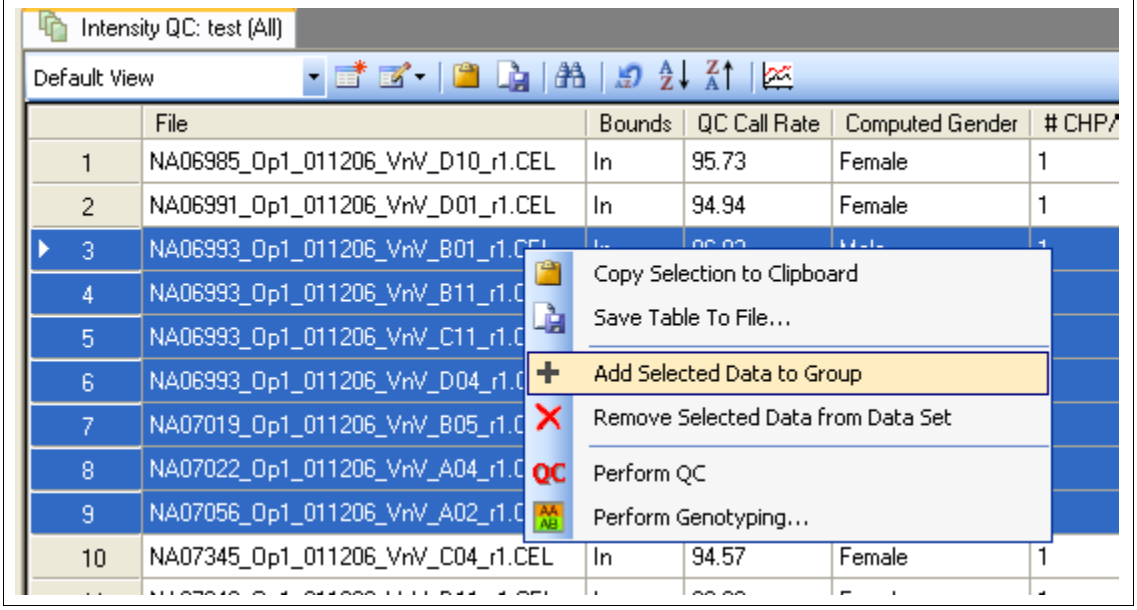
- *Creating a Custom Intensity Group from the CHP File Data* on page 117
- *Creating Custom Intensity Data Groups Using the SNP Cluster Graph* on page 162

This feature enables you to group Axiom CEL files processed using the same reagent version before genotyping analysis.

To make a custom group of intensity data files:

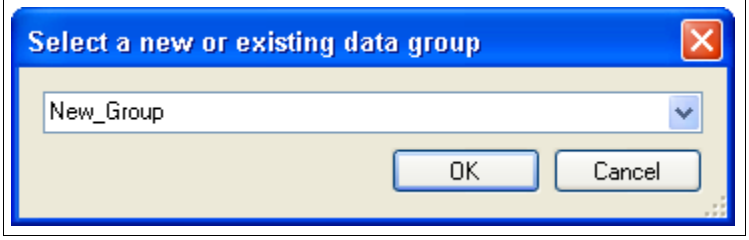
1. Select the row(s) to be added to the new group from an open Intensity QC table.
   See *Table Features* on page 198 for information on sorting the table by metrics values and selecting rows.
2. Right-click on the selected rows and select **Add Selected Data to Group** (Figure 6.11).

Figure 6.11 Add Selected Data to Group



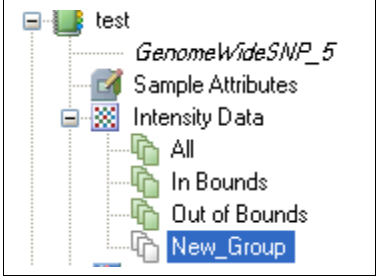
The Select a new or existing data group dialog box opens (Figure 6.12).

Figure 6.12 Select data group



3. Enter a name or select an existing data group in the drop-down list and select **OK**.
   The new group will be displayed in the tree. Custom groups are indicated by white icons, while the default groups are indicated by green icons (Figure 6.13).

Figure 6.13 Intensity data groups in the GTC data tree



Custom Intensity groups can be re-named by right-clicking on the group and selecting **Rename Intensity Data Group**.

Custom Intensity groups can be deleted by right-clicking on the group and selecting **Remove Intensity Data Group**.

NOTE: Removing a custom Intensity Data Group does not remove the data from the Data Set. To remove Intensity data, see *Removing Data from a Data Set* on page 60.

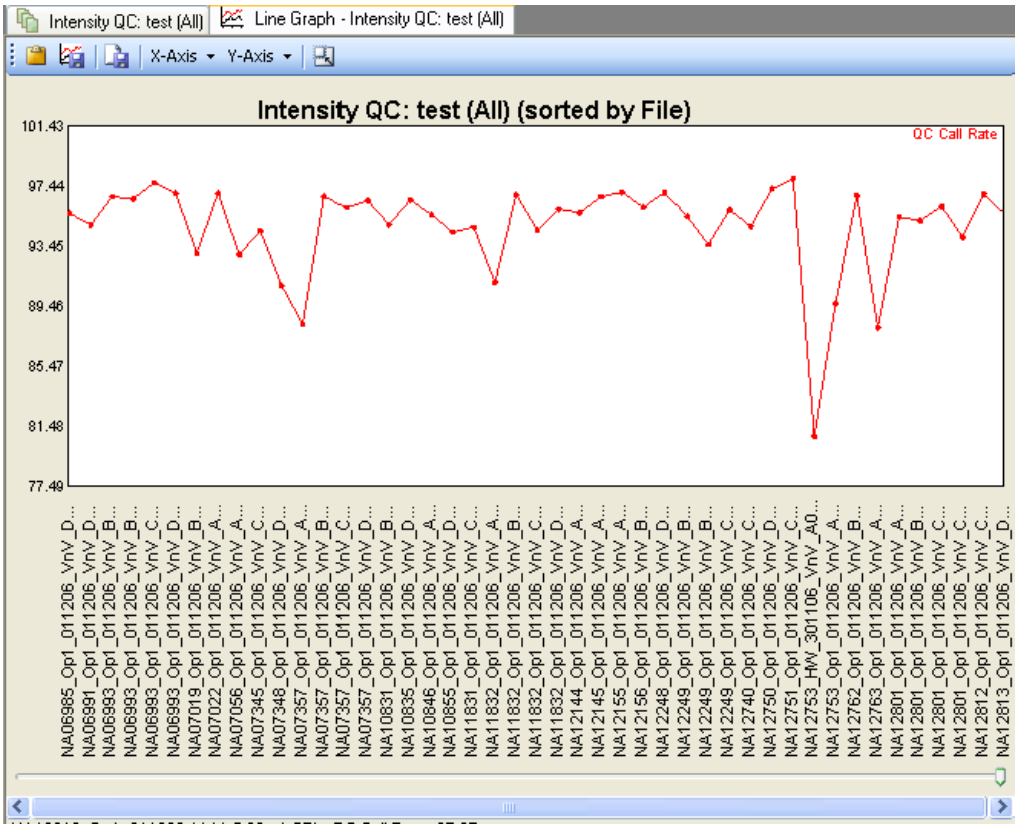
Graphing QC Results

In addition to the tabular display of the metrics, the QC results can be displayed in a line graph. The graphical display is useful in identifying outlier samples.

To open a line graph:

- Click on the Line Graph shortcut on the Intensity QC Table tool bar.

Figure 6.14 Graph of QC data



For more information, see *Graph Features* on page 203.

NOTE: Values displayed in tables or exported to a text file are only done with a certain number of digits after the decimal. Filtering is performed using the full precision stored in the SNP statistics file.

Signature Genotypes

During the QC step in Genotyping Console, a set of SNPs are genotyped using the QC algorithm shown in the table below (Table 6.6). These SNPs can be used to verify a sample's identity by comparing the genotype calls to the SNP calls made using a different technology, for example, genotyping by PCR, or other references.

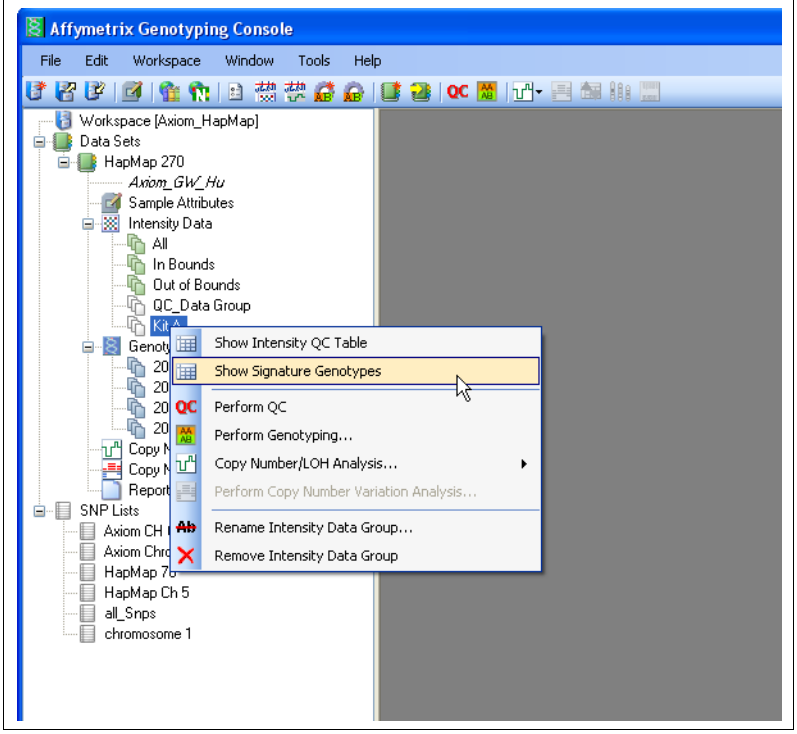
Table 6.6 Algorithms used to make Signature SNP genotype calls

Array Type	Number of Signature SNPs	Signature SNP Genotyping Generated Using:
Human Mapping 100K Array	31	Dynamic Model (DM) algorithm
Human Mapping 500K Array	50	Dynamic Model (DM) algorithm
Genome-Wide Human SNP Array 5.0	72	Dynamic Model (DM) algorithm
Genome-Wide Human SNP Array 6.0	72	Contrast QC (CQC) is the primary QC method, as Dynamic Model (DM) algorithm was also used for QC
Axiom Genome-Wide CEU Array Plate Axiom Genome-Wide ASI Array Plate	83 88	Dish QC (DQC) followed by measuring the genotyping call rate as generated during the 2nd pass genotyping with the Axiom GT1 algorithm
Axiom Genome-Wide BOS 1 Array Plate	116	Dish QC (DQC) followed by measuring the genotyping call rate as generated during the 2nd pass genotyping with the Axiom GT1 algorithm

To see the Signature SNP genotypes:

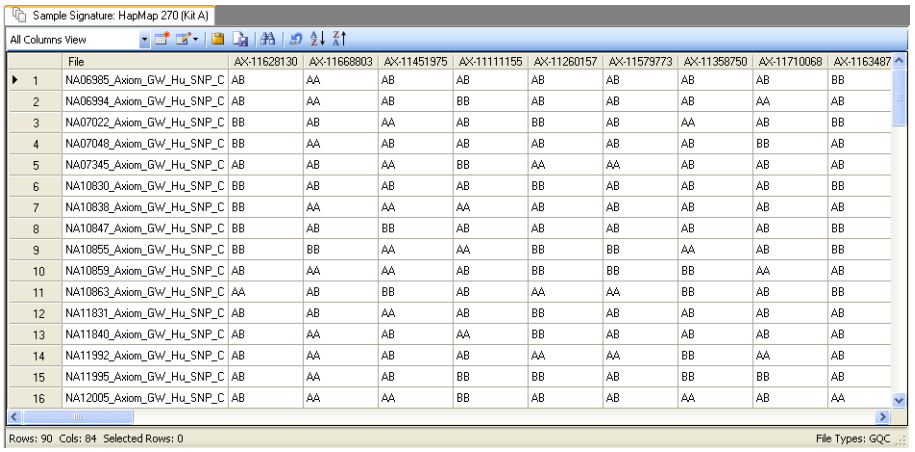
- Right-click an Intensity QC group and select **Show Signature Genotypes** (Figure 6.15).

Figure 6.15 Show Signature Genotypes menu item



The Sample Signature table opens showing the genotype calls for the Signature SNPs (Figure 6.16).

Figure 6.16 Sample Signature table with Signature SNP genotypes



	File	AX-11628130	AX-11668803	AX-11451975	AX-11111155	AX-11260157	AX-11579773	AX-11358750	AX-11710068	AX-1163487
1	NA06985_Axiom_GW_Hu_SNP_C	AB	AA	AB	AB	AB	AB	AB	AB	BB
2	NA06994_Axiom_GW_Hu_SNP_C	AB	AA	AB	BB	AB	AB	AB	AA	AB
3	NA07022_Axiom_GW_Hu_SNP_C	BB	AB	AA	AB	BB	AB	AA	AB	BB
4	NA07048_Axiom_GW_Hu_SNP_C	BB	AA	AB	AB	AB	AB	AB	BB	AB
5	NA07345_Axiom_GW_Hu_SNP_C	AB	AB	AA	BB	AA	AA	AB	AB	AB
6	NA10830_Axiom_GW_Hu_SNP_C	BB	AB	AB	AB	BB	AB	AB	AB	BB
7	NA10838_Axiom_GW_Hu_SNP_C	BB	AA	AA	AA	AB	AB	AB	AB	AB
8	NA10847_Axiom_GW_Hu_SNP_C	BB	AB	BB	AB	AB	AB	AB	AB	BB
9	NA10855_Axiom_GW_Hu_SNP_C	BB	BB	AA	AA	BB	BB	AA	AB	BB
10	NA10859_Axiom_GW_Hu_SNP_C	AB	AA	AA	AB	BB	BB	BB	AA	AB
11	NA10863_Axiom_GW_Hu_SNP_C	AA	AB	BB	AB	AA	AA	BB	AB	BB
12	NA11831_Axiom_GW_Hu_SNP_C	AB	AB	AA	AB	BB	AB	AB	AB	AB
13	NA11840_Axiom_GW_Hu_SNP_C	AB	AA	AB	AA	BB	AB	AB	AB	AB
14	NA11992_Axiom_GW_Hu_SNP_C	AB	AA	AB	AB	AA	AA	BB	AA	AB
15	NA11995_Axiom_GW_Hu_SNP_C	AB	AA	AB	BB	BB	AB	BB	BB	AB
16	NA12005_Axiom_GW_Hu_SNP_C	AB	AA	AA	BB	AB	AB	AA	AB	AA

By default the following columns are displayed:

File - file name

AFFX-SNP_# - Probe set ID for signature SNP

Annotations for these signature SNPs can be obtained either from NetAffx, or by first *importing a custom SNP list* containing the listed Probe Set IDs. For more information on displaying data in the Sample Signature Table see *Table Features* on page 198.

Genotyping Analysis

Genotyping Console supports genotyping analysis for the following algorithms and arrays, as described in the table below (Table 7.1):

Table 7.1 CHP Algorithms and array types for genotyping analysis

Algorithm	Array Type
BRLMM	Human Mapping 100K Array Human Mapping 500K Array
BRLMM-P	Genome-Wide Human SNP Array 5.0 Rat and Mouse Arrays
Birdseed v1 or Birdseed v2	Genome-Wide Human SNP Array 6.0

Axiom GT1	Axiom Arrays, including:
	• Axiom Human Arrays:
	  • Axiom Genome-Wide Human Arrays
	  • Axiom Genome-Wide CEU 1 Array
	  • Axiom Genome-Wide ASI 1 Array
	  • Axiom Genome-Wide YRI 1 Array set
	  • Axiom myDesign Custom Arrays
	• Axiom Genome-Wide BOS 1 Array

The following sections describe:

Performing Genotyping Analysis

Association studies are designed to identify SNPs with subtle allele frequency differences between different populations. Genotyping errors, differences in sample collection and processing, and population differences are among the many things that can contribute to false positives or false negatives. Efforts should be made to minimize or account for technical or experimental differences. For example, randomization of cases and controls prior to genotyping can reduce or eliminate any possible effects from running cases and controls under different conditions.

IMPORTANT: **Affymetrix recommends that you perform genotyping and QC analysis with all files stored locally. For more details on the hard disk space requirements to perform genotyping, see Appendix J, *Hard Disk Requirements* on page 363.**

The two-step genotyping workflow can be used to optimize genotyping calls. See *Two-Step Genotyping Workflow* on page 123.

This section includes:

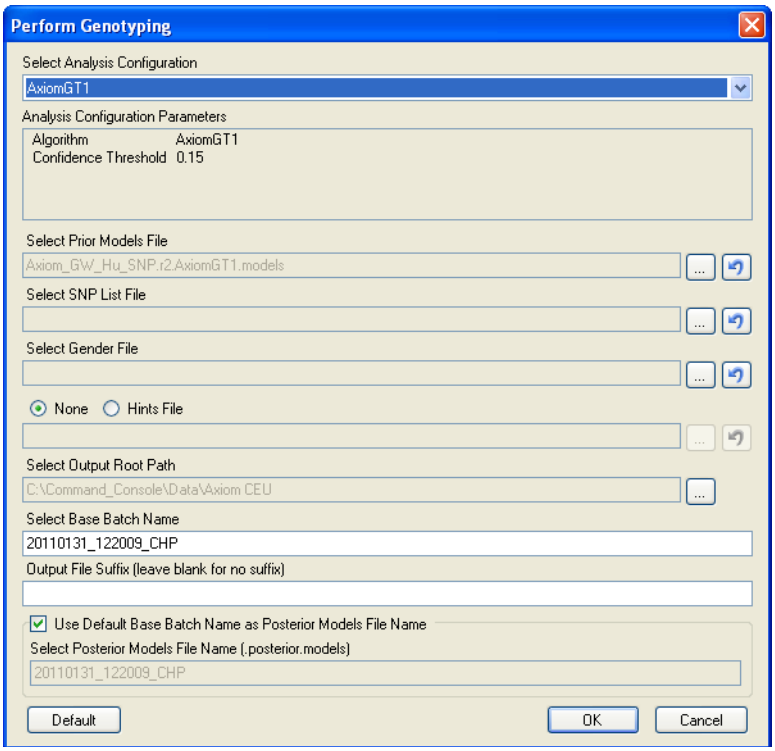
- *Running a Genotyping Analysis* on page 93

Intro to Genotyping Options

GTC provides multiple options for performing genotyping.

Genotyping options are selected in the Perform Genotyping dialog box (Figure 7.1) prior to initializing the genotyping.

Figure 7.1 Perform Genotyping dialog box for Axiom GT1 algorithm



Not all options are available for all the different types of analyses and arrays.

The following options are common to all analyses and are described in *Running a Genotyping Analysis* on page 93:

- Select Output Root Path
- Select Base Batch Name
- Output File Suffix

You can create and select a new analysis configuration for all array types, but the specific options vary from array to array. See *Analysis Configuration Options* on page 100 for more information.

See *Other Genotyping Options* on page 103 for information about the other options.

Selecting the Number of Samples for Analysis

See the notes below for information on determining the number of samples for analysis.

IMPORTANT: **See the BRLMM white paper, BRLMM-P white paper and Birdseed references on Affymetrix.com for recommendations on minimum number of samples to run. In general, more samples are better, 44 per batch is recommended for these algorithms, though fewer may yield acceptable results.**

100K/500K

NOTE: For Human Mapping 100K/500K array sets, the algorithm is run on each array type separately. Therefore, the CHP files are grouped in two batch results, and the CHP Summary data for each array type will be displayed in its own table. Each table will have the appropriate array type appended to its base batch name. You may choose to create a custom group that contains all CHP files.

NOTE: The BRLMM algorithm requires at least two observations of each genotype to create a prior, so 6 is the absolute minimum number of samples required to run this algorithm. However, running it with this small a number is not advised. Performance has been seen to peak when running 50 or so samples. Depending on sample quality, fewer can yield acceptable results.

SNP5/SNP6

NOTE: For BRLMM-P and Birdseed (v1), there is no minimum required number of CEL files. You can run either on a single CEL file, although performance may be poor. Running Birdseed v2 requires a minimum of two samples, although performance may be poor. It is recommended that each BRLMM-P or Birdseed (v1) or Birdseed v2 clustering run consist of at least 44 samples.

SNP 6 Only

NOTE: Birdseed v2 uses the EM algorithm to derive a max likelihood fit of a 2-dimensional Gaussian mixture model in A vs. B space. A key difference between Birdseed (v1) and Birdseed v2 is that v1 uses SNP-specific models or priors only as an initial condition from which the EM fit is free to wander- on rare occasions this allows for mislabeling of the clusters. For Birdseed v2 the SNP-specific priors are used not only as initial conditions for EM, but are incorporated into the likelihood as Bayesian priors. This constrains the extent to which the EM fit can wander off. Correctly labeling SNP clusters, whose centers have shifted relative to the priors, is problematic for both Birdseed versions. However, given the additional constraint on the EM fit, Birdseed v2 is more likely than Birdseed to either correctly label the clusters or set genotypes to No Calls.

NOTE: For Birdseed or Birdseed v2, chromosome X and Y performance within each gender will be influenced by the number of samples of that gender in the clustering. For example, clustering a single female with males will yield typical high performance on autosomal SNPs for all samples, but performance on the X chromosome for the female may be poor. For good performance on X in females, it is recommended that at least 15 female samples be included in the clustering run. For X or Y in males there is no minimum requirement.

Axiom

NOTE: Running Axiom GT1 requires a minimum of 20 distinct samples with either zero female samples or at least 10 distinct female samples. See **Appendix H, *Best Practices Axiom Analysis Workflow* on page 361** for more details.

NOTE: Running Axiom GT1 with generic priors for Axiom myDesign™ arrays requires a minimum of 90 distinct samples with either zero female samples or at least 30 distinct female samples.

IMPORTANT: Axiom array plates can be processed with more than one reagent version. You can only perform batch genotyping analysis on CEL files processed using the same reagent version. You can create custom intensity data groups with CEL files processed using the same reagent version after performing QC (see *Creating Custom Intensity Data Groups using Intensity QC Data* on page 84).

Special Considerations for Axiom Data

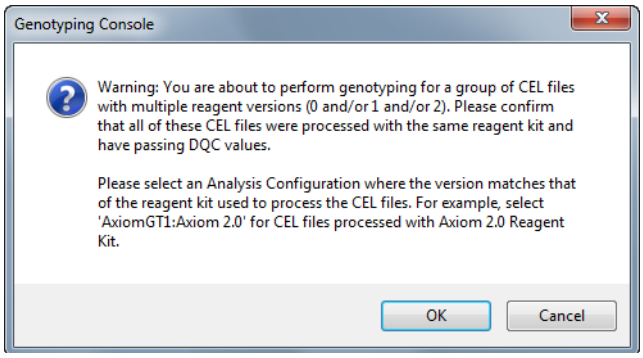
Axiom array plates can be processed with more than one reagent version. CEL files processed with either reagent version can be distinguished during intensity quality control, which can detect the reagent version used.

Genotyping analysis should only be performed with CEL files processed using the same reagent version.

Occasionally during QC, CEL files processed with a single reagent version will be assigned to the incorrect reagent version during QC, resulting in CEL files with mixed reagent versions.

If CEL files with mixed reagent versions are grouped together for genotyping analysis, the following notice appears (Figure 7.2).

Figure 7.2 Warning for multiple reagent versions during Axiom Genotyping



If this warning shows up, do one of the of the following:

- If the CEL files were processed using the same reagent kit, but QC provided mixed reagent versions you can still genotype all the data together by choosing the correct matching genotyping configuration. For example, choose “AxiomGT1: Axiom 2.0” for CEL files processed with Axiom 2.0 Reagent Kit.
- If the CEL files were indeed processed using different reagent kits, create custom intensity data groups with CEL files processed using the same reagent version, and then perform genotyping for each new data group separately.

Running a Genotyping Analysis

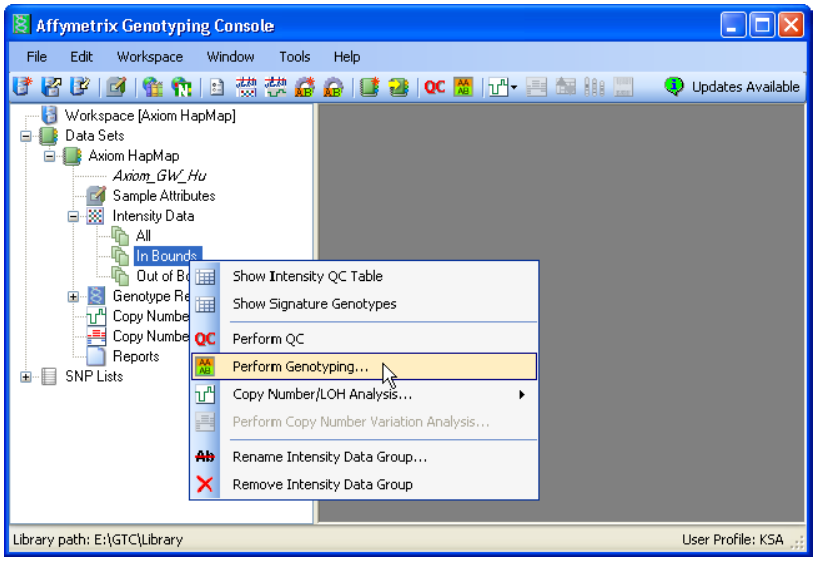
IMPORTANT: Affymetrix recommends that you perform genotyping and QC analysis with all files stored locally. For more details on the hard disk space requirements to perform genotyping, see **Appendix J, *Hard Disk Requirements* on page 363**.

To initiate genotyping analysis:

1. Right-click on a CEL intensity group (e.g. In Bounds or Custom Group) and select **Perform Genotyping…** (Figure 7.3).

IMPORTANT: You can only perform batch genotyping analysis on Axiom CEL files processed using the same reagent version. You can create custom intensity data groups with CEL files processed using the same reagent version after performing QC. See ***Creating a Custom Intensity Group from the CHP File Data* on page 117**.

Figure 7.3 Perform Genotyping…



If Axiom CEL files that have been assigned to multiple reagent versions are selected in a data group for genotyping analysis, a warning notice appears (Figure 7.2).

See *Special Considerations for Axiom Data* on page 92 for more information.

The Perform Genotyping dialog box opens.

The Perform Genotyping dialog box has different options for different array types (Figure 7.4 through Figure 7.9).

Figure 7.4 Perform Genotyping dialog box, Axiom™ Genome-Wide Array Plate (processed with Reagent Version 1)

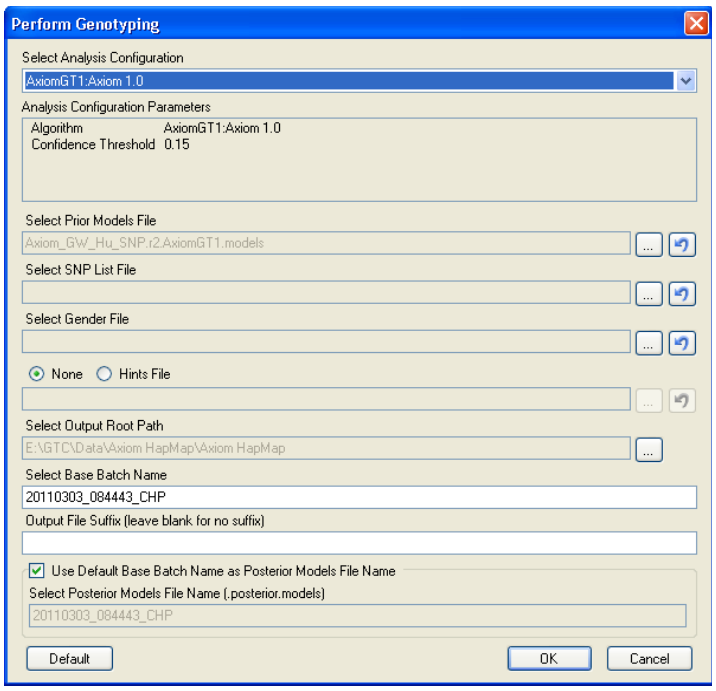


Figure 7.5 Perform Genotyping dialog box, Genome-Wide Human SNP Array 6.0

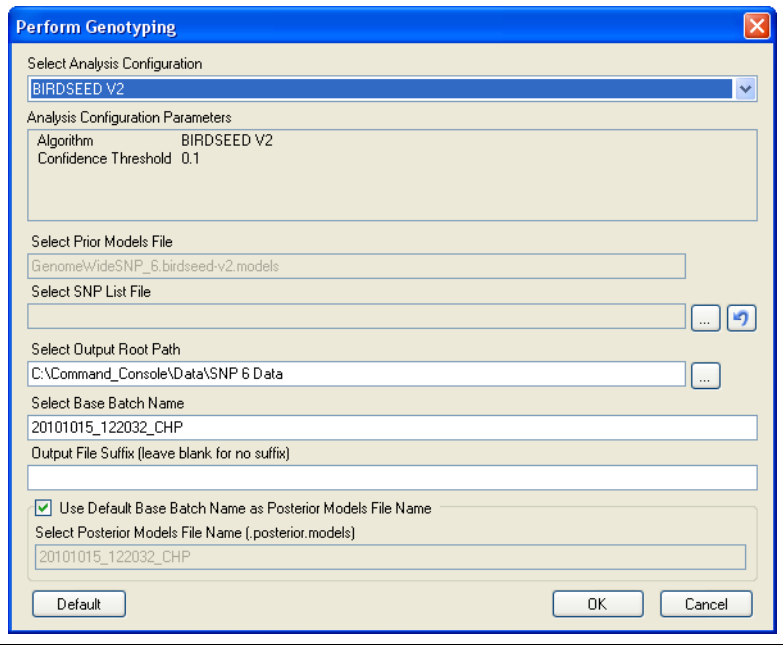


Figure 7.6 Perform Genotyping dialog box, Genome-Wide Human SNP Array 5.0

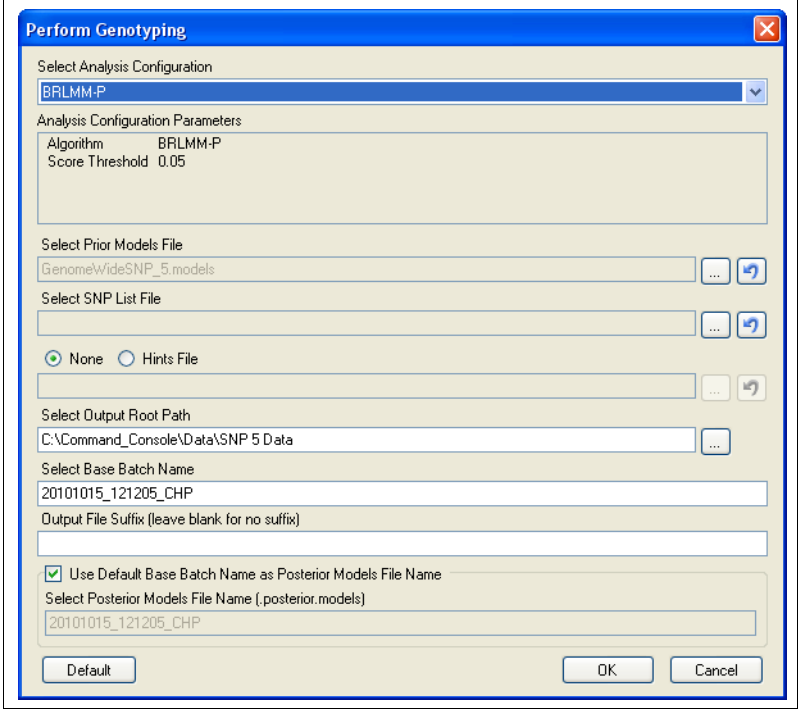


Figure 7.7 Perform Genotyping dialog box, Affymetrix® Mouse Diversity Genotyping Array

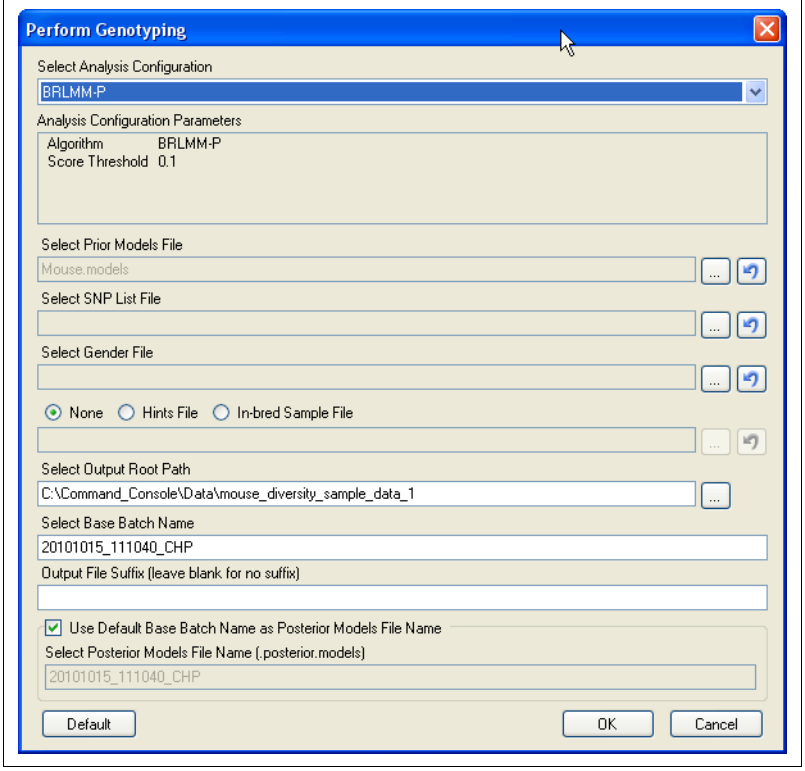


Figure 7.8 Perform Genotyping dialog box, GeneChip Human Mapping 500K Array Set

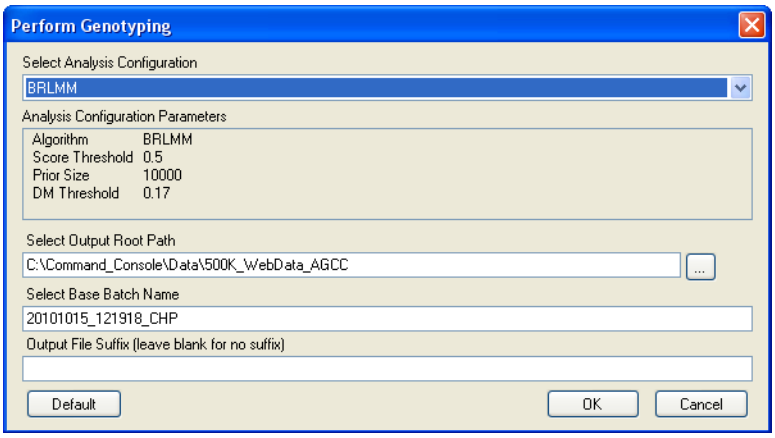
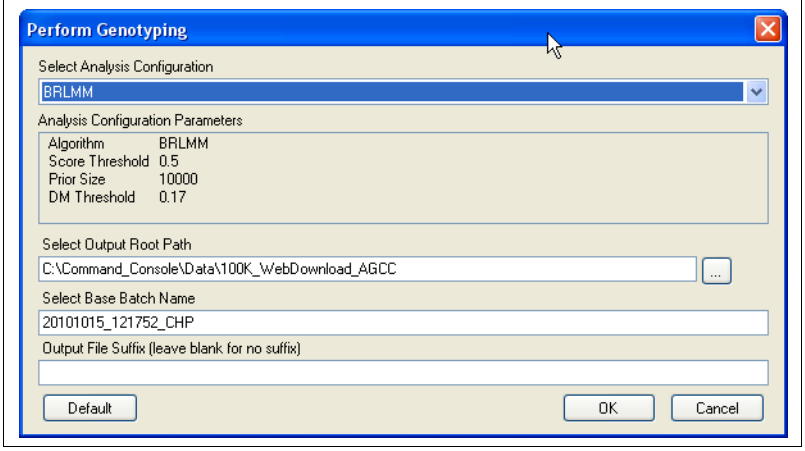


Figure 7.9 Perform Genotyping dialog box, GeneChip Human Mapping 100K Array Set

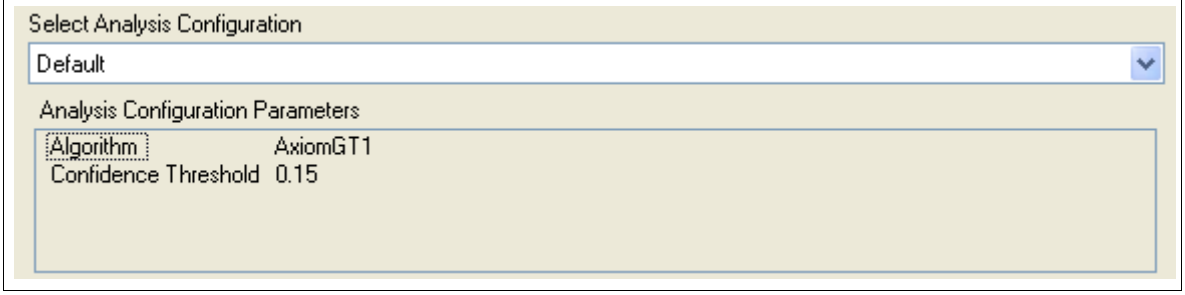


2. Select the Analysis Configuration.

   A different analysis configuration can be selected for each type of analysis, but the available parameters vary depending upon the analysis (see *Parameter Definition and Default Settings* on page 100).

   The available analysis configurations are available from the drop down menu (Figure 7.10).

Figure 7.10 Select Analysis Configuration, Axiom GT1

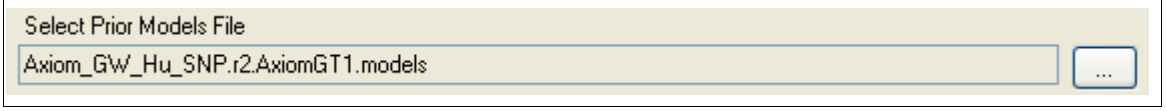


   The current settings are displayed below the menu.

To modify the default settings, see *Analysis Configuration Options* on page 100.

3. Select a Prior Model File if the option is available (Figure 7.11).

Figure 7.11 Select Prior Models File, Axiom GT1



This option can be used for:

- BRLMM-P (SNP 5, mouse, and rat)
- Axiom GT1 (including custom Axiom™ myDesign™ arrays and Axiom™ Genome-Wide BOS 1 arrays)

The currently selected model file is displayed in the Select Prior Models Files box.

See *Select Prior Models File* on page 105 for more information on selecting a prior model file.

See *Model Files Options* on page 104 for a discussion of the types of model files and how they are used in genotyping.

4. Select a SNP List file if the option is available (Figure 7.12).

Figure 7.12 Select SNP List File

The SNP List option can be selected for:

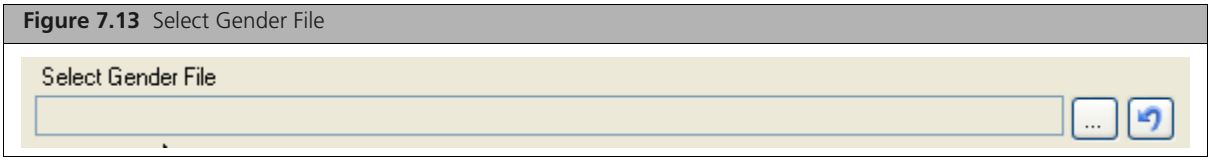
- BRLMM-P (SNP 5, mouse and rat)
- Birdseed V1 and Birdseed V2
- Axiom GT1 (including custom Axiom™ myDesign™ arrays and Axiom™ Genome-Wide BOS 1) arrays

See *Select SNP List File* on page 106.

NOTE: You can use the SNP List file name as a suffix for the CHP files to distinguish CHP files generated with different SNP lists.

5. Select a Gender File if the option is available (Figure 7.13).

Figure 7.13 Select Gender File



A Gender file is a list of the samples with gender calls.

This option is available for:

- BRLMM-P for mouse and rat only
- Axiom GT1 (including custom Axiom™ myDesign™ arrays and Axiom™ Genome-Wide BOS 1 arrays)

See *Gender File* on page 107.

6. Select a Hints File or In-bred Sample File if the option is available (Figure 7.14, Figure 7.15.)

NOTE: **The Hints and In-bred Sample file options are mutually exclusive.**

Figure 7.14 Select Hints File (Axiom GT and BRLMM-P only)

Figure 7.15 Select Hints or In-Bred Sample File (BRLMM-P and Axiom for non-human arrays only)

The Hints file option can be selected for:

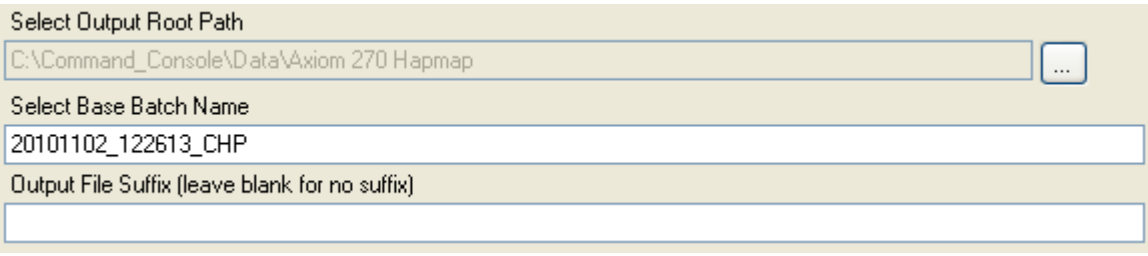
- BRLMM-P (including mouse and rat)
- Axiom GT1 (including custom Axiom™ myDesign™ arrays and Axiom™ Genome-Wide BOS 1 arrays)

The In-bred Sample File option is available only for non-human arrays.

See *Hints and In-bred Sample File Options* on page 108.

7. Change the output options if desired (Figure 7.16).

Figure 7.16 Analysis output options



Change the following if desired:

- Output Root Path: location of the Genotyping Results Group folder.
- Base Batch Name: Name of the Genotyping Results Group and its folder.

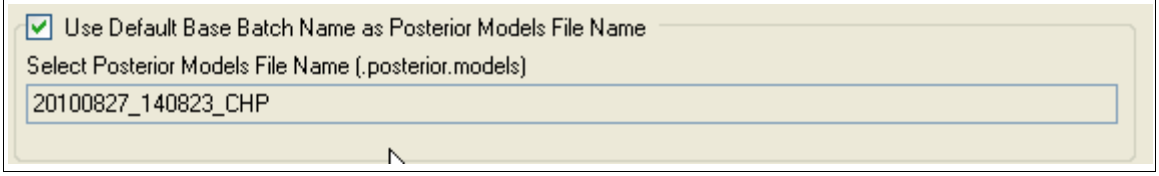
NOTE: **This folder is the location where the different Data Results files are kept. You can access the folder through Windows Explore to view report files.**

- Output File Suffix: suffix added to distinguish output file names.

NOTE: **The default batch name includes the date and time; therefore, it is unique for each run.**

8. Select the Posterior File options if available (Figure 7.17).

Figure 7.17 Select Posterior File Options



This option is available for:

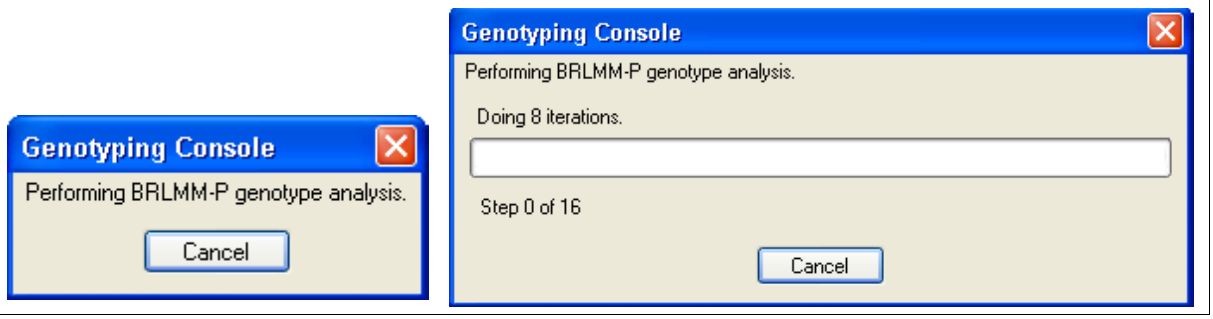
- BRLMM-P
- Birdseed v1 and Birdseed V2
- Axiom GT1

See *Posterior File Options* on page 106.

9. Click **OK**.

   Once the genotyping analysis is initiated, several windows will be displayed showing the progress of the algorithm (Figure 7.18):

Figure 7.18 Progress dialog boxes for Genome-Wide Human SNP Array 5.0



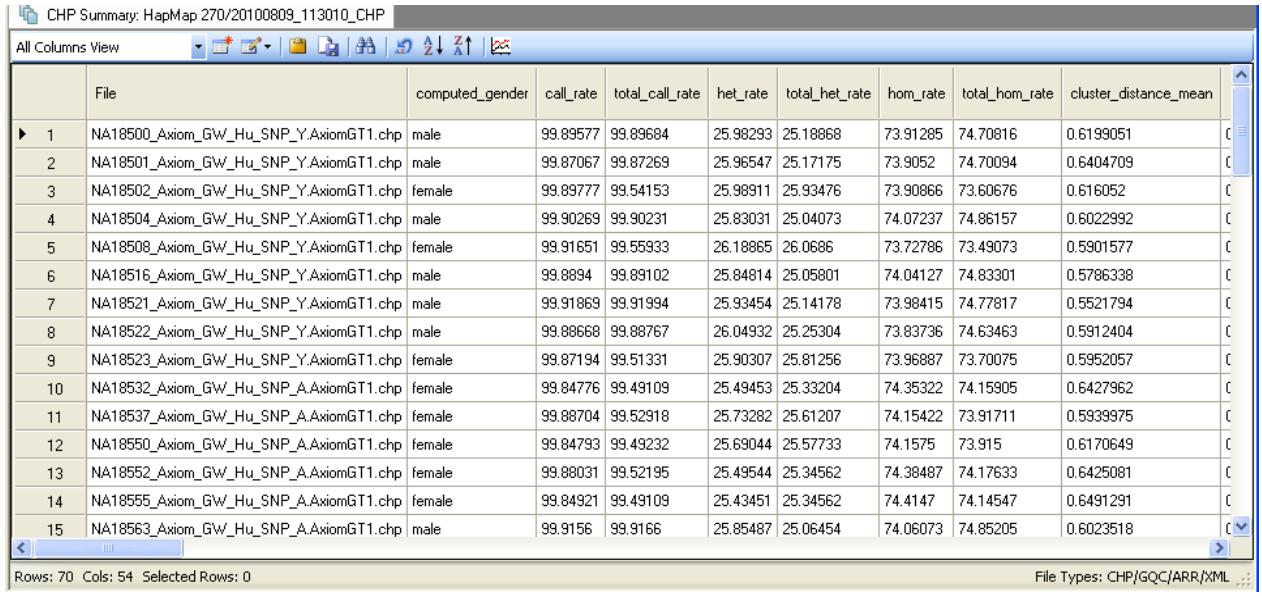
NOTE: The status messages window also displays information regarding the algorithm process.

NOTE: For fastest run time, Affymetrix recommends performing genotyping analysis with all files stored locally.

IMPORTANT: Batches of up to 800 CEL files (Axiom™ Genome-Wide Human Array) have been successfully run on the recommended workstation.

When the algorithm completes the genotyping analysis, GTC automatically displays the CHP Summary Table.

Figure 7.19 CHP Summary table


CHP Summary: HapMap 270/20100809_113010_CHP

All Columns View

	File	computed_gender	call_rate	total_call_rate	het_rate	total_het_rate	hom_rate	total_hom_rate	cluster_distance_mean
1	NA18500_Axiom_GW_Hu_SNP_Y.AxiomGT1.chp	male	99.89577	99.89684	25.98293	25.18868	73.91285	74.70816	0.6199051
2	NA18501_Axiom_GW_Hu_SNP_Y.AxiomGT1.chp	male	99.87067	99.87269	25.96547	25.17175	73.9052	74.70094	0.6404709
3	NA18502_Axiom_GW_Hu_SNP_Y.AxiomGT1.chp	female	99.89777	99.54153	25.98911	25.93476	73.90866	73.60676	0.616052
4	NA18504_Axiom_GW_Hu_SNP_Y.AxiomGT1.chp	male	99.90269	99.90231	25.83031	25.04073	74.07237	74.86157	0.6022992
5	NA18508_Axiom_GW_Hu_SNP_Y.AxiomGT1.chp	female	99.91651	99.55933	26.18865	26.0686	73.72786	73.49073	0.5901577
6	NA18516_Axiom_GW_Hu_SNP_Y.AxiomGT1.chp	male	99.8894	99.89102	25.84814	25.05801	74.04127	74.83301	0.5786338
7	NA18521_Axiom_GW_Hu_SNP_Y.AxiomGT1.chp	male	99.91869	99.91994	25.93454	25.14178	73.98415	74.77817	0.5521794
8	NA18522_Axiom_GW_Hu_SNP_Y.AxiomGT1.chp	male	99.88668	99.88767	26.04932	25.25304	73.83736	74.63463	0.5912404
9	NA18523_Axiom_GW_Hu_SNP_Y.AxiomGT1.chp	female	99.87194	99.51331	25.90307	25.81256	73.96887	73.70075	0.5952057
10	NA18532_Axiom_GW_Hu_SNP_A.AxiomGT1.chp	female	99.84776	99.49109	25.49453	25.33204	74.35322	74.15905	0.6427962
11	NA18537_Axiom_GW_Hu_SNP_A.AxiomGT1.chp	female	99.88704	99.52918	25.73282	25.61207	74.15422	73.91711	0.5939975
12	NA18550_Axiom_GW_Hu_SNP_A.AxiomGT1.chp	female	99.84793	99.49232	25.69044	25.57733	74.1575	73.915	0.6170649
13	NA18552_Axiom_GW_Hu_SNP_A.AxiomGT1.chp	female	99.88031	99.52195	25.49544	25.34562	74.38487	74.17633	0.6425081
14	NA18555_Axiom_GW_Hu_SNP_A.AxiomGT1.chp	female	99.84921	99.49109	25.43451	25.34562	74.4147	74.14547	0.6491291
15	NA18563_Axiom_GW_Hu_SNP_A.AxiomGT1.chp	male	99.9156	99.9166	25.85487	25.06454	74.06073	74.85205	0.6023518

Rows: 70 Cols: 54 Selected Rows: 0 File Types: CHP/GQC/ARR/XML

For more information, see *CHP Summary Table* on page 110.

Analysis Configuration Options

Certain genotyping algorithm parameters can be changed to match experimental conditions. You can modify or create a configuration for all analysis and array types, but the particular parameters that can be changed will vary.

This section provides information on:

Parameter Definition and Default Settings

The following parameters can be changed:

Score/Confidence Threshold	The maximum value of confidence for which the algorithm will make a genotype call. Calls with confidence scores less than the threshold are assigned a call. For example, if the threshold is 0.15, then any SNP with confidence < 0.15 is called, and any SNP with confidence ≥ 0.15 is not called. If the threshold is increased (maximum = 1), then additional SNPs in which there is less confidence (higher confidence score) will be called.
Prior Size (100K/500K only)	How many probe sets to use for determining prior.
DM Threshold (100K/500K only)	DM confidence threshold used for seeding clusters.

The table below (Table 7.2) lists the default settings for configuration parameters for the different types of arrays and algorithms:

Table 7.2 Algorithm Parameters

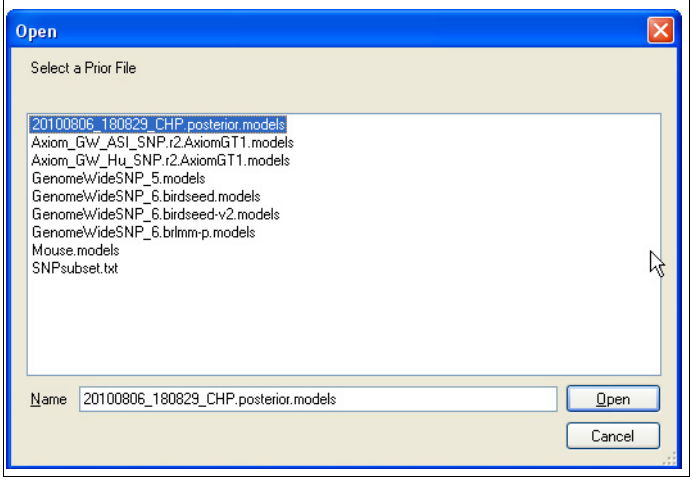
Algorithm	BRLMM	BRLMM-P		Birdseed	Birdseed 2	Axiom GT	
Array	100K/ 500K	SNP 5	Rat and Mouse	SNP 6	SNP 6	Human	Bovine
Score/confidence Threshold	0.5	0.5	0.1	0.1	0.1	0.15	0.15
Prior Size	10000	N/A	N/A	N/A	N/A	N/A	N/A
DM Threshold	0.17	N/A	N/A	N/A	N/A	N/A	N/A

Modifying the Parameters

To modify the default algorithm settings:

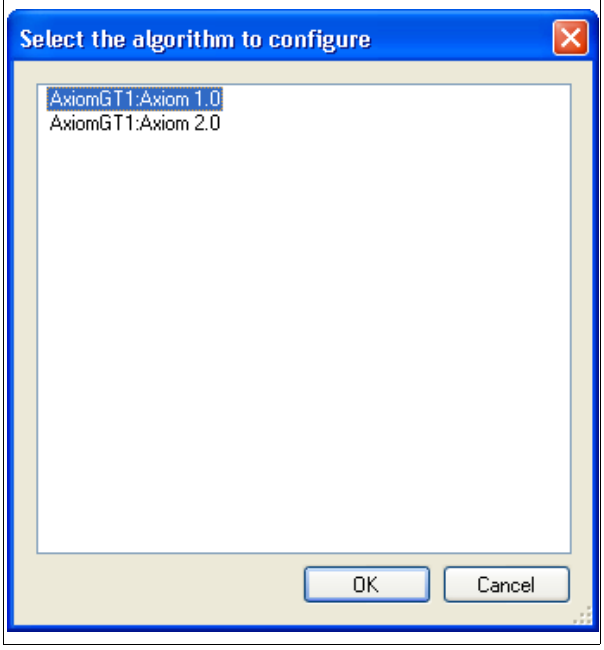
1. Select the **New Genotyping Configuration** shortcut on the main tool bar, or
   From the Edit menu, select **Genotyping Configurations > New Configuration**.
   The Select a Probe Array Type dialog box opens (Figure 7.20).

Figure 7.20 Select a Probe Array Type dialog box



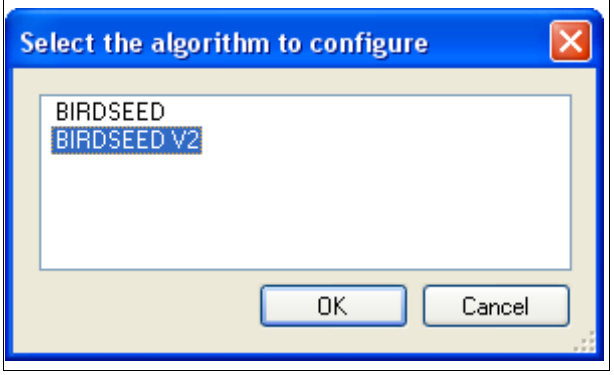
2. Select the array type from the list and click **Select**.
   For the Axiom Genome-Wide ASI Array Plate, you will be asked to choose whether to edit the configuration for AxiomGT1 or AxiomGT2 (Figure 7.21).
   Select “GT1” for Axiom CEL files processed with Reagent Version 1. Select “GT2” for Axiom CEL files processed with Reagent Version 2.

Figure 7.21 Select Axiom GT version



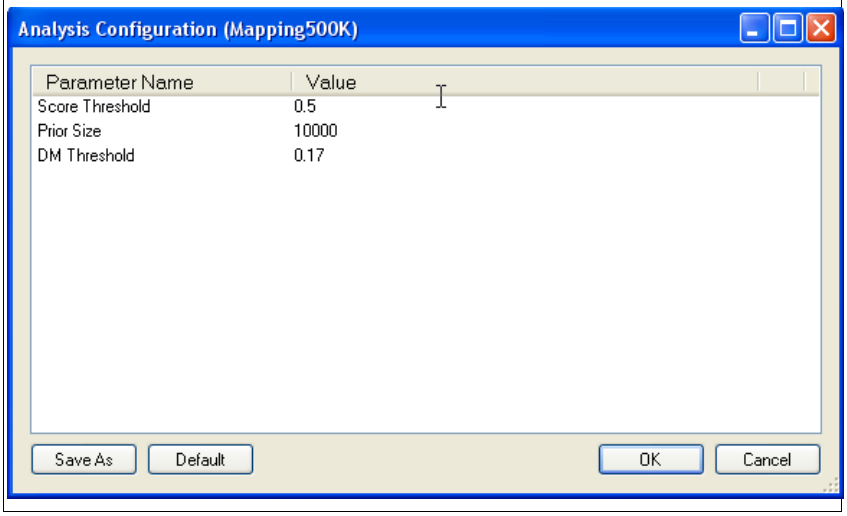
For the Genome-wide Human SNP Array 6.0, you will be asked to choose whether to edit the configuration for Birdseed (v1) or Birdseed v2 (Figure 7.22).

Figure 7.22 Select Birdseed version



3. Next, for all array types, the appropriate Analysis Configuration dialog box opens (Figure 7.23). The default algorithm parameters available for editing will be displayed.

Figure 7.23 Analysis Configuration dialog box



For information about the parameters and settings for different array types, see *Parameter Definition and Default Settings* on page 100

4. Enter a new value for the parameter(s) you wish to change and select **OK**.

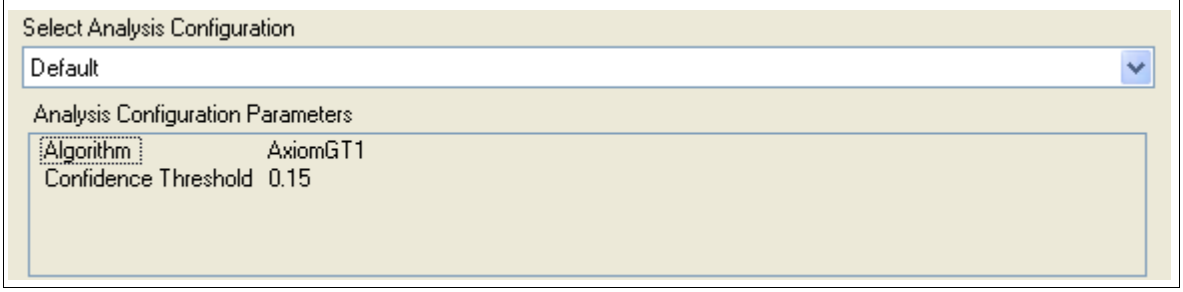
   Click the **Default** button to return to the default settings.

   You will be asked to provide a name for the new genotyping analysis configuration.

Selecting a New Configuration

The default and modified analysis configurations are available from the drop down menu (Figure 7.24) in the Perform Genotyping dialog box.

Figure 7.24 Select Analysis Configuration, Axiom GT1



The current settings are displayed below the menu.

Other Genotyping Options

The table below lists the options that vary by analysis and array type. Additional information can be found at the links.

None of these options are available for 100K/500K arrays.

Table 7.3 Other Genotyping options

	Algorithms	BRLMM-P		Birdseed V1 and V2	Axiom GT1	
	Array Types	**SNP5**	**Non-Human**	**SNP6**	**Axiom Human Arrays Including myDesign**	**Axiom Non-Human**
Parameters	*Select Prior Models File* on page 105	Yes	Yes	No	Yes	Yes
	Posterior File Options on page 106	Yes	Yes	Yes	Yes	Yes
	Hints Files on page 109	Yes	Yes	No	Yes	Yes
	Select SNP List File on page 106)	Yes	Yes	Yes	Yes	Yes
	Inbred Sample File on page 110	No	Yes	No	No	Yes
	Gender File on page 107	No	Yes	No	No	Yes

Model Files Options

GTC 4.2 enables you to select model files for the following genotyping analyses:

- BRLMM-P
- Axiom GT1

These model files contain cluster location information that is used in generating genotyping calls.

We can define genotyping model files in two different ways:

- The methods and data used to create them.
- How they are used in the Genotyping process.

There are three different ways Genotyping model files can be created:

1. Generic model files have generic cluster location information: every diploid and haploid SNP uses the same cluster coordinates.

   NOTE: In some cases a generic model file may have cluster location data for specific SNPs.

2. SNP Specification Model files: have cluster location information based on information from the Affymetrix training data, but not experimental data. These files are provided by Affymetrix.

   SNP Specification Model files contain the best estimate for where genotype clusters are located before using any data in the current experimental dataset. This estimate can come from general principles (the BB genotype should have more intensity in the B probe than in the A probe) or from specific training data (for any given SNP in HapMap), the BB genotypes had the following average intensities). This also incorporates a measure of precision – estimates taken from general principles are treated as being vague and easily overridden by observed data, and estimates from specific training are treated as precise and difficult to override. Note that some clusters for low MAF SNPs may have many observations in the training data and be precise, while the rare homozygous allele cluster may not be known to high precision because of a lack of training data.

3. User-generated posterior model files: contain cluster location information generated during genotyping using:

- data on the cluster location information contained in a model file that was selected prior to genotyping,
- information in the Hints file
- the current experimental data set selected for genotyping.
- For animal samples only:
  - Information in the Inbred Sample file
  - Information in the Gender File

The cluster data in the user-generated posterior model file is then used to produce the reported genotype call for the samples.

Posterior models file contain the best estimate for where genotype clusters are located after the data in the current experimental data set is combined with the prior model information. This posterior set of cluster properties is used to generate the genotype calls. Clusters that are known with high precision in the prior will not change much unless there is a large amount of observed data contradicting that cluster location, clusters that are known with low precision in the prior will easily adapt to observed data. This prevents clusters from being 'mislabeled' as one of the other genotypes, while allowing some flexibility to adapt to the current dataset.

These user-generated files are saved in the same folder as the results CHP files. If you want to use a previously created posterior models file as a prior models file for future genotyping, you will need to copy the posterior models file from the result folder to the current library folder

The model files can be used in different stages of Genotyping:

- Prior Model File: Selected before genotyping begins, used as the starter for the process.

  Prior model files can be any of the three types of model files:
  - Generic
  - SNP Specification
  - User-generated
- Posterior Model File: created during genotyping and used to generate the final calls.

  Posterior model files are always user-generated model files, but not every user-generated file will be used as a prior file for future genotyping.

When viewing SNP data in the SNP Cluster Graph, both the prior model and the posterior model cluster information can be displayed as ellipses or lines for BRLMM-P, Birdseed v1 and V2, and Axiom GT1 data. See Chapter 9, *Using the SNP Cluster Graph* on page 148.

Select Prior Models File

For BRLMM-P and Axiom GT1 analyses you can select a different prior models file than the default model file provided with the library files, including any of the following:

- Generic model file
- SNP Specification file
- Previously generated posterior file

The currently selected model file is displayed in the Select Prior Models Files box (Figure 7.25).

Figure 7.25 Select Prior Models File

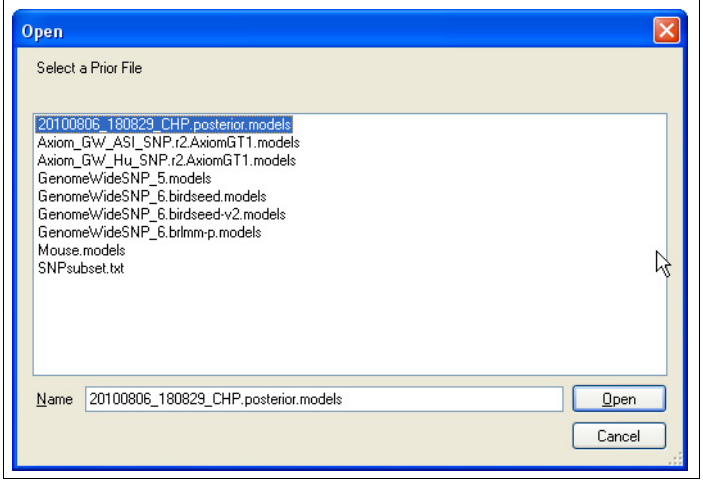
To select a different model file:

1. Make sure that the model file you wish to select is in the GTC Library folder.

   You can copy a posterior file to the folder.

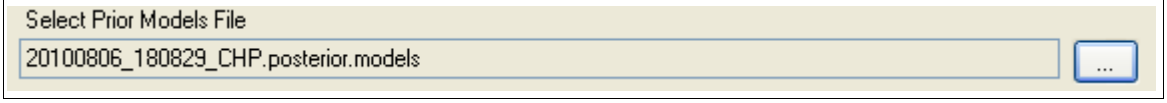
2. Click the **Browse** button.
   The Select a Prior File dialog box opens (Figure 7.26).

Figure 7.26 Select a Prior File dialog box



3. Select the prior model file you wish to use.
4. Click **Open**.
   The new model file name is displayed in the Select Prior Models File box (Figure 7.27).

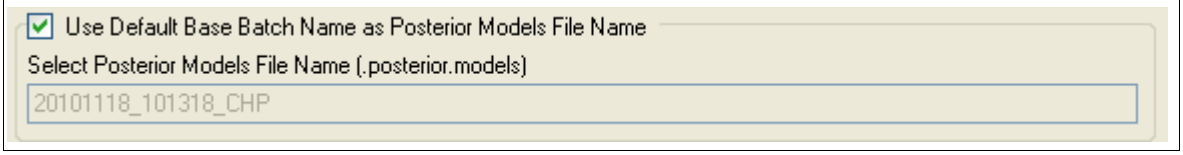
Figure 7.27 New Prior Models File selected



Posterior File Options

The posterior file options allow you to change the name of the posterior models file (Figure 7.28).

Figure 7.28 Posterior file options



- Deselect the Use Default Base Batch Name checkbox and enter a new posterior models file name.

Select SNP List File

The Select SNP List File option enables you to genotype only the SNPs of interest, instead of all the SNPs on the array.

The Select SNP List File option is available for the following genotyping algorithms:

- Axiom GT (human and animal)
- Birdseed/Birdseed 2
- BRLMM-P (human and animal)

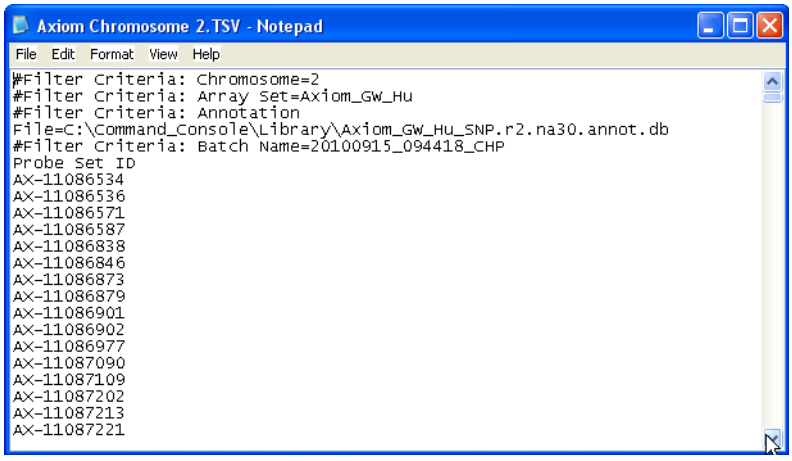
The probe sets genotyped during the analysis will be restricted to those in the SNP List. The genotyping call rate metrics will be calculated only using the probe sets in the SNP List. For Axiom™ myDesign™ Genotyping Arrays and Axiom™ Genome-Wide CHB Array, the call rate metrics will be calculated using the probe sets in the SNP list plus the 3000 SNPQC probe sets. The contents of the resultant CHP files will contain analysis results for those probe sets included in the SNP List instead of all of the SNP probe sets found in the CDF file.

You can use a SNP list generated by GTC or one created in the following tab-separated value (TSV) format:

- Comment lines starting with the hash "#" symbol.
- Probe Set ID
- List of probe sets to be genotyped.

SNP list files can be created and edited using simple text editing programs like Microsoft® Notepad (Figure 7.29).

Figure 7.29 SNP List File viewed in Microsoft® Notepad



To select a SNP list:

1. Make sure that the SNP list file you wish to select is in the GTC Library folder.
   You can copy a file to the folder.

Figure 7.30 Select SNP List File

2. Enter the path and file name (Figure 7.30); or
   Click the **Browse** button and select a SNP list from the dialog box.

Gender File

The Select Gender File option allows you to improve the clustering performance of an algorithm by providing information on the gender of the individual from which the sample was taken.

The gender information in the Gender file substitutes for the computed gender in all respects, including the choice of model used for special SNPs.

The Select Gender File option can be used with:

- Axiom GT (human and non-human arrays)

- BRLMM-P arrays (non-human arrays only)

The file uses the following format:

The first row in the file has the following headings:

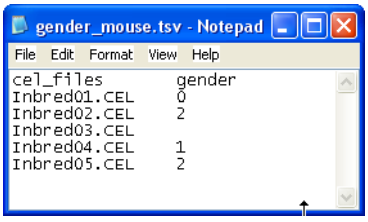
- cel_files: the name of the CEL file corresponding to the samples for which gender info is being provided
- gender: value for the gender call (Table 7.4).

Table 7.4 Sample Gender code

Sample Gender	Code
Unknown	0
Male	1
Female	2

Each following row lists a CEL file name and gender for the individual (Figure 7.31).

Figure 7.31 Gender File information viewed in Microsoft Notepad

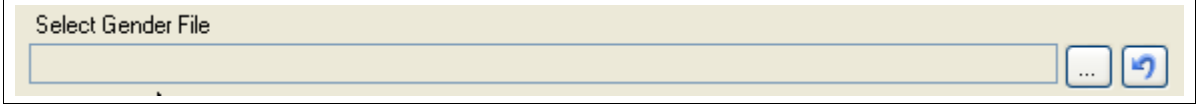


All CEL files need to be listed. Files without gender information should have a '0' in the gender column. Empty value will be treated as '0'.

To select a Gender file:

1. Make sure that the file you wish to select is in the GTC Library folder.
   You can copy a file to the folder.

Figure 7.32 Select Gender File



2. Enter the path and file name (Figure 7.32); or
   Click the **Browse** button and select the file from the dialog box.

Hints and In-bred Sample File Options

These options are only available for certain algorithms and arrays, as described below.

The Hints file and the In-Bred Sample file options are mutually exclusive. You may only choose one or the other, not both.

Hints Files

The Hints File allows you to refine the clustering performance of an algorithm by incorporating reference data. If some data points have known genotypes, the genotype cluster locations may be adapted towards clusters that reproduce the supplied genotypes (even if incorrect). The Hints file data may not change the cluster properties if the existing cluster properties are too strong, as the existing information will override the information in the Hints file.

The supplied genotypes are not used in making genotype calls. Only the resulting genotype cluster properties (i.e., the resulting posterior files) will be used to make genotype calls that can be exported.

You can use the Hints file option for:

- Axiom GT (human and bovine)
- BRLMM-P (Human, rat, and mouse)

The file uses the following format (Figure 7.33 and Table 7.5):

The first row in the file lists has the following headings:

- Probeset ID: Identifier for the SNP probe set
- CEL file Name: the Cel File the calls are provided for

Each following row lists a probe set ID and a genotyping call for each CEL file, using the following code:

Table 7.5 Hints code

Number	Call
-1	No call (in this case, no reference)
0	AA
1	AB
2	BB

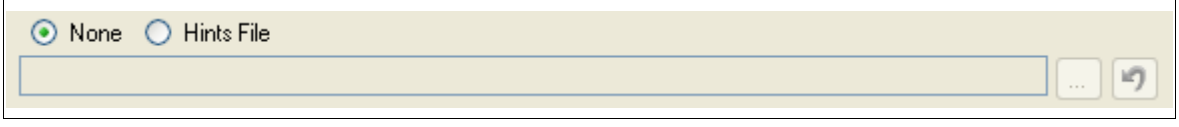
Figure 7.33 Hints file viewed in Microsoft Excel

	A	B	C	D	E
	probeset_id	NA06985.CEL	NA06991_	NA06993_	NA06994_
	AX-11088282	2	2	2	2
	AX-11092203	2	2	2	2
	AX-11157151	2	2	2	2
	AX-11161133	0	0	0	0
	AX-11203782	2	2	2	2
	AX-11225288	0	0	0	0
	AX-11227475	1	2	2	2
	AX-11286614	2	2	2	2

To select a Hints file:

1. Make sure that the file you wish to select is in the GTC Library folder.
   You can copy a file to the folder.
2. Click the Hints file button (Figure 7.34).

Figure 7.34 Select Hints File



3. Enter the path and file name; or

Click the **Browse** button and select the file from the dialog box.

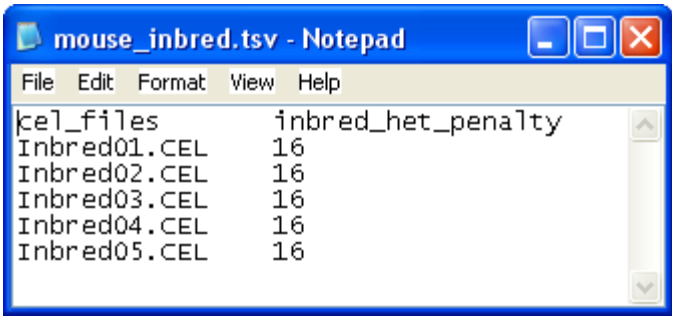
Inbred Sample File

The In-bred Sample File allows you to improve the clustering performance of an algorithm by providing additional information about the degree of increased homozygosity (or decreased heterozygosity) expected from in-bred samples.

This option is available only for non-human data (Axiom Bovine, Mouse, and Rat).

The inbred sample data is provided by the user in a TSV file (Figure 7.35).

Figure 7.35 In-bred Sample File format viewed in Microsoft Notepad



The In-bred Samples file uses the following format:

The first row in the file lists has the following headings:

- cel_files: CEL file that the penalty information is provided for
- inbred_het_penalty: The inbreeding penalty value that controls how much to bias clustering against having heterozygous calls in samples which may be inbred. “0” = no penalty (normal sample), 1 = mild penalty, 16 = maximum penalty (inbred sample). The Inbred Sample file shall include all the samples and each sample shall have an inbreeding penalty value (use 0 for normal samples).

Each following row lists a cell file name and penalty value for the file.

To select an inbred sample data file:

1. Make sure that the file you wish to select is in the GTC Library folder.
   You can copy a file to the folder.
2. Click the In-bred Sample File button (Figure 7.36).

Figure 7.36 Select In-bred Sample File

3. Enter the path and file name; or
   Click the **Browse** button and select the file from the dialog box.

CHP Summary Table

The CHP Summary Table contains a summary of the batch genotyping results.

See *Table Features* on page 198 for more information on customizing the table view.

The tables below provide definitions for items in the CHP Summary Table:

- Table 7.6, *CHP Summary table, items common to all arrays,* on page 111

Table 7.6 CHP Summary table, items common to all arrays

Item (common to all arrays)	Definition
File	File name.
computed_gender	Computed gender for the sample. For more information about the processes used to compute gender for the different array types, Appendix E, *Gender Calling in GTC* on page 353.
call_rate	BRLMM/BRLMM-P/Birdseed/Axiom call rate at the default or user-specified threshold for autosomal SNPs.
total-call_rate	BRLMM/BRLMM-P/Birdseed/Axiom call rate at the default or user-specified threshold for all SNPs.
het_rate	Percentage of SNPs called AB (i.e. the heterozygosity) for autosomal SNPs.
total_het_rate	Percentage of SNPs called AB (i.e., the heterozygosity) for all SNPs.
hom_rate	Percentage of SNPs called AA or BB (i.e. the homozygosity) for autosomal SNPs.
total_hom_rate	Percentage of SNPs called AA or BB (i.e. the homozygosity) for all SNPs.

NOTE: **Genotyping Console 4.1 uses a new method to calculate call_rate, het_rate and hom_rate metrics after genotyping. Instead of using all SNPs to calculate these metrics, the new method only uses autosomal SNPs to calculate these metrics. When using a SNP list for genotyping, only the autosomal SNPs included in the list will be used to calculate these metrics. For Axiom™ myDesign™ Genotyping Arrays and Axiom™ Genome-Wide CHB Array, these metrics will be calculated using the autosomal SNPs in the list plus the 3000 SNPQC SNPs. The old metrics calculated using all SNPs will be hosted under 3 new metrics named as total_call_rate, total_het_rate, and total_hom_rate respectively.**

The results will vary, depending on the array and the sample. Overall, if the array has chromosome Y SNPs, the call rates for female samples could improve slightly and the call rates for male samples could go down very slightly. If the array does not have chromosome Y SNPs, the call rates could go down very slightly for both male and female samples. This is because male and female samples have a tendency to have homozygous calls on X & Y chromosomes. Removing them could slightly reduce the homozygous call rate and therefore raise the heterozygous call rate and reduce the overall call rate.

Table 7.7 CHP Summary table, items common to all arrays

Item (common to all arrays)	Definition
cluster_distance_mean	Average distance to the cluster center for the called genotype.
cluster_distance_stdev	Standard deviation of the distance to the cluster center for the called genotype.
raw_intensity_mean	Average of the raw PM probe intensities.
raw_intensity_stdev	Standard deviation of the raw PM probe intensities.
allele_summarization_mean	Average of the allele signal estimates (log2 scale).
allele_summarization_stdev	Standard deviation of the allele signal estimates (log2 scale).

Table 7.7 CHP Summary table, items common to all arrays

Item (common to all arrays)	Definition
allele_deviation_mean	Average of the absolute difference between the log2 allele signal estimate and its median across all arrays.
allele_deviation_stdev	Standard deviation of the absolute difference between the log2 allele signal estimate and its median across all arrays.
allele_mad_residuals_mean	Average of the median absolute deviation (MAD) between observed probe intensities and probe intensities fitted by the model.
allele_mad_residuals_stdev	Standard deviation of the median absolute deviation (MAD) between observed probe intensities and probe intensities fitted by the model.
em cluster chrX het contrast_gender	Gender call made by the em-cluster-chrX-het-contrast_gender method. This method estimates the heterozygosity rate (% AB genotypes) of SNPs on the X chromosome. If the heterozygosity is above a threshold, then the gender call is female, otherwise the gender call is male.
em cluster chrX het contrast_gender_chrX_het_rate	The estimated heterozygosity rate (% AB genotypes) of SNPs on the X chromosome.
pm_mean	Average of the PM probe signals.
File Date	The date and time the CHP file was last modified.
QC Call Rate	Computed QC Call Rate for all QC SNPs (not available for Axiom).

Table 7.8 CHP Summary table, items for 100K/500K arrays

Item (100K/500K)	Definition
dm chrX het rate_gender	Gender call based on ChrX Het rate using DM calls.
dm chrX het rate_gender_chrX_het_rate	The DM based ChrX Het rate from which the gender call is based.
dm listener call rate	DM call rate.

NOTE: For Human Mapping 100K/500K arrays, the CHP Summary data for the different array types (used for different enzyme sets) will be displayed in separate tables. Each table will have the appropriate array type appended to its base batch name. Separate results sets are displayed in the Genotype Result, as shown in the figure below (Figure 7.37).

Figure 7.37 Genotype results files for Human Mapping 500K arrays with paired enzyme sets

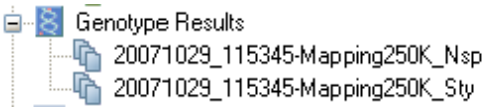


Table 7.9 CHP Summary table, items for SNP 5 arrays

Item (SNP 5)	Definition
QC Call Rate (NSP)	Computed QC Call Rate (via DM algorithm) for SNPs located only on NSP restriction fragments.
QC Call Rate (Nsp/Sty Overlap)	Computed QC Call Rate (via DM algorithm) for SNPs located on both NSP and STY restriction fragments.
QC Call Rate (Sty)	Computed QC Call Rate (via DM algorithm) for SNPs located only on STY restriction fragments.

Table 7.10 CHP Summary table, items for SNP 6 arrays

Item (SNP 6)	Definition
QC cn probe chrXY ratio_gender_meanX	The average probe intensity (raw, untransformed) of X chromosome nonpolymorphic probes.
QC cn probe chrXY ratio_gender_meanY	The average probe intensity (raw, untransformed) of Y chromosome nonpolymorphic probes.
QC cn probe chrXY ratio_gender_ratio	Gender ratio Y/X = cn probe chrXY-ratio_gender_meanY/ cn probe chrXY ratio_gender_meanX.
QC Computed Gender	Computed gender. For more details, see Appendix E, *Gender Calling in GTC* on page 353. Gender calls made by the cn-probe-chrXY-ratio_gender method. If the cn-probe-chrXY-ratio_gender_ratio is less than the lower cutoff the gender call is female. If the cn-probe-chrXY-ratio_gender_ratio is greater than the upper cutoff, then the gender call is male. If the cn-probe-chrXY-ratio_gender_ratio is between the lower and upper cutoffs, then the gender call is unknown.
Contrast QC	Computed Contrast QC for all QC SNPs.
Contrast QC (Random)	Contrast QC for 10K random autosomal SNPs.
Contrast QC (Nsp)	Contrast QC for QC 20K SNPs on Nsp fragments.
Contrast QC (Sty)	Contrast QC for QC 20K SNPs on Sty fragments.
Contrast QC (Nsp/Sty Overlap)	Contrast QC for QC 20K SNPs on both an Nsp and Sty fragment.
QC Call Rate (NSP)	Computed QC Call Rate (via DM algorithm) for SNPs located only on NSP restriction fragments.
QC Call Rate (Nsp/Sty Overlap)	Computed QC Call Rate (via DM algorithm) for SNPs located on both NSP and STY restriction fragments.
QC Call Rate (Sty)	Computed QC Call Rate (via DM algorithm) for SNPs located only on STY restriction fragments.

Table 7.11 CHP Summary table, items for Axiom Arrays

Item (Axiom)	Definition
QC cn probe chrXY ratio_gender_meanX	The average probe intensity (raw, untransformed) of X chromosome nonpolymorphic probes.
QC cn probe chrXY ratio_gender_meanY	The average probe intensity (raw, untransformed) of Y chromosome nonpolymorphic probes.
QC cn probe chrXY ratio_gender_ratio	Gender ratio Y/X = cn probe chrXY-ratio_gender_meanY/ cn probe chrXY ratio_gender_meanX.
QC Computed Gender	Computed gender. For more details, see Appendix E, *Gender Calling in GTC* on page 353. Gender calls made by the cn-probe-chrXY-ratio_gender method. If the cn-probe-chrXY-ratio_gender_ratio is less than the lower cutoff the gender call is female. If the cn-probe-chrXY-ratio_gender_ratio is greater than the upper cutoff, then the gender call is male. If the cn-probe-chrXY-ratio_gender_ratio is between the lower and upper cutoffs, then the gender call is unknown.
QC axiom_signal_contrast_AT_B_IQR	Interquartile range of control GC probe raw intensities (background intensities) in the AT channel.
QC axiom _signal_contrast_AT_B	Mean of the control GC probe raw intensities (background intensities) in the AT channel.
QC AT Channel FLD	Linear Discriminant for signal and background in the AT channel, defined as (median_of_GC_probe_intensities – median_of_AT_probe_intensities)^2 / [0.5 * (Axiom_signal_contrast_AT_B_IQR^2 + Axiom_signal_contrast_AT_S_IQR^2)].
QC axiom_signal_contrast_AT_SBR	Signal to background ratio in the AT channel, defined as Axiom_signal_contrast_AT_S / Axiom_signal_contrast_AT_B.
QC axiom_signal_contrast_AT_S_IQR	The interquartile range of control AT probe raw intensities (signal intensities) in the AT channel.
Qc axiom_signal_contrast_AT_S	Mean of the control AT probe raw intensities (signal intensities) in the AT channel.
QC axiom_signal_contrast_A_signal_mean	Mean of the control A probe raw intensities in the AT channel.
QC axiom_signal_contrast_C_signal_mean	Mean of the control C probe raw intensities in the GC channel.
QC axiom_signal_contrast_GC_B_IQR	The interquartile range of control AT probe raw intensities (background intensities) in the GC channel.
QC Axiom_signal_contrast_GC_B	Mean of control AT probe raw intensities (background intensities) in the GC channel.
QC GC Channel FLD	Linear Discriminant for signal and background in the GC channel, defined as (median_of_GC_probe_intensities – median_of_AT_probe_intensities)^2 / [0.5 * (Axiom_signal_contrast_GC_B_IQR^2 + Axiom_signal_contrast_GC_S_IQR^2)].
QC Axiom_signal_contrast_GC_SBR	Signal to background ratio in the GC channel, defined as Axiom_signal_contrast_GC_S / Axiom_signal_contrast_GC_B.
QC axiom_signal_contrast_GC_S_IQR	Interquartile range of control GC probe raw intensities (signal intensities) in the GC channel.
QC Axiom_signal_contrast_GC_S	Mean of control GC probe raw intensities (signal intensities) in the GC channel.
QC axiom_signal_contrast_G_signal_mean	Mean of the control G probe raw intensities in the GC channel.
QC axiom_signal_contrast_T_signal_mean	Mean of the control T probe raw intensities in the AT channel.

Table 7.11 CHP Summary table, items for Axiom Arrays

Item (Axiom)	Definition
Dish QC	A QC metric that evaluates the overlap between the two homozygous peaks (AT versus GC) using normalized intensities of control non-polymorphic probes from both channels. It is defined as the fraction of AT probes not within two standard deviations of the GC probes in the contrast space.
SNP QC call rate	Call rate for approximately 3000 SNPs that Affymetrix provides as positive controls on custom arrays (Axiom myDesign only).
Log Difference QC	A cross channel QC metric, defined as mean(log(AT_SBR))/std(log(AT_SBR)) + mean(log(GC_SBR))/std(log(GC_SBR)), where signal and background are calculated for control non-polymorphic probes after intensity normalization.
QC axiom_varscore_CV_GC	Median of the coefficient of variation for each control GC probe set in the GC channel.
QC axiom_varscore_CV_AT	Median of the coefficient of variation for each control AT probe set in the AT channel.

In addition to the tabular display of the metrics, the CHP results can be displayed in a line graph.

To open a line graph:

- Click the line graph shortcut on the CHP Summary table tool bar.
  See *Graph Features* on page 203 for more information.

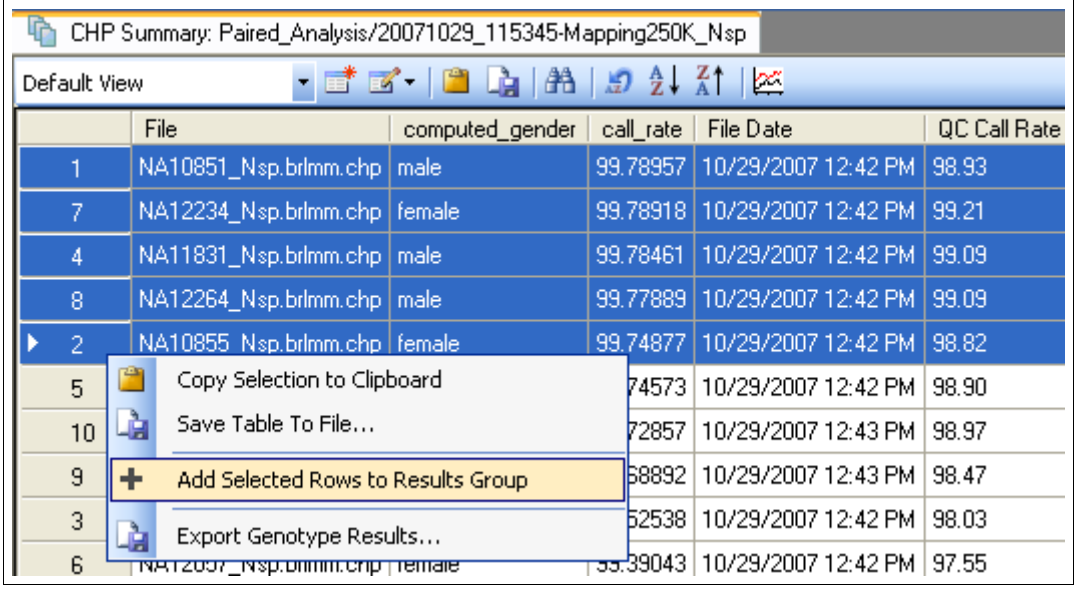
Creating Genotyping Results Custom Groups

Genotyping Console enables you to create custom groupings of genotyping results.

To make a custom group of genotyping results:

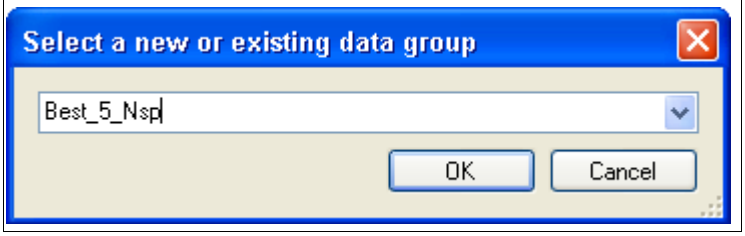
1. Select the row(s) from an open CHP Summary table to be added to the new group (Figure 7.38). Rows can be selected individually or by call rate or other parameter in the table.
2. Right-click and select **Add Selected Rows to Results Group** (Figure 7.38).

Figure 7.38 Add Selected Rows to Results Group



The Select a new or existing data group dialog box appears (Figure 7.39).

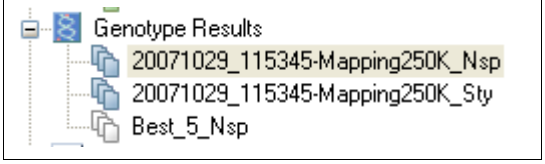
Figure 7.39 Select new or existing data group



3. Enter a name or select an existing data group and select **OK**.

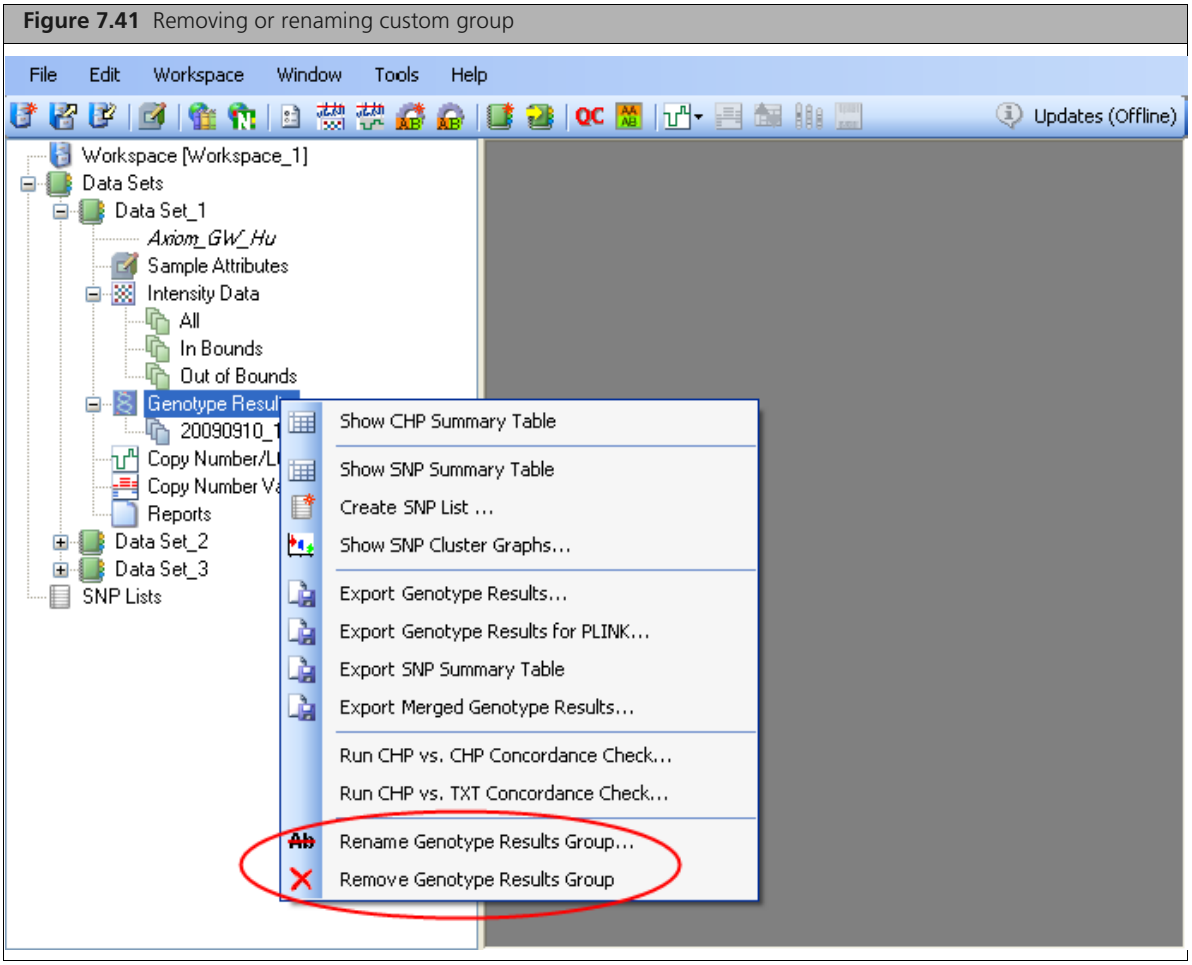
   The new genotype results custom group will be displayed in the tree (Figure 7.40). Custom groups are indicated by white icons.

Figure 7.40 Custom Group in tree



Custom results groups can be renamed or deleted by right-clicking the group and selecting **Rename Genotype Results Group** or **Remove Genotype Results Group** (Figure 7.41).

Figure 7.41 Removing or renaming custom group



NOTE: Removing a custom Genotyping Results Group does not remove the data from the Data Set. To remove Genotype Results data, see ***Removing Data from a Data Set*** on page 60.

NOTE: If a custom Genotype Results group is selected for displaying SNP summary results or SNP cluster graphs, the first time the SNP summary table or SNP cluster graph is generated, Genotyping Console will prompt you to save the summary statistics file.

Creating a Custom Intensity Group from the CHP File Data

You can create a custom intensity group from the CHP summary table using information in the CHP files. This is useful when performing the two-step genotyping workflow.

The creation of custom groups is based on checks implemented using the properties listed below. These properties need to match for the CHPs being added to the custom group in the following order:

- Algorithm family (always present)
- Array type (always present)
- CDF GUID (Can be 'Default' or the GUID from the CDF file)
- Probelist Checksum (Can be None if no probelist file is used or the MD5 checksum of the used probelist file)

There are two methods to do this:

- *Creating a Custom Intensity Data Group Using the CHP Summary Table* on page 117
- *Creating a Custom Intensity Data Group Using Thresholds Filtering* on page 120

You can also use either of the following options to create a custom intensity data group:

- *Creating a Custom Intensity Group from the CHP File Data* on page 117
- *Creating Custom Intensity Data Groups Using the SNP Cluster Graph* on page 162

The custom intensity group can be used for the *Two-Step Genotyping Workflow* on page 123.

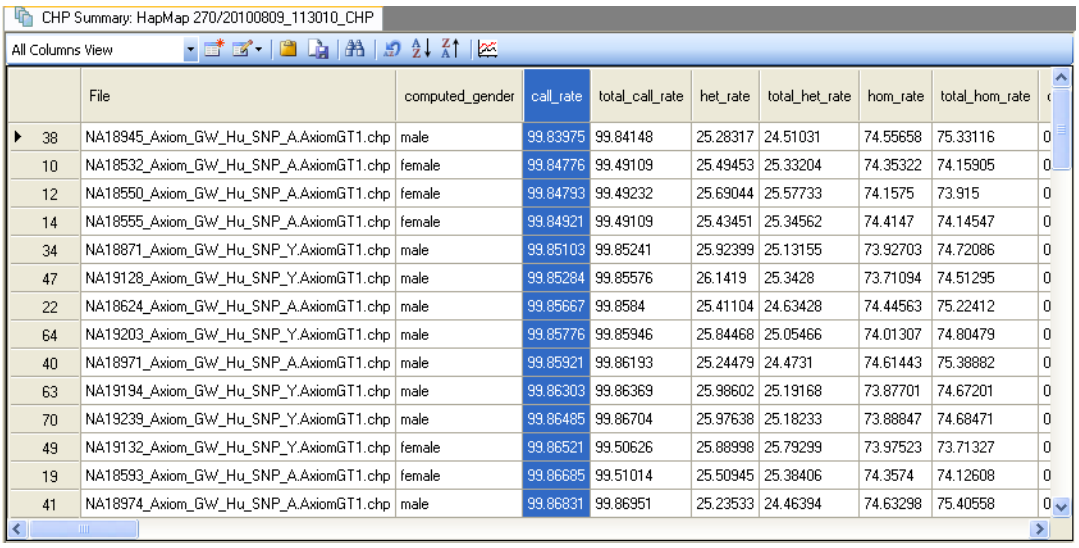
Creating a Custom Intensity Data Group Using the CHP Summary Table

To create a custom intensity data group using the CHP Summary Table:

1. Group the array files you wish to exclude or include using the Sort functions of the CHP Summary Table tool bar.

   A. Click in the column header for the column you wish to sort by (Figure 7.42).

Figure 7.42 CHP Summary Table with Call Rate column selected



	File	computed_gender	call_rate	total_call_rate	het_rate	total_het_rate	hom_rate	total_hom_rate
38	NA18945_Axiom_GW_Hu_SNP_A.AxiomGT1.chp	male	99.83975	99.84148	25.28317	24.51031	74.55658	75.33116
10	NA18532_Axiom_GW_Hu_SNP_A.AxiomGT1.chp	female	99.84776	99.49109	25.49453	25.33204	74.35322	74.15905
12	NA18550_Axiom_GW_Hu_SNP_A.AxiomGT1.chp	female	99.84793	99.49232	25.69044	25.57733	74.1575	73.915
14	NA18555_Axiom_GW_Hu_SNP_A.AxiomGT1.chp	female	99.84921	99.49109	25.43451	25.34562	74.4147	74.14547
34	NA18871_Axiom_GW_Hu_SNP_Y.AxiomGT1.chp	male	99.85103	99.85241	25.92399	25.13155	73.92703	74.72086
47	NA19128_Axiom_GW_Hu_SNP_Y.AxiomGT1.chp	male	99.85284	99.85576	26.1419	25.3428	73.71094	74.51295
22	NA18624_Axiom_GW_Hu_SNP_A.AxiomGT1.chp	male	99.85667	99.8584	25.41104	24.63428	74.44563	75.22412
64	NA19203_Axiom_GW_Hu_SNP_Y.AxiomGT1.chp	male	99.85776	99.85946	25.84468	25.05466	74.01307	74.80479
40	NA18971_Axiom_GW_Hu_SNP_A.AxiomGT1.chp	male	99.85921	99.86193	25.24479	24.4731	74.61443	75.38882
63	NA19194_Axiom_GW_Hu_SNP_Y.AxiomGT1.chp	male	99.86303	99.86369	25.98602	25.19168	73.87701	74.67201
70	NA19239_Axiom_GW_Hu_SNP_Y.AxiomGT1.chp	male	99.86485	99.86704	25.97638	25.18233	73.88847	74.68471
49	NA19132_Axiom_GW_Hu_SNP_Y.AxiomGT1.chp	female	99.86521	99.50626	25.88998	25.79299	73.97523	73.71327
19	NA18593_Axiom_GW_Hu_SNP_A.AxiomGT1.chp	female	99.86685	99.51014	25.50945	25.38406	74.3574	74.12608
41	NA18974_Axiom_GW_Hu_SNP_A.AxiomGT1.chp	male	99.86831	99.86951	25.23533	24.46394	74.63298	75.40558

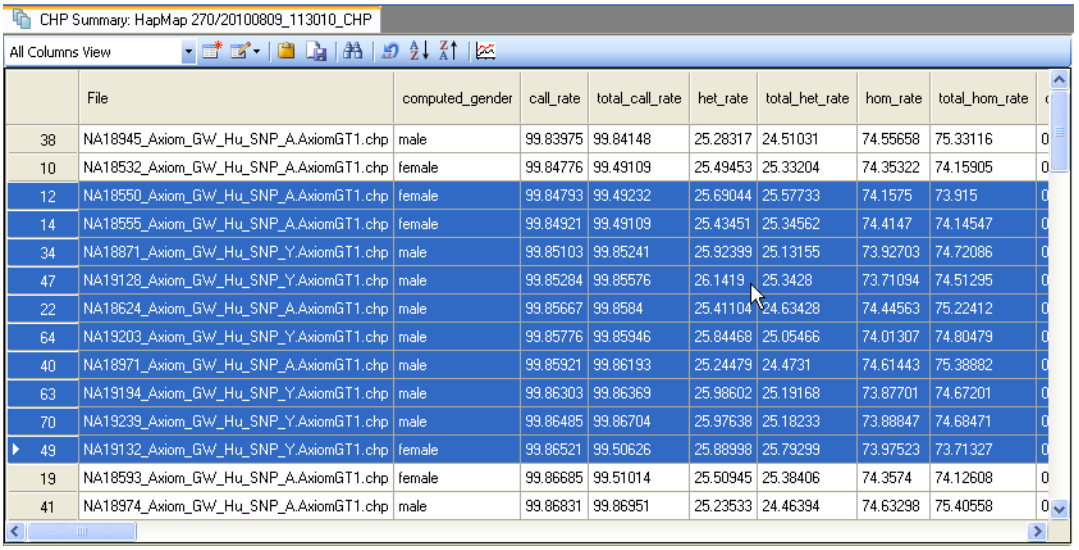
B. Sort the table by clicking the Sort Ascending or Sort Descending buttons in the table tool bar.

The files are sorted by the column parameter.

2. Select the rows in the table that you wish to include or exclude.

You must select the rows by clicking in the rows label column. (Figure 7.43)

Figure 7.43 Selecting CHP files



	File	computed_gender	call_rate	total_call_rate	het_rate	total_het_rate	hom_rate	total_hom_rate
38	NA18945_Axiom_GW_Hu_SNP_A.AxiomGT1.chp	male	99.83975	99.84148	25.28317	24.51031	74.55658	75.33116
10	NA18532_Axiom_GW_Hu_SNP_A.AxiomGT1.chp	female	99.84776	99.49109	25.49453	25.33204	74.35322	74.15905
12	NA18550_Axiom_GW_Hu_SNP_A.AxiomGT1.chp	female	99.84793	99.49232	25.69044	25.57733	74.1575	73.915
14	NA18555_Axiom_GW_Hu_SNP_A.AxiomGT1.chp	female	99.84921	99.49109	25.43451	25.34562	74.4147	74.14547
34	NA18871_Axiom_GW_Hu_SNP_Y.AxiomGT1.chp	male	99.85103	99.85241	25.92399	25.13155	73.92703	74.72086
47	NA19128_Axiom_GW_Hu_SNP_Y.AxiomGT1.chp	male	99.85284	99.85576	26.1419	25.3428	73.71094	74.51295
22	NA18624_Axiom_GW_Hu_SNP_A.AxiomGT1.chp	male	99.85667	99.8584	25.41104	24.63428	74.44563	75.22412
64	NA19203_Axiom_GW_Hu_SNP_Y.AxiomGT1.chp	male	99.85776	99.85946	25.84468	25.05466	74.01307	74.80479
40	NA18971_Axiom_GW_Hu_SNP_A.AxiomGT1.chp	male	99.85921	99.86193	25.24479	24.4731	74.61443	75.38882
63	NA19194_Axiom_GW_Hu_SNP_Y.AxiomGT1.chp	male	99.86303	99.86369	25.98602	25.19168	73.87701	74.67201
70	NA19239_Axiom_GW_Hu_SNP_Y.AxiomGT1.chp	male	99.86485	99.86704	25.97638	25.18233	73.88847	74.68471
49	NA19132_Axiom_GW_Hu_SNP_Y.AxiomGT1.chp	female	99.86521	99.50626	25.88998	25.79299	73.97523	73.71327
19	NA18593_Axiom_GW_Hu_SNP_A.AxiomGT1.chp	female	99.86685	99.51014	25.50945	25.38406	74.3574	74.12608
41	NA18974_Axiom_GW_Hu_SNP_A.AxiomGT1.chp	male	99.86831	99.86951	25.23533	24.46394	74.63298	75.40558

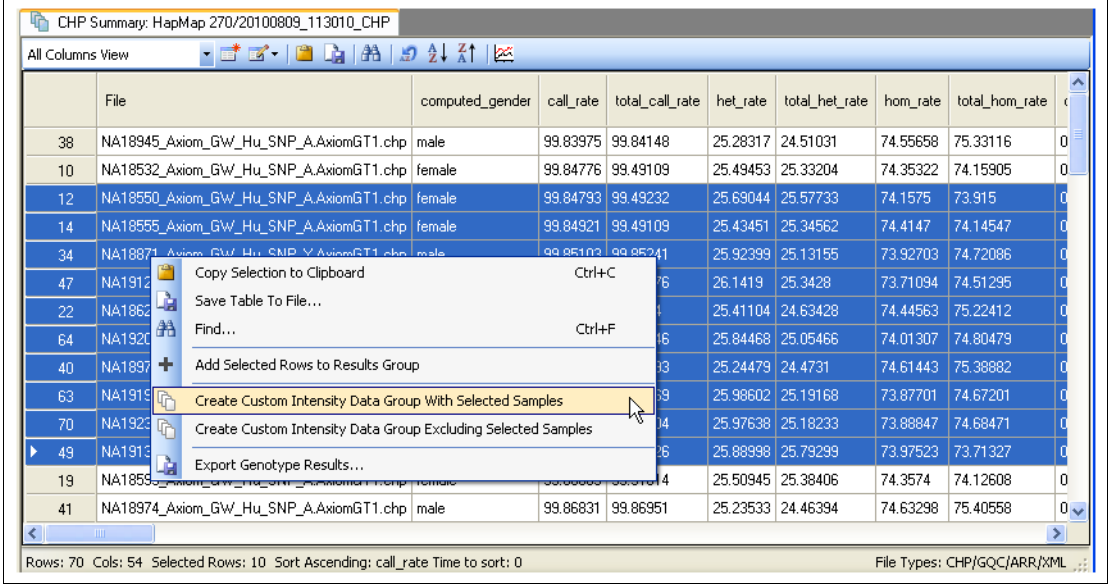
Select contiguous rows by clicking in the top and bottom rows while holding down the Shift key

Select multiple non-contiguous roles by clicking in the rows while holding down the CTRL key.

3. Right click on the selected cells and select the desired option from the menu (Figure 7.44):

- **Create Custom Intensity Group With Selected Results**
- **Create Custom Intensity Group Excluding Selected Samples**

Figure 7.44 CHP files selected in table and right-click menu

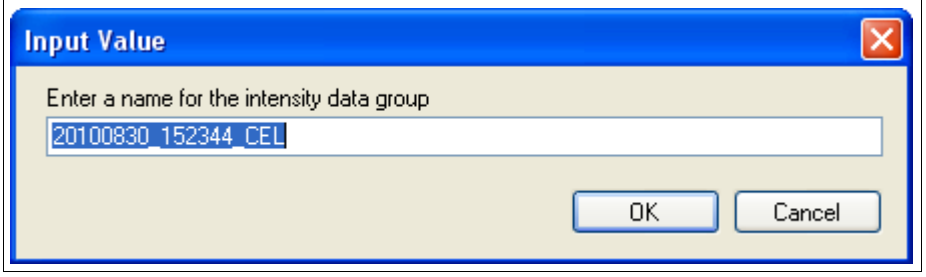


If you have selected improperly, you will see the following error message (Figure 7.45):

Figure 7.45 Error message when rows not selected properly

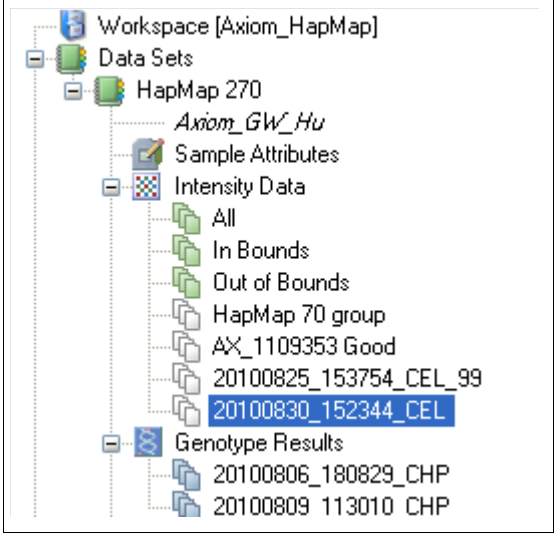
If you have selected properly, the Custom Intensity Group Name Dialog box opens (Figure 7.46).

Figure 7.46 Custom Intensity Data Group Name dialog box



4. Enter a name for the intensity group and click **OK**.
   You can also use the default name that appears in the dialog box.
   The new group is displayed in the data tree. Custom Groups are indicated by white icons (Figure 7.47).

Figure 7.47 Data tree displaying regular (in green) and custom (in white) data intensity groups



Custom Intensity Groups can be re-named by right-clicking on the group and selecting Rename Intensity Data Group.

Custom Intensity groups can be deleted by right-clicking on the group and selecting Remove Intensity Data Group.

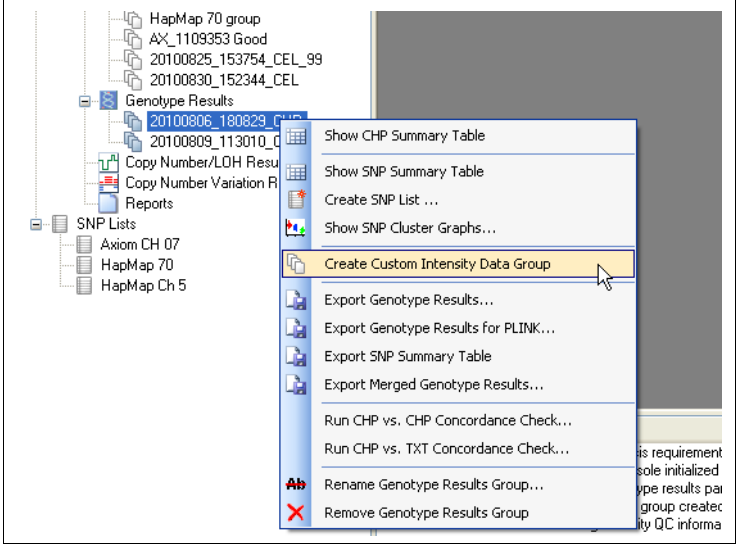
Creating a Custom Intensity Data Group Using Thresholds Filtering

You can also use thresholds filtering on different metrics to create an intensity data group.

To create an intensity data group using thresholds filtering:

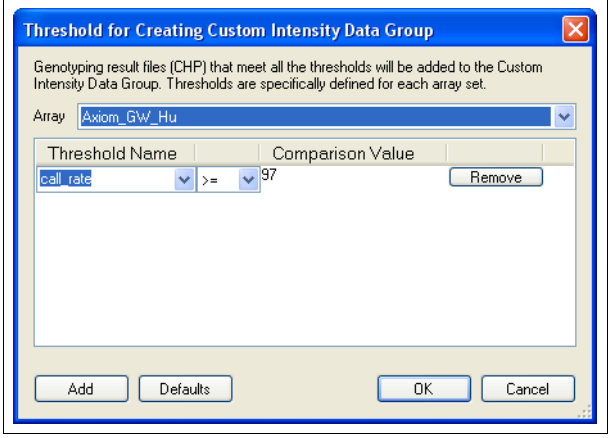
1. Right-click on the Genotype Results set you wish to filter and select **Create Custom Intensity Group** from the right-click menu (Figure 7.48).

Figure 7.48 Data Tree and Right-click menu



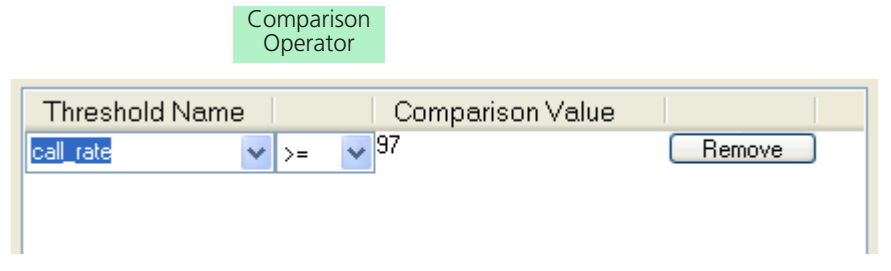
The Threshold dialog box opens (Figure 7.49).

Figure 7.49 Threshold dialog box



2. Select the metric, the comparison operator (less than (<), less than or equal to (≤), greater than (>), greater than or equal to (≥), equal to (=), or not equal to (!=)), and the value (Figure 7.50).

Figure 7.50 Threshold box



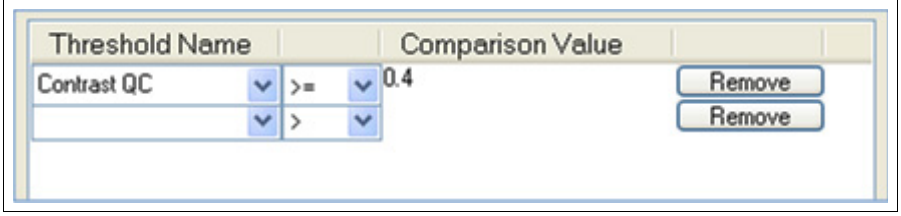
To use a different metric, select the text in the "Threshold Name" filed and type the exact name, case-sensitively, of the new metric in this field. For metrics to be applied, they must exist in the *CHP Summary Table* on page 110 when All Columns View is selected in the table.

3. Enter a new Threshold Name if desired:

   A. Click **Add**.

   A new row appears in the dialog box (Figure 7.51).

Figure 7.51 New Row Added



   B. Enter a new threshold name. The metric must exist in the *CHP Summary Table* on page 110 when All Columns View is selected in the table.

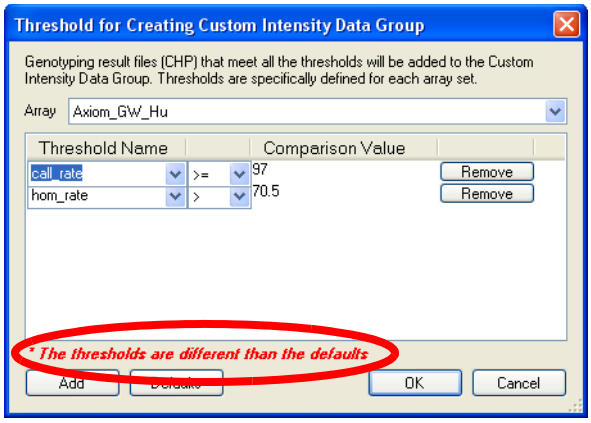
   C. Select a comparison operator.

   D. Enter the comparison value.

   If you enter more than one threshold, the samples must meet both thresholds to be included in the new group.

4. To delete a threshold item, click **Remove**.
   A notice appears in the dialog box when you have changed the thresholds (Figure 7.52).

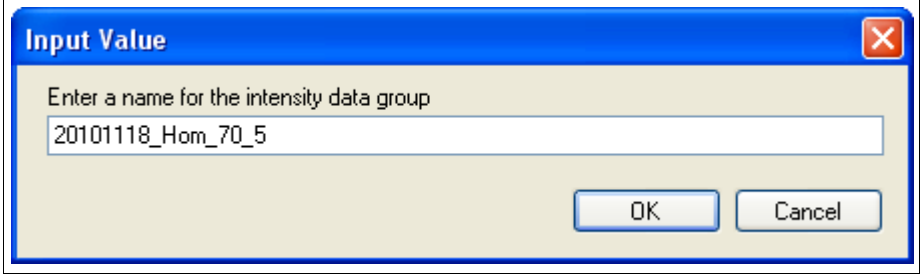
Figure 7.52 Flag in the Thresholds dialog box indicating that thresholds have been changed



   The threshold name must be an algorithm attribute; you cannot filter on sample or user attributes.

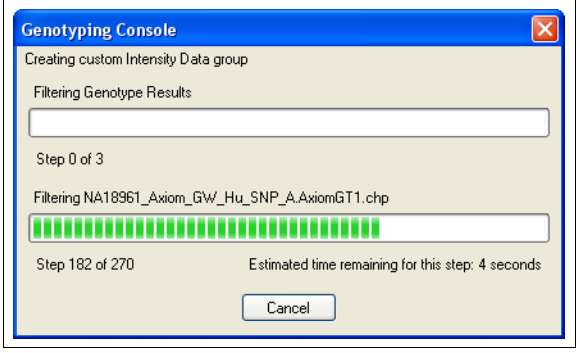
5. Click **OK**.
   The Enter a name for the intensity data group dialog box opens (Figure 7.53).

Figure 7.53 Enter a name for the intensity data group dialog box



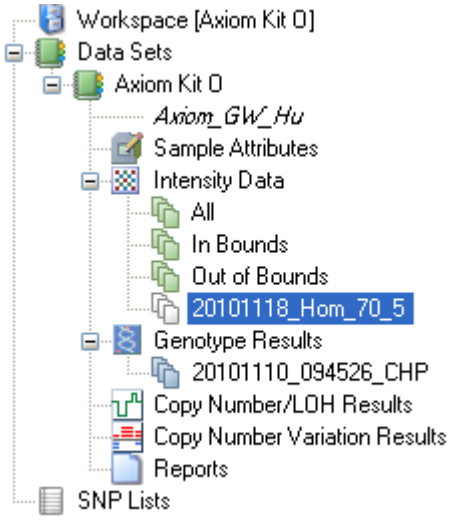
6. Enter a name and click **OK**.
   The Progress box displays the progress of the filtering operation (Figure 7.54).

Figure 7.54 Input Value



When the filtering operation is finished, the new intensity group is displayed in the data tree with a white icon (Figure 7.55).

Figure 7.55 New intensity data group in data tree



Two-Step Genotyping Workflow

The Two-Step Genotyping workflow helps maximize the quality of the resulting genotypes by implementing a workflow with two different QC steps to obtain optimal call rates when working with genotyping data.

The two-step workflow requires two different QC steps:

1. Remove samples based on single sample intensity QC metrics, such as Dish QC, as described in Chapter 6, *Intensity Quality Control for Genotyping Analysis* on page 74.
2. Perform a first round of genotyping, as described in *Performing Genotyping Analysis* on page 89.
3. Remove the outlier samples with call rates (for Axiom, Affymetrix recommends using < 97% as a cutoff) by using either of the following methods:
   - *Creating a Custom Intensity Data Group Using the CHP Summary Table* on page 117
   - *Creating a Custom Intensity Data Group Using Thresholds Filtering* on page 120)
4. Perform a second round of genotyping on the remaining samples.

Review the Genotyping Results

This chapter describes the options for performing an initial review of the genotyping results.

It contains the following sections:

Genotyping QC Steps

Before conducting downstream analysis of genotyping results it is essential to perform thorough QC of both SNPs and samples. There is no single 'best' way to do the QC, but some steps that are generally helpful in a broad range of circumstances are outlined below.

1. Per-sample QC filtering
   - Pre-clustering
     - Samples failing the per-array QC metric should be excluded prior to clustering, as described in Chapter 6, *Intensity Quality Control for Genotyping Analysis* on page 74.
     - Sample swaps which may have occurred during handling should be identified and resolved or removed. One way to do this is to generate a 'fingerprint' by typing all samples on a subset of a dozen or more SNPs which intersect with the SNPs reported in the *Signature Genotypes* on page 87. Another is to use known pedigree information (where appropriate) to confirm expected relatedness patterns.
   - Post-clustering
     - Remove samples with outlier clustering call rates or heterozygosity (which will tend to be low-performing samples that escaped the QC call rate filter).
     - Depending on the downstream analysis to be applied, consider identifying any cryptic relatedness and removing related samples.
     - Depending on the downstream analysis to be applied, consider controlling for population structure possibly be removing samples that are clearly from different populations from the bulk of the collection.
2. Per-SNP QC filtering
   - Remove SNPs with per-SNP call rates (sometimes referred to as completeness) less than some threshold. Commonly-used values for the per-SNP call rate threshold range from 90% to 95%.
   - Consider removing SNPs with minor allele frequency (MAF) below a certain threshold (for example, 1%).
   - Depending on circumstances, consider removing SNPs significantly out of Hardy Weinberg equilibrium in cases and/or controls. A p-value threshold in the range of 10^{-7} is sometimes used.

Once the genotyping results are generated, you can:

- *Create a SNP List* on page 125
- *Import a Custom SNP List* on page 130
- Display *SNP Summary Table* on page 131
- Perform *Concordance Checks* on page 139

You can also review the individual SNP calls in the SNP Cluster Graph (see Chapter 9, *Using the SNP Cluster Graph* on page 148).

Create a SNP List

For many genotyping applications, poorly performing SNPs can lead to an increase in false positives and a decrease in power. Such under-performing SNPs can be caused by systematic or sporadic errors that occur due to stochastic, sample, or experimental factors. Prior to downstream analysis it is prudent to apply some SNP filtering criteria to remove SNPs that are not performing ideally in the data set in question.

The subject of SNP filtering is an area of current research and best practices are still being developed by the community. Some common filters used will:

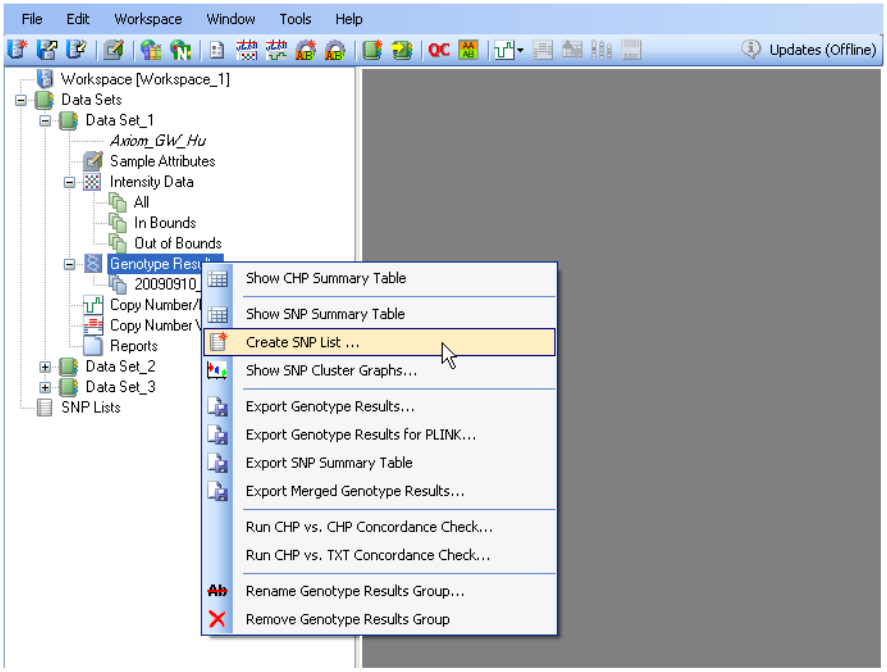
- Remove SNPs with a significantly low per SNP call rate
- Remove SNPs significantly out of HW equilibrium in cases and/or controls
- Remove SNPs with significantly different call rates in cases and controls
- Remove SNPs with Mendelian errors

Studies on multiple data sets have shown that SNPs with a lower per SNP call rate tend to have a higher error rate, and disproportionately contribute to the overall error rate in the experiment. Most importantly, though they may constitute a very small fraction of the total pool of SNPs, if the errors happen to stratify by case/control status then these low per-SNP call rate SNPs are more likely to show up as apparent associations.

To create a SNP list for filtering SNPs:

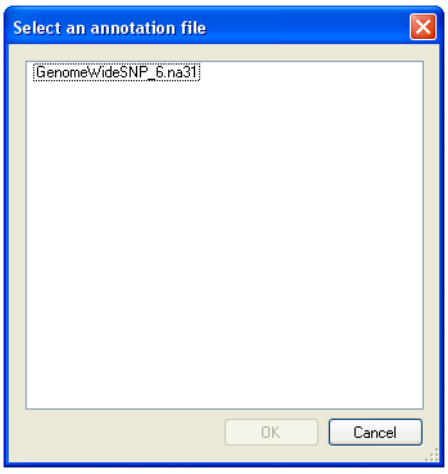
1. Right-click a genotyping batch results and select **Create SNP List** (Figure 8.1).

Figure 8.1 Creating a SNP list to enable per-SNP QC



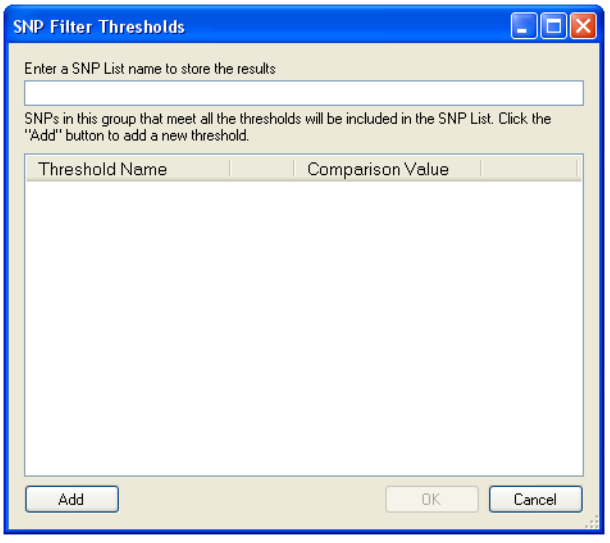
The Select an annotation file dialog box opens (Figure 8.2).

Figure 8.2 Select an annotation file dialog box



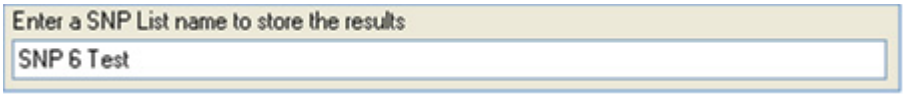
2. Select the annotation file to be used with the list and click OK. The SNP Filters Threshold window box opens (Figure 8.3).

Figure 8.3 SNP Filter Thresholds



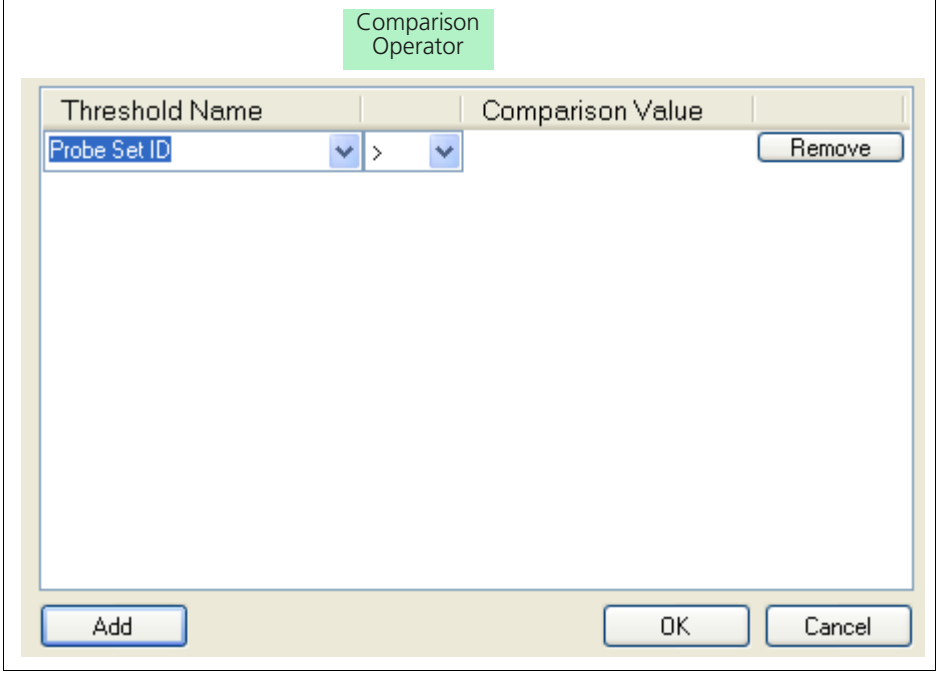
3. Enter a name for the SNP List (Figure 8.4).

Figure 8.4 Enter a SNP List name



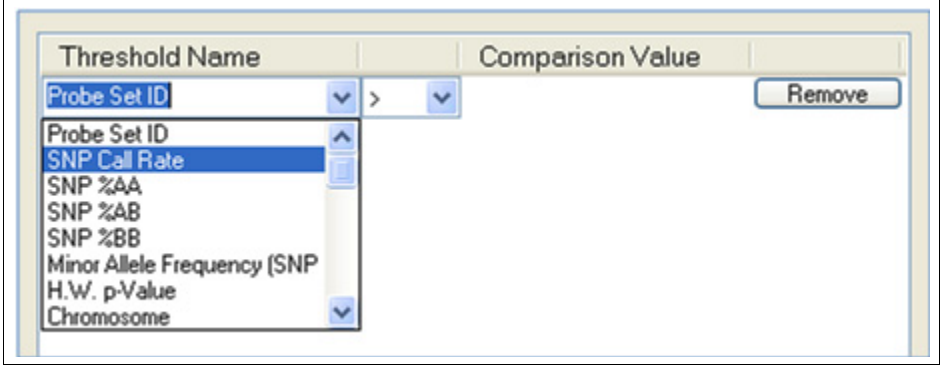
4. Click the **Add** button (Figure 8.5).

Figure 8.5 Adding a threshold



5. Select the Threshold Name from the drop-down list (Figure 8.6).

Figure 8.6 Selecting a Threshold Name

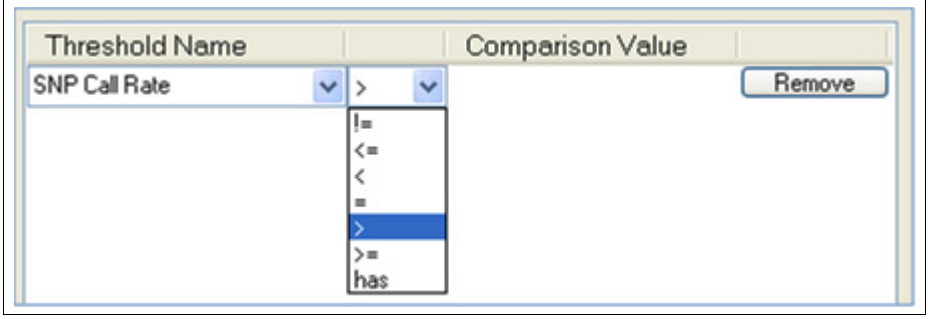


The list displays:

- Metrics displayed in the SNP Summary Table (see page 131)
- Some of the annotations in the annotations file

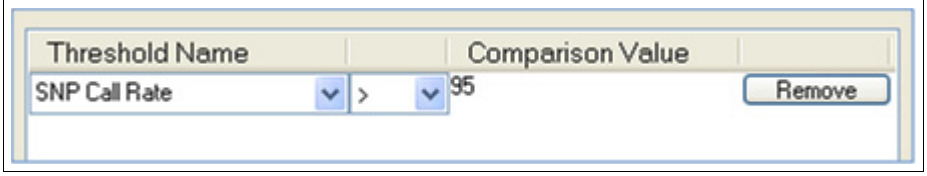
6. Choose the operator (e.g. =, >, has) (Figure 8.7). The “has” option is used when the category being filtered is text based (e.g. Associated Gene, In HapMap, etc.).

Figure 8.7 Selecting a comparison operator



7. Enter a Comparison Value (e.g. 99, YES, etc.) (Figure 8.8).

Figure 8.8 Entering a comparison value



8. Repeat steps 4 through 7 to add another threshold; or

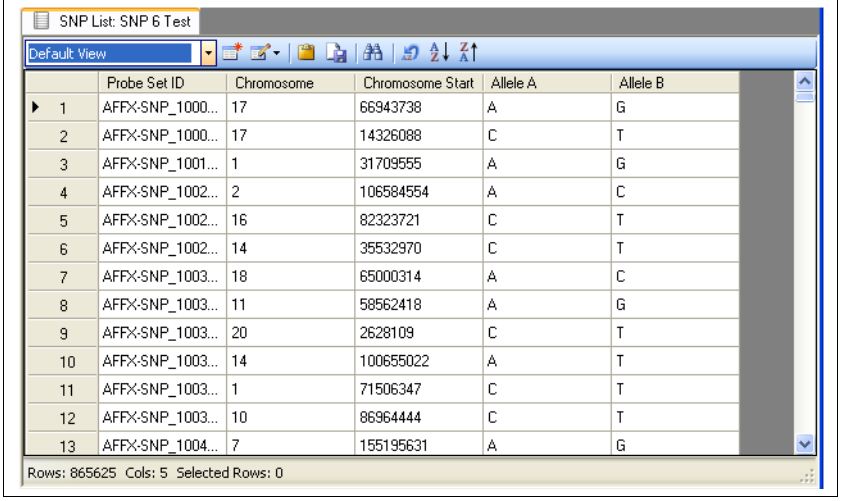
   Select **OK**.

   To remove filter criteria, select the **Remove** button.

   If you enter more than one threshold, the SNPs must meet all thresholds to be included in the new SNP list.

   The resulting SNP List will be automatically displayed (Figure 8.9). If some SNPs in the list have db NULL values, those SNPs will not be returned as results.

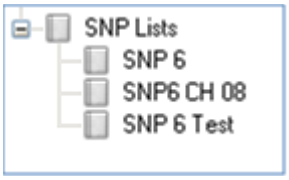
Figure 8.9 Custom SNP table



	Probe Set ID	Chromosome	Chromosome Start	Allele A	Allele B
1	AFFX-SNP_1000...	17	66943738	A	G
2	AFFX-SNP_1000...	17	14326088	C	T
3	AFFX-SNP_1001...	1	31709555	A	G
4	AFFX-SNP_1002...	2	106584554	A	C
5	AFFX-SNP_1002...	16	82323721	C	T
6	AFFX-SNP_1002...	14	35532970	C	T
7	AFFX-SNP_1003...	18	65000314	A	C
8	AFFX-SNP_1003...	11	58562418	A	G
9	AFFX-SNP_1003...	20	2628109	C	T
10	AFFX-SNP_1003...	14	100655022	A	T
11	AFFX-SNP_1003...	1	71506347	C	T
12	AFFX-SNP_1003...	10	86964444	C	T
13	AFFX-SNP_1004...	7	155195631	A	G

The list is added to the SNP List in the Tree (Figure 8.10).

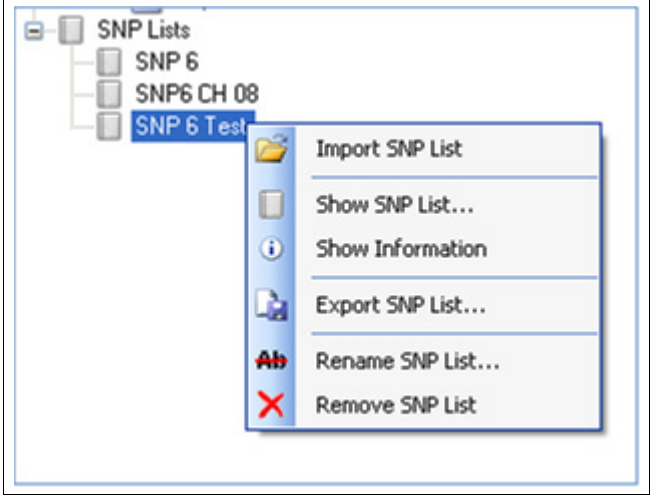
Figure 8.10 Custom SNP table and list in data tree



For more information on displaying data in SNP Lists see Chapter 11, *Table & Graph Features* on page 198.

SNP Lists can be exported, renamed, or removed by right-clicking on the SNP List and selecting the appropriate action (Figure 8.11).

Figure 8.11 SNP list menu



To view a SNP List, select the **Show SNP List** option. To review the filter criteria for a SNP List, select the **Show Information** option.

NOTE: When the criteria used to create a custom SNP list are unknown (e.g. an imported SNP List), the Show Information option will only indicate the SNP count.

NOTE: SNP lists are created based on a batch and the filters apply to the original batch on which they are based. For example, filtering by call rate on batch A will contain SNPs that pass this threshold. If this SNP list is used with a different batch, SNPs in the list may now demonstrate call rates below the threshold.

After creating a SNP List, you can apply any SNP List to generate a *SNP Cluster Graph* (page 124) or during *Export Genotype Results* (page 181).

Import a Custom SNP List

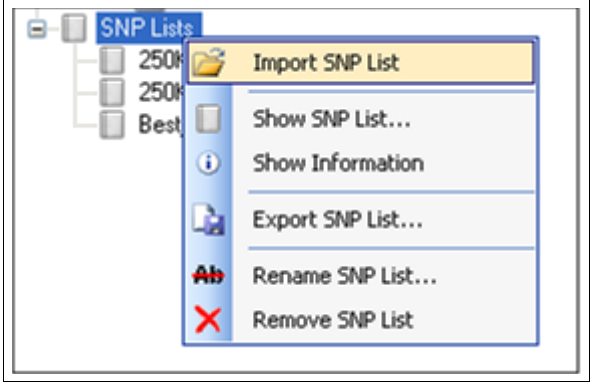
The Import Custom SNP List option enables you to import custom SNP lists that you may receive from other users.

The SNP List file must be a text file and contain a column labeled “Probe Set ID”. The file can contain additional columns although they will be ignored by the software. A SNP List can be generated by NetAffx: see the Advanced Workflow example *Analyzing Genotyping Results of Specific Gene Lists* on page 347.

To import a SNP List:

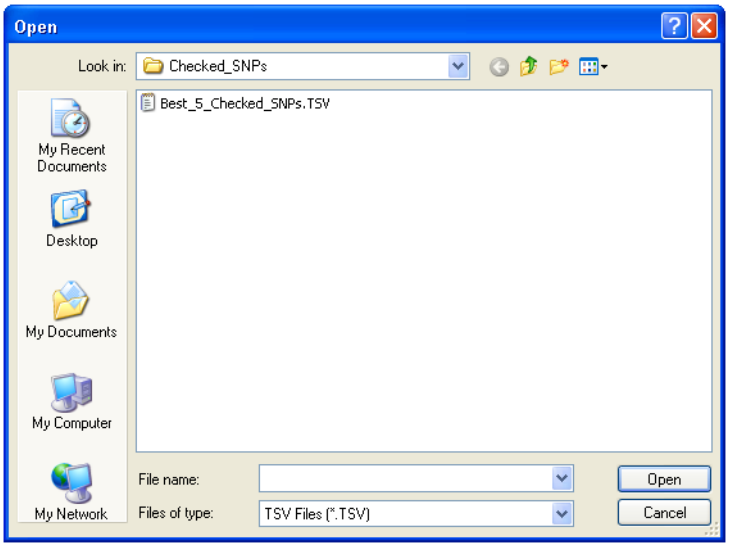
1. Right-click on SNP Lists in the data tree and select **Import SNP List** (Figure 8.12).

Figure 8.12 SNP List menu



The Open dialog box appears (Figure 8.13).

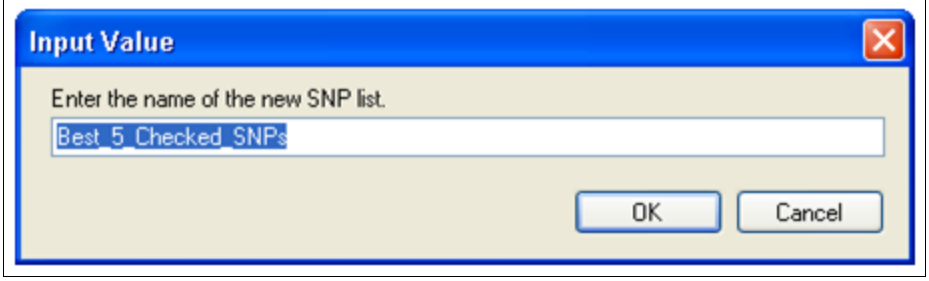
Figure 8.13 Open dialog box



2. Navigate to the location of the SNP List and select a list.
3. Click **Open**.

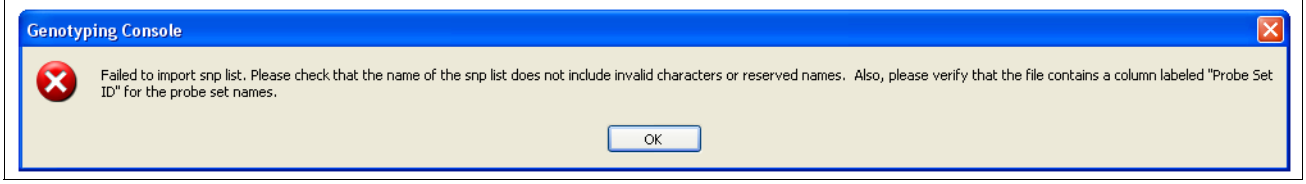
   The Input Value dialog box opens (Figure 8.14).

Figure 8.14 Input Value box for new SNP List name



4. Enter a name of the SNP List and click **OK**.
   If the import fails, the following notice appears (Figure 8.15):

Figure 8.15 Import Failure notice



   Click **OK** and correct the problem with the file.
   If the import succeeds, the SNP List will be displayed in the data tree. (Figure 8.16)

Figure 8.16 Imported SNP List

SNP Summary Table

IMPORTANT: You cannot display the SNP Summary Table until you have created a SNP list. See *Create a SNP List* on page 125 for more information.

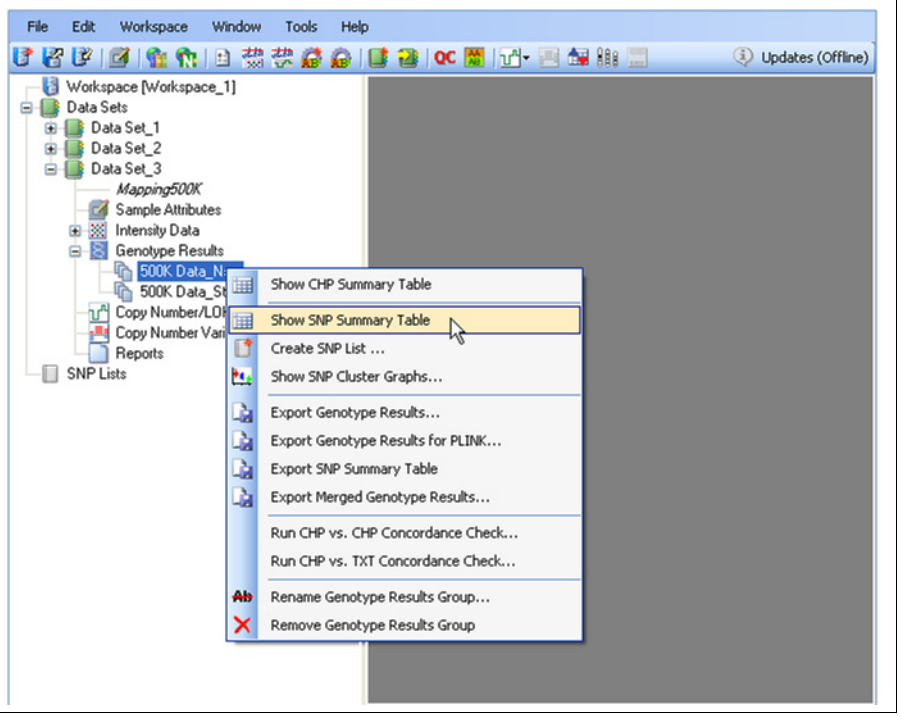
The SNP Summary Table contains SNP level statistics based on the batch of CHP files.

Genotyping Console stores the SNP summary information in a binary file. By generating this file, Genotyping Console can more quickly display the data each subsequent time the results are displayed. This file is usually generated during genotyping analysis, but if the CHP files were imported into GTC, or if the batch folder selected is for a newly created custom Genotype Results group, you will be prompted to save a SNP Statistics Summary file.

To open the SNP Summary Table:

1. Right-click a Genotype Results batch file in the GTC data tree and select **Show SNP Summary Table** (Figure 8.17).

Figure 8.17 Show SNP Summary Table

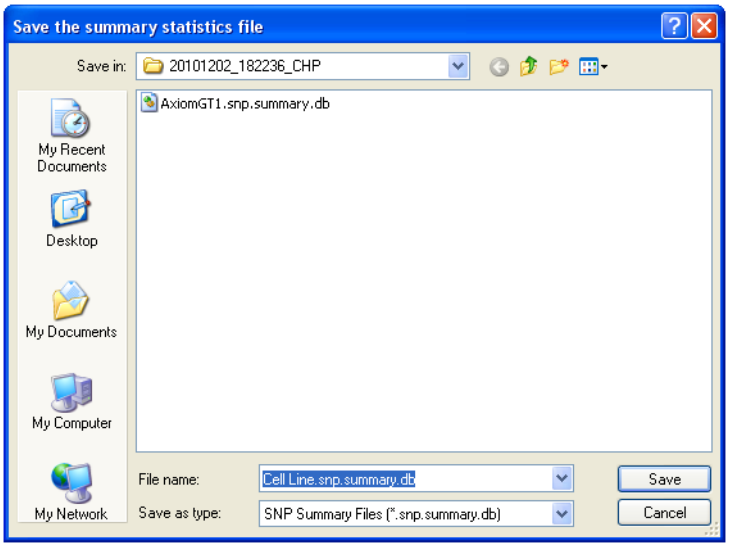


If the SNP Statistics have not been calculated for the CHP files, the SNP Statistics dialog box opens (Figure 8.18).

Figure 8.18 SNP Statistics Calculation dialog

2. Click **OK** in the SNP Statistics Calculation dialog box.
   The Save the summary statistics file dialog box opens (Figure 8.19).

Figure 8.19 Save the summary statistics file dialog box

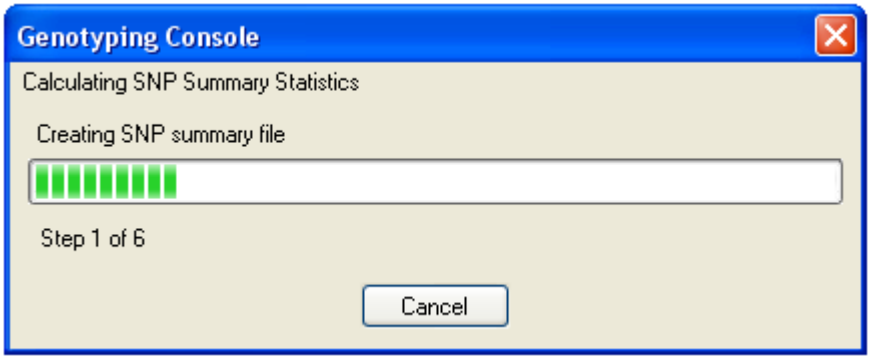


The dialog box is open to the batch results folder and prompts you to save a summary file with the name of the batch folder.

3. Click **Save** in the Save the summary Statistics file dialog box.

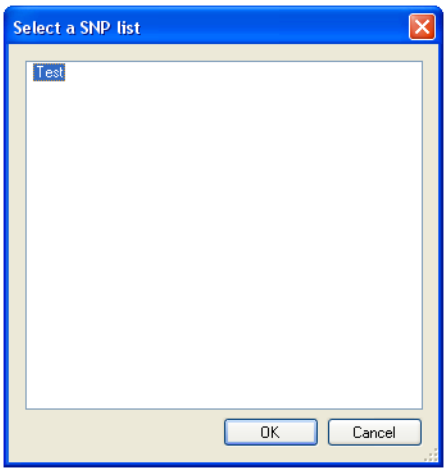
   The Calculating SNP Summary Statistics dialog box appears (Figure 8.20).

Figure 8.20 Calculating SNP Summary Statistics dialog box



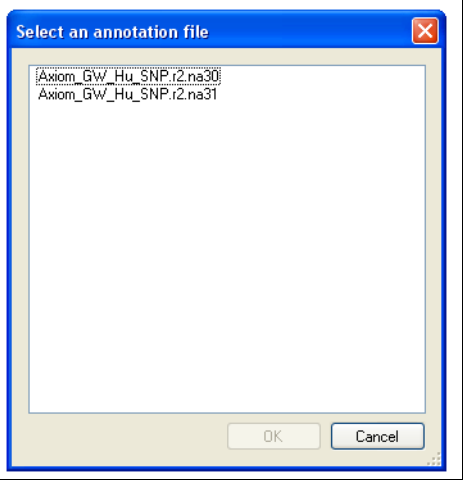
When the SNP Summary Statistics have been calculated, the Select SNP List dialog box opens (Figure 8.21).

Figure 8.21 Select a SNP list dialog box



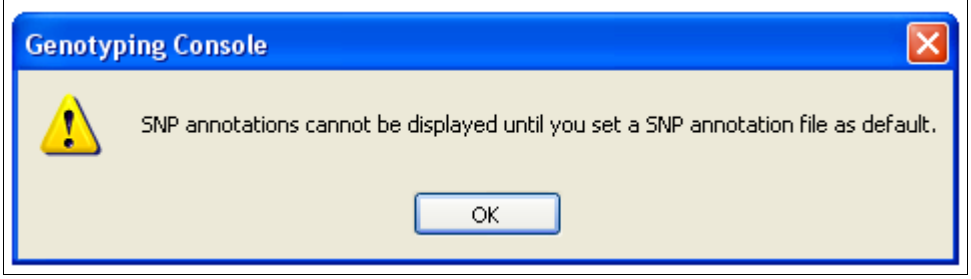
4. Select a SNP list and click **OK**.
   If a default SNP Annotation file has not been selected, the Select an annotation file dialog box opens (Figure 8.22).

Figure 8.22 Select an annotation file dialog box



   See *Annotation Options* on page 40 for instructions on setting a default SNP annotations file.
5. Select an annotation file and click **OK** in the Select an annotation file dialog box.
   If you click **Cancel**, the following notice appears (Figure 8.23).

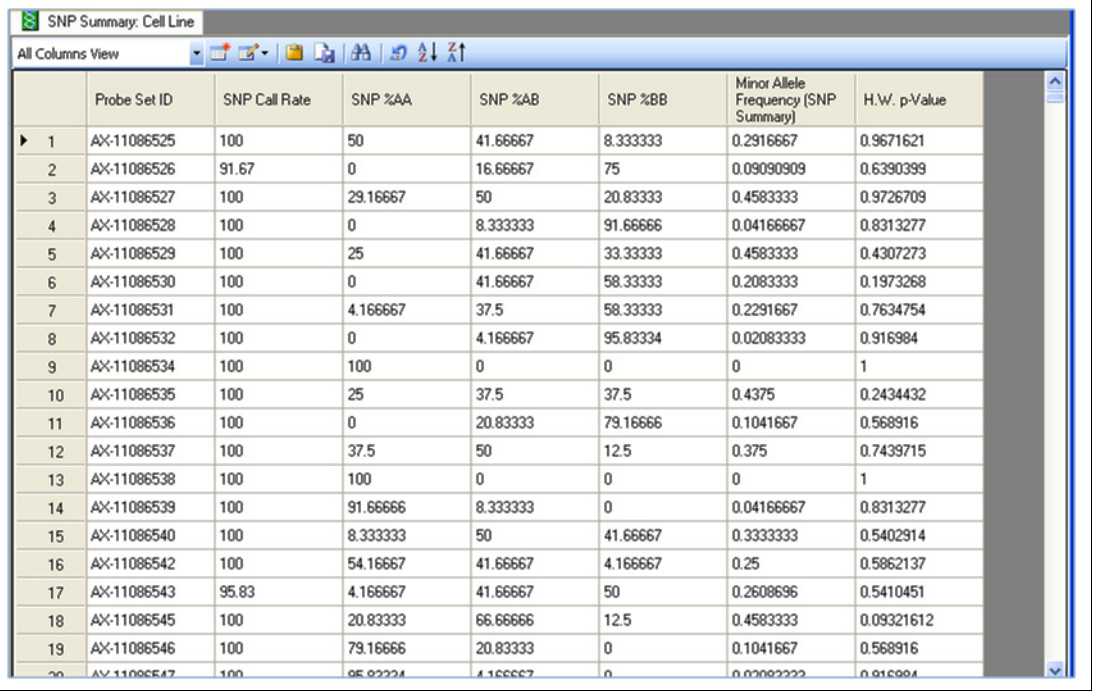
Figure 8.23 SNP Annotations notice



Click **OK** in the SNP Annotations notice to display the SNP Summary without annotation information.

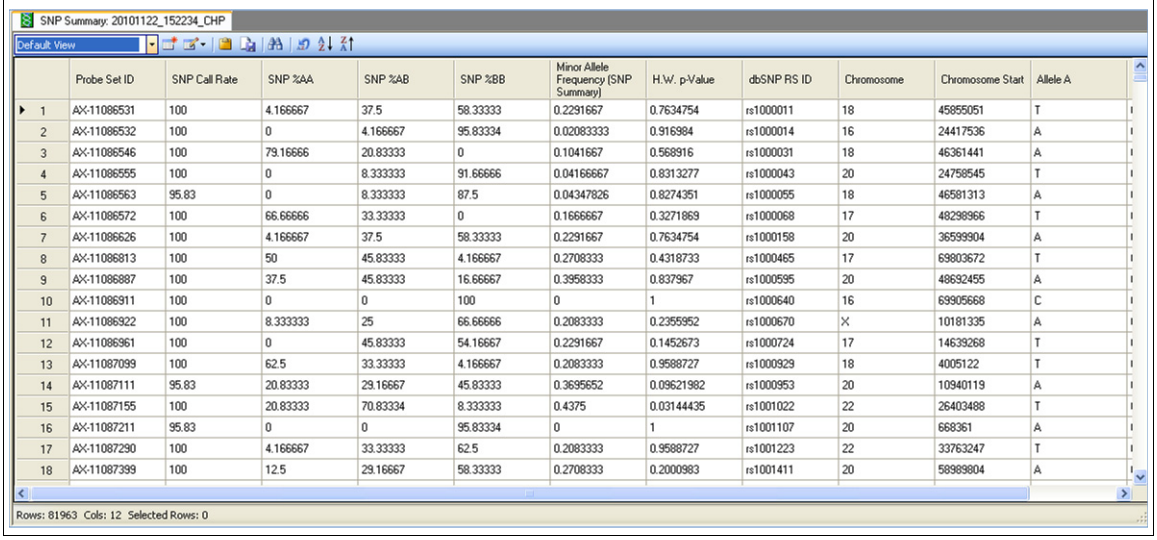
The SNP Summary Table opens (Figure 8.24, Figure 8.25).

Figure 8.24 SNP Summary table without annotation data



	Probe Set ID	SNP Call Rate	SNP %AA	SNP %AB	SNP %BB	Minor Allele Frequency (SNP Summary)	H.W. p-Value
1	AX-11086525	100	50	41.66667	8.333333	0.2916667	0.9671621
2	AX-11086526	91.67	0	16.66667	75	0.09090909	0.6390399
3	AX-11086527	100	29.16667	50	20.83333	0.4583333	0.9726709
4	AX-11086528	100	0	8.333333	91.66666	0.04166667	0.8313277
5	AX-11086529	100	25	41.66667	33.33333	0.4583333	0.4307273
6	AX-11086530	100	0	41.66667	58.33333	0.2083333	0.1973268
7	AX-11086531	100	4.166667	37.5	58.33333	0.2291667	0.7634754
8	AX-11086532	100	0	4.166667	95.83334	0.02083333	0.916984
9	AX-11086534	100	100	0	0	0	1
10	AX-11086535	100	25	37.5	37.5	0.4375	0.2434432
11	AX-11086536	100	0	20.83333	79.16666	0.1041667	0.568916
12	AX-11086537	100	37.5	50	12.5	0.375	0.7439715
13	AX-11086538	100	100	0	0	0	1
14	AX-11086539	100	91.66666	8.333333	0	0.04166667	0.8313277
15	AX-11086540	100	8.333333	50	41.66667	0.3333333	0.5402914
16	AX-11086542	100	54.16667	41.66667	4.166667	0.25	0.5862137
17	AX-11086543	95.83	4.166667	41.66667	50	0.2608696	0.5410451
18	AX-11086545	100	20.83333	66.66666	12.5	0.4583333	0.09321612
19	AX-11086546	100	79.16666	20.83333	0	0.1041667	0.568916

Figure 8.25 SNP Summary table with annotation data



	Probe Set ID	SNP Call Rate	SNP %AA	SNP %AB	SNP %BB	Minor Allele Frequency (SNP Summary)	H.W. p-Value	dbSNP RS ID	Chromosome	Chromosome Start	Allele A
1	AX-11086531	100	4.166667	37.5	58.33333	0.2291667	0.7634754	rs1000011	18	45855051	T
2	AX-11086532	100	0	4.166667	95.83334	0.02083333	0.916984	rs1000014	16	24417536	A
3	AX-11086546	100	79.16666	20.83333	0	0.1041667	0.568916	rs1000031	18	46361441	A
4	AX-11086555	100	0	8.333333	91.66666	0.04166667	0.8313277	rs1000043	20	24758545	T
5	AX-11086563	95.83	0	8.333333	87.5	0.04347826	0.8274351	rs1000055	18	46581313	A
6	AX-11086572	100	66.66666	33.33333	0	0.1666667	0.3271869	rs1000068	17	48298966	T
7	AX-11086626	100	4.166667	37.5	58.33333	0.2291667	0.7634754	rs1000158	20	36599904	A
8	AX-11086813	100	50	45.83333	4.166667	0.2708333	0.4318733	rs1000465	17	69803672	T
9	AX-11086887	100	37.5	45.83333	16.66667	0.3958333	0.837967	rs1000595	20	48692455	A
10	AX-11086911	100	0	0	100	0	1	rs1000640	16	69905668	C
11	AX-11086922	100	8.333333	25	66.66666	0.2083333	0.2355952	rs1000670	X	10181335	A
12	AX-11086961	100	0	45.83333	54.16667	0.2291667	0.1452673	rs1000724	17	14639268	T
13	AX-11087099	100	62.5	33.33333	4.166667	0.2083333	0.9588727	rs1000929	18	4005122	T
14	AX-11087111	95.83	20.83333	29.16667	45.83333	0.3695652	0.09621982	rs1000953	20	10940119	A
15	AX-11087155	100	20.83333	70.83334	8.333333	0.4375	0.03144435	rs1001022	22	26403488	T
16	AX-11087211	95.83	0	0	95.83334	0	1	rs1001107	20	668361	A
17	AX-11087290	100	4.166667	33.33333	62.5	0.2083333	0.9588727	rs1001223	22	33763247	T
18	AX-11087399	100	12.5	29.16667	58.33333	0.2708333	0.2000983	rs1001411	20	58989804	A

NOTE: You can see additional annotations by switching to "All Columns View".

NOTE: For readability, metrics are not displayed at full precision, and tables saved to file contain the same precision as is displayed in Genotyping Console. However, SNP filtering is performed using the full precision stored in the binary SNP summary file.

The SNP Summary Table contains the SNP level results and metrics (Table 8.1) for more information on performing genotyping.

NOTE: For Human Mapping 100K/500K, the SNP Summary data for the different array types will be displayed in different tables with different names.

See Chapter 7, *Genotyping Analysis* on page 89 for more information on performing genotyping. See *Table Features* on page 198 for more information on customizing the table view.

Table 8.1 SNP Summary table metrics

Column Header	Description
SNPID	The Affymetrix unique identifier for the set of probes used to detect a particular Single Nucleotide Polymorphism (SNP).
SNP Call Rate	Call Rate for that SNP across all samples in the batch. $$SNPCallRate = \frac{\#AA + \#AB + \#BB}{Total\ \#\ CHP\ Files}$$
SNP %AA	Percentage of AA calls for this SNP in this batch. $$\%AA = \frac{\#\ AA\ Calls}{Total\ \#\ CHP\ Files}$$
SNP %AB	Percentage of AB calls for this SNP in this batch. $$\%AB = \frac{\#\ AB\ Calls}{Total\ \#\ CHP\ Files}$$
SNP %BB	Percentage of BB calls for this SNP in this batch. $$\%BB = \frac{\#\ BB\ Calls}{Total\ \#\ CHP\ Files}$$

Table 8.1 SNP Summary table metrics

Column Header	Description
Minor Allele Frequency	The allele frequency for the A allele is calculated as: $$P_A = \frac{(\#\ AA\ Calls + 0.5 * AB\ Calls)}{Total\ \#\ Calls}$$ Where the Total # Calls does not include the No Calls. The B allele frequency is . $$P_B = 1 - P_A$$ The minor allele frequency is the $Min(P_A, P_B)$.

Table 8.1 SNP Summary table metrics

Column Header	Description
H-W p-value	Hardy Weinberg p-value is a measure of the significance of the discrepancy between the observed ratio or heterozygote calls in a population and the ratio expected if the population was in Hardy Weinberg equilibrium. The Hardy Weinberg p-value is calculated from the likelihood ratio: $$x^2 = \frac{(f^2 aa - fa)^2}{f^2 aa} + \frac{(2faafbb - fab)^2}{2faafbb} + \frac{(f^2 bb - fb)^2}{f^2 bb}$$ Where: $$fa = \frac{(\#AA\ Calls)}{Total\ \#\ Calls}$$ $$fb = \frac{(\#BB\ Calls)}{Total\ \#\ Calls}$$ $$faa = \frac{(\#AA\ Calls + 0.5 * \#\ AB\ Calls)}{Total\ \#\ Calls}$$ $$fbb = \frac{(\#BB\ Calls + 0.5 * \#\ AB\ Calls)}{Total\ \#\ Calls}$$ $$fab = \frac{(\#AB\ Calls)}{Total\ \#\ Calls}$$ The Hardy Weinberg p-value is . $$P_{HW} = CDF(x^2)$$ Where CDF is the Cumulative Distributive Function for the chi-squared distribution.
dbSNP RS ID	The dbSNP ID that corresponds to this probe set or SNP. The dbSNP at the National Center for Biotechnology Information (NCBI) attempts to maintain a unified and comprehensive view of known single nucleotide polymorphisms (SNPs), small scale insertions/deletions, polymorphic repetitive elements, and microsatellites from TSC and other sources. The dbSNP is updated periodically, and the dbSNP version used for mapping is given in the dbSNP version field. For more information, please see: http://www.ncbi.nlm.nih.gov/SNP/.
Chromosome	The chromosome on which the SNP is located on the current Genome Version.
Physical Position	The nucleotide base position where the SNP is found. The genomic coordinates given are in relation to the current genome version and may shift as subsequent genome builds are released.

Table 8.1 SNP Summary table metrics

Column Header	Description
Allele A	The allele of the SNP that is in lower alphabetical order. When comparing the allele data on NetAffx to the allele data for the corresponding RefSNP record in dbSNP, the alleles reported here could be different from the alleles reported for the corresponding RefSNP on the dbSNP web site. This difference arises mainly from the reference genomic strand that was chosen to define the alleles by Affymetrix. To choose the reference genomic strand, we follow a convention based on the alphabetic ordering of the sequence surrounding the SNP. Sometimes the reference strand on the dbSNP is different from NetAffx, and the alleles could represent reverse complement of those provided on dbSNP.
Allele B	The allele of the SNP that is in higher alphabetical order. When comparing the allele data on NetAffx to the allele data for the corresponding RefSNP record in dbSNP, the alleles reported here could be different from the alleles reported for the corresponding RefSNP on the dbSNP web site. This difference arises mainly from the reference genomic strand that was chosen to define the alleles by Affymetrix. To choose the reference genomic strand, we follow a convention based on the alphabetic ordering of the sequence surrounding the SNP. Sometimes the reference strand on the dbSNP is different from NetAffx, and the alleles could represent reverse complement of those provided on dbSNP.

NOTE: You can display additional annotations by selecting the "All Columns View". For complete descriptions on all available annotations columns in the SNP Summary table, see Appendix D.

See *Performing Genotyping Analysis* on page 89 for more information on performing genotyping analyses. See *Table Features* on page 198 for more information on customizing the table view.

NOTE: The SNP Summary table does not support line graphs.

Concordance Checks

The concordance checks enable you to compare the SNP calls in different files. You can perform:

- *CHP vs. Text Concordance Check* on page 140: Compares the SNP calls in a CHP file with the SNP calls in a previously created text file. In this check you can compare multiple CHP files to the same text file. You can use a Text reference file, such as the 500K Ref_103 file provided on the Affymetrix website, or create your own reference file. Reference files for Concordance Checks must have "ProbeSet ID" as the first column and "Call" or "Consensus" as the second column.
- *CHP vs. CHP Concordance Check* on page 144: Compares the SNP calls in one CHP file to the SNP Calls in another CHP file. This comparison is done on a paired basis; you can perform the check on multiple pairs of CHP files in the same analysis. The output for both checks is a single "report" file that can be displayed as a table.

In both cases the check compares the SNPs that are common to both sample and reference files and have genotype calls. SNPs that are not shared between the files, and SNPs that do not have calls, are not included in the comparison.

IMPORTANT: The definition of allele A and allele B (call codes) is different among different arrays. For some arrays, all SNPs are mapped to the forward strand of the genome. For other arrays, SNPs can be on the forward strand or reverse strand of the genome. This means that a particular SNP that is present in two different arrays can have different call codes for the same base calls. For example, the same GG base call can be AA in SNP 6.0 results or BB in Axiom™ Genome-Wide array results. The 'Strand' column in the annot.db file lists the strand information for all the SNPs.

Table 8.2 Examples of the different call codes and base calls made for the same G/C SNP on the Genome-Wide Human SNP Array 6.0 and on Axiom Genome-Wide Human Array

dbSNP XX	Genome-Wide Human SNP Array 6.0			Axiom Genome-Wide CEU 1 Array	
	Annotation file Base Call (Reverse Strand)	Export Base Call (Forward Strand)	Affymetrix Call Code	Annotation file and export Base Call (Forward Strand)	Affymetrix Call Code
Allele 1	C	G	A	G	B
Allele 2	G	C	B	C	A

IMPORTANT: When performing a CHP vs. Text Concordance Check between Axiom™ Genome-Wide Human Arrays and other arrays, the data must be carefully compared. You cannot simply look at "%Concordance" numbers. For call code comparisons of SNPs on the reverse strand of the genome, the AA calls = BB calls in Axiom, AB calls = BA calls in Axiom, BB calls = AA calls in Axiom (Table 8.3). For SNPs on the forward strand of the genome, the AA calls = AA calls in Axiom; AB calls = AB calls in Axiom; BB calls = BB calls in Axiom.

Table 8.3 Possible genotypes for an example G/C SNP

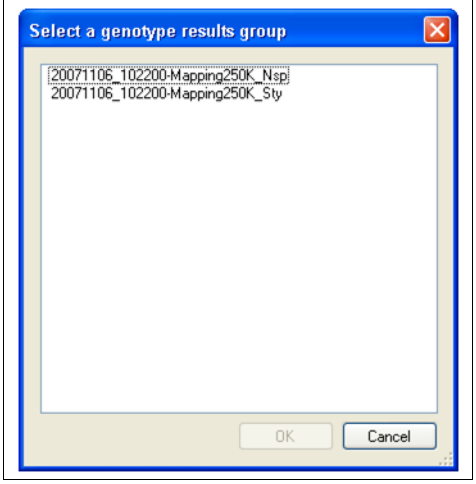
Genome-Wide Human SNP Array 6.0		Axiom™ Genome-Wide CEU 1 Array	
Forward Strand Base Call	Affymetrix Call Code	Forward Strand Base Call	Affymetrix Call Code
GG	AA	CC	AA
GC	AB	CG	AB
CC	BB	GG	BB

CHP vs. Text Concordance Check

To perform a reference concordance check:

1. Open the Workspace and select the Data Set with the data for analysis.
2. Select the Genotype Results file set.
3. From the Workspace menu, select **Genotype Results > Run CHP vs. TXT Concordance Check**; or
   Right-click the Genotype Results file set and select **Run CHP vs. TXT Concordance Check** from the pop-up menu.
   If you have not previously selected a Results file set, the Select Genotype Results Groups dialog box opens (Figure 8.26).

Figure 8.26 Select Genotype Results Group dialog box

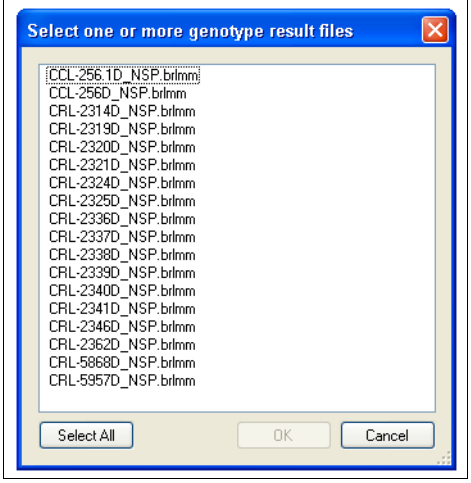


4. Select a results group from the list and click **OK**.

NOTE: You will be able to select arrays from only one enzyme set at a time when performing a CHP vs. Text Concordance Check.

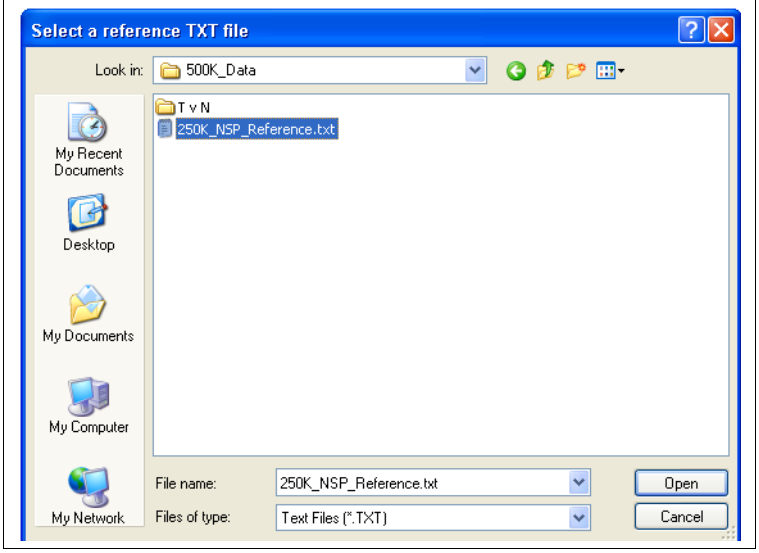
The Select files dialog box opens (Figure 8.27).

Figure 8.27 Select Results files dialog box



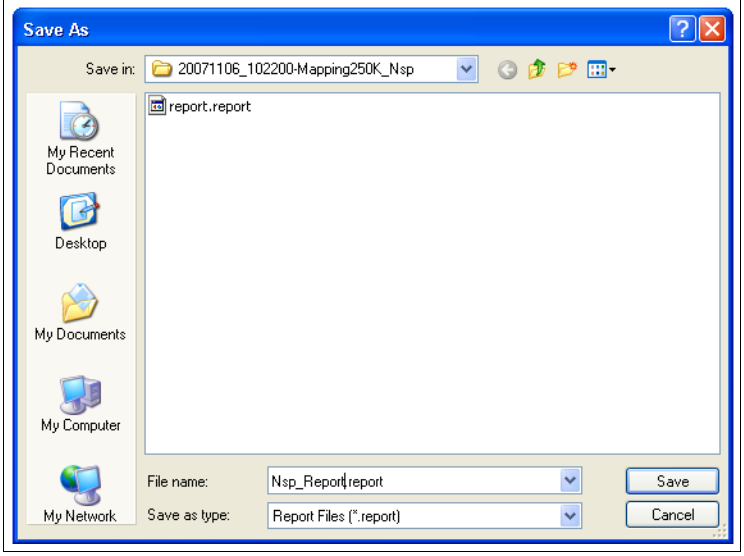
5. Select the files for concordance check and click **OK**.
   The Select Reference File opens (Figure 8.28).

Figure 8.28 Select Reference File dialog box



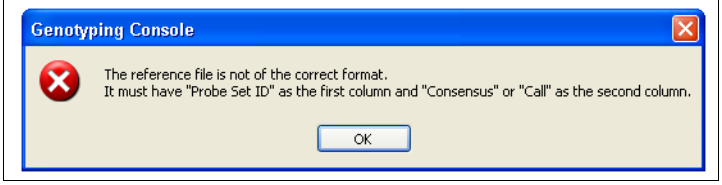
6. Browse to the location with the reference file you wish to use and select the file.
   See *Reference File Format* on page 143 for more information.
7. Click **Open**.
   The Save As dialog box opens (Figure 8.29).

Figure 8.29 Save As dialog box



8. Browse to the location where you want to save the report and enter a file name for the report.
9. Click **Save**.
   If the reference file does not have the correct format, the following error message appears (Figure 8.30).

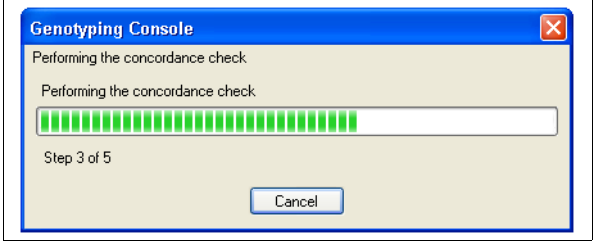
Figure 8.30 Error message



If this message appears, click **OK** to cancel the operation and then fix the file format problem. See *Reference File Format* on page 143 for more information.

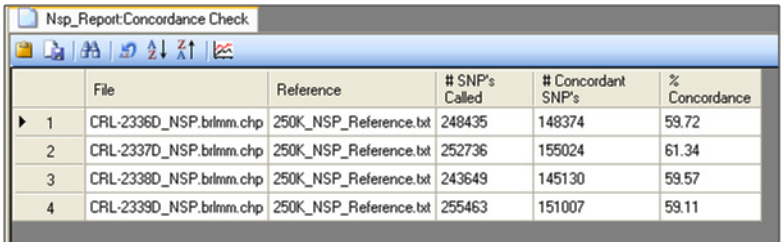
If the reference file is correct, the Progress bar appears (Figure 8.31).

Figure 8.31 Progress bar displaying the progress of the concordance check



When the analysis is finished, the Reference Concordance Report table appears (Figure 8.32).

Figure 8.32 Reference Concordance Report table



	File	Reference	# SNP's Called	# Concordant SNP's	% Concordance
1	CRL-2336D_NSP.brlmm.chp	250K_NSP_Reference.txt	248435	148374	59.72
2	CRL-2337D_NSP.brlmm.chp	250K_NSP_Reference.txt	252736	155024	61.34
3	CRL-2338D_NSP.brlmm.chp	250K_NSP_Reference.txt	243649	145130	59.57
4	CRL-2339D_NSP.brlmm.chp	250K_NSP_Reference.txt	255463	151007	59.11

You can also open the concordance report from the data tree. The Reference Concordance report table contains the following information:

- File – Sample file name
- Reference – Reference file name
- #SNP's Called – Number of SNPs common to both sample and reference files with genotype calls
- # Concordant SNP's – Number of called SNPs that have the same genotype call
- % Concordance – Percentage of called SNPs that have the same genotype call

You can:

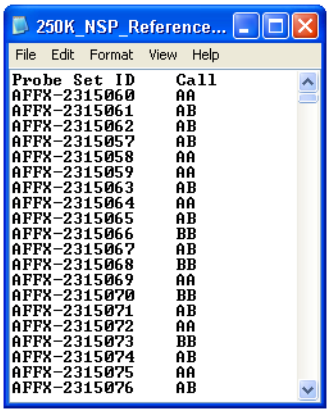
- Copy selected data in the table to the clipboard.
- Save the entire table as a text file.

Reference File Format

The reference file is a tab-delimited text file with two columns (Figure 8.33):

- First column must be titled "Probe Set ID"
- Second column must be titled "Consensus" or "Call"

Figure 8.33 Example reference file for Reference Concordance check



A reference file can be created by editing a *genotyping results file* (page 181).

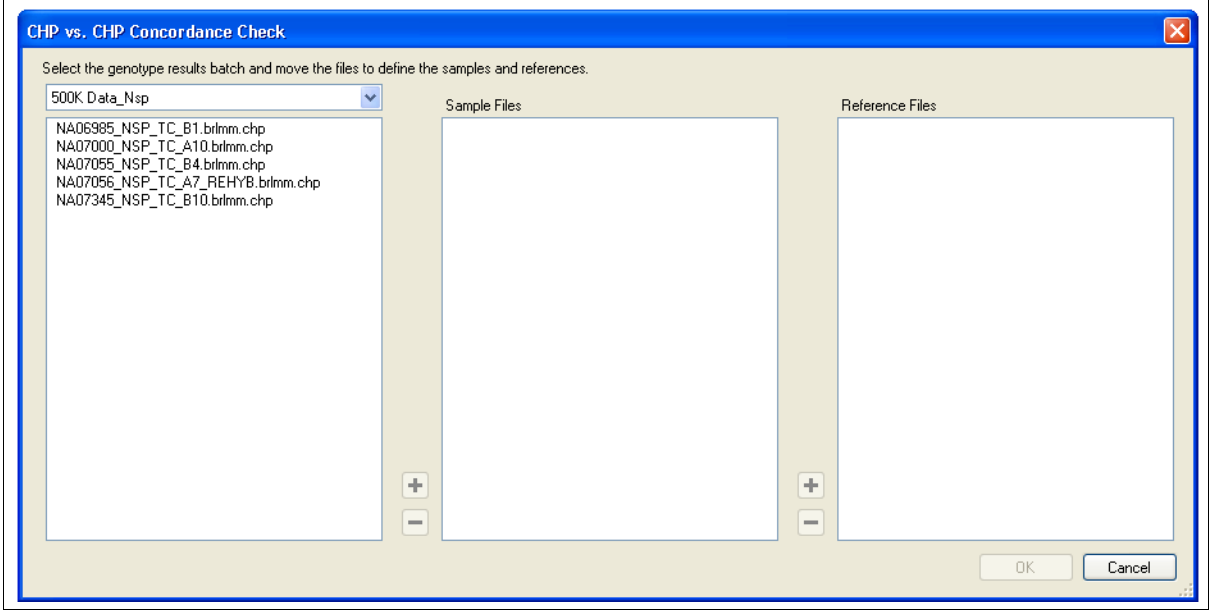
NOTE: The column headers must be capitalized as shown in the figure above **(Figure 8.33)**.

CHP vs. CHP Concordance Check

To perform a CHP vs. CHP concordance check:

1. Open the Workspace and select the Data Set with the data for analysis.
2. Select the Genotype Results file set (optional).
3. From the Workspace menu, select **Genotype Results > Run CHP vs. CHP Concordance Check**; or
   Right-click the Genotype Results file set and select **Run CHP vs. CHP Concordance Check** from the pop-up menu.
   The Select files dialog box opens (Figure 8.34).

Figure 8.34 Select Results files dialog box

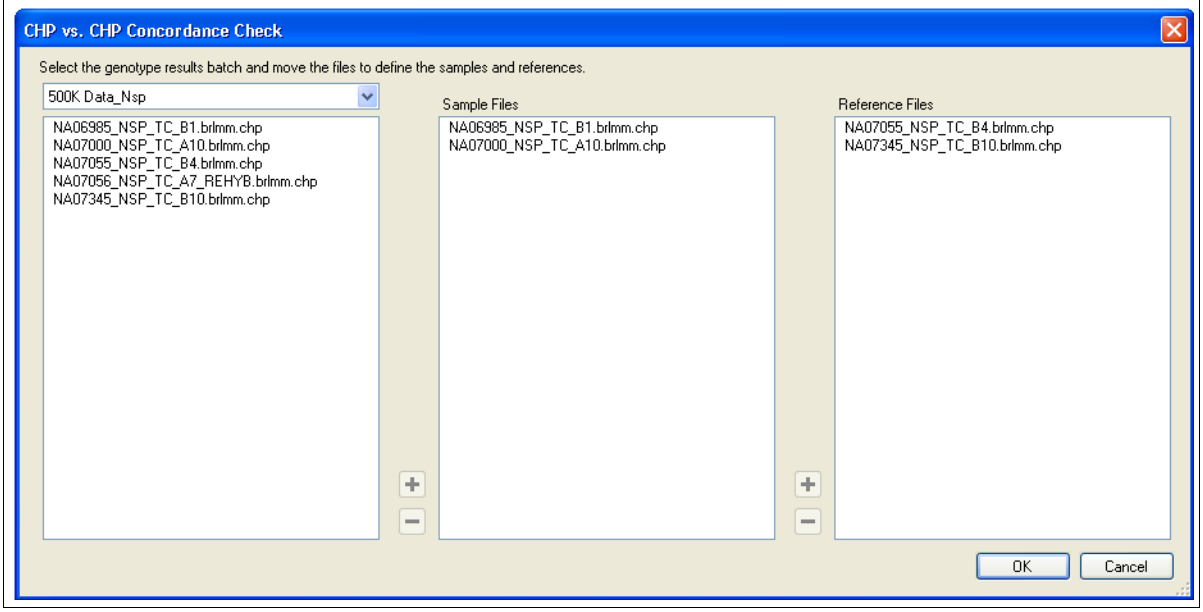


4. Select files in the Available Files list.
   Click the Add button ⊕ to add data to the sample or reference list.
   Click the Remove button ⊖ to remove data from a list.

NOTE: The first file in the Sample Files list is compared to the first file in the Reference Files list. The second files in both lists are compared to each other, and so on, as shown in the figure below (Figure 8.35).

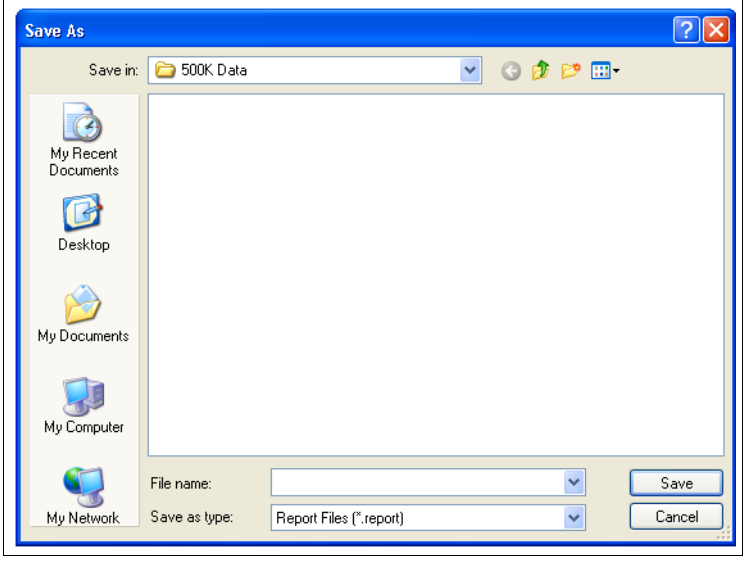
NOTE: You can pair files from different enzyme sets for Human Mapping 100K/500K array sets; this allows you to compare the signature SNPs for arrays with different enzyme sets.

Figure 8.35 Files paired for CHP vs. CHP concordance check



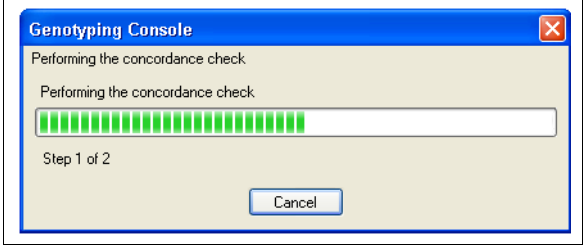
5. 5.When you have selected the files for the concordance check, click **OK**.
   The Save As dialog box opens (Figure 8.36).

Figure 8.36 Save As dialog box



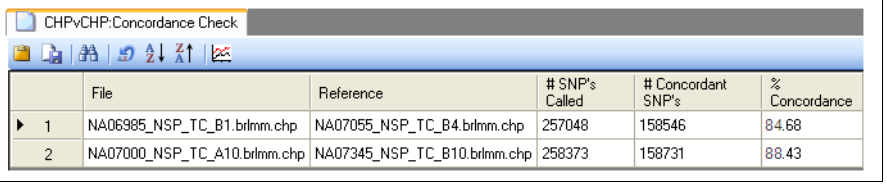
6. Browse to the location where you want to save the report and enter a file name for the report.
7. Click **Save**.
   If the reference file is correct, the Progress bar appears (Figure 8.37).

Figure 8.37 Save As dialog box



When the analysis is finished, the Concordance Report table appears (Figure 8.38).

Figure 8.38 Reference Concordance Report table



	File	Reference	# SNP's Called	# Concordant SNP's	% Concordance
1	NA06985_NSP_TC_B1.brlmm.chp	NA07055_NSP_TC_B4.brlmm.chp	257048	158546	84.68
2	NA07000_NSP_TC_A10.brlmm.chp	NA07345_NSP_TC_B10.brlmm.chp	258373	158731	88.43

You can also open the concordance report from the data tree.

The Reference Concordance report table contains the following information:

File – Sample CHP file name

Reference – Reference CHP file name

SNP's Called– Number of SNPs common to both sample and reference files with genotype calls

Concordant SNP's – Number of called SNPs that have the same genotype call

% Concordance – Percentage of called SNPs that have the same genotype call

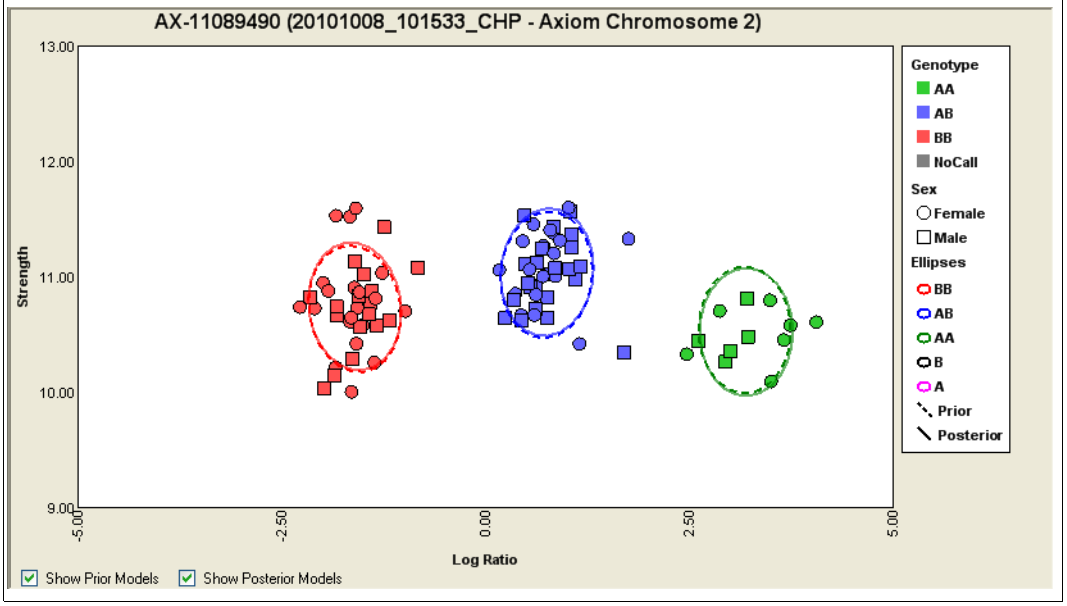
You can:

- Copy selected data in the table to the clipboard.
- Save the entire table as a text file.

Using the SNP Cluster Graph

The SNP Cluster Graph (Figure 9.1) displays the SNP calls for selected samples as a set of points in the clustering space used for making the calls. It allows you to perform a visual inspection of the SNP calls, aids in identifying problematic SNPs, and to manually change calls.

Figure 9.1 SNP cluster graph for data from an Axiom Genome-Wide array.



See *Parts of the SNP Cluster Graph* on page 155 to learn more about the SNP Cluster Graph components.

The SNP Cluster Graph is described in the following sections:

- *Introduction*
- *Generating SNP Cluster Graphs* on page 152
- *Parts of the SNP Cluster Graph* on page 155
- *Changing a SNP's Call* on page 160
- *Changing the Display* on page 168
- *Saving Cluster Graph Information* on page 173

Introduction

While applying per-SNP filters helps remove the majority of problematic SNPs, no filtering scheme is perfect. Even with stringent filtering, a small proportion of poorly performing SNPs may remain. Moreover, the poorly performing SNPs are often the ones most likely to perform differently between cases and controls. The list of significantly associated SNPs is often enriched for such problematic SNPs.

The SNP filtering process greatly reduces the occurrence of these false positives, but given their tendency to end up in the list of associated SNPs, it is likely that some will remain. Before carrying forth SNPs to subsequent phases of analysis, visual inspection of the SNPs in the clustering space is strongly recommended, since this inspection can help identify problematic SNPs.

The SNP Cluster Graph displays SNP clusters and allows you to change a SNP's original call. When editing a SNP call the software stores a copy of the original call in the CHP file. This back up of the original call will be stored in the for all arrays analyzed with GTC 4.2. For CHP files created with earlier versions of the software the backup data will be saved when adding CHP files to the workspace. For CHP

files already in the workspace you can create the backup by optimizing for the cluster graph, or the cluster graph itself will create it. It is recommended to optimize before using the cluster graph for the best performance.

In the cluster graph, user-selected colors and shapes can be assigned to genotype and sample call data and to other attributes. For example, the cluster graph in displays genotype by color and uses different shapes to indicate gender.

NOTE: Samples must have a sample file (ARR) in order to display user attributes by color or shape. If sample files are not available (for example, CHP files generated in GCOS), then only array plate information, fluidics instrument information or scanner information (if available in the CHP file) can be displayed using color or shape.

The graph can also display the prior and posterior cluster location information used to make the calls.

NOTE: SNP filtering uses the full precision of stored metrics. The displayed precision in tables is less than this for readability.

The Clustering space is calculated differently depending upon the analysis applied to the data:

- *BRLMM and BRLMM-P Data* on page 149
- *Birdseed Data* on page 150
- *Axiom Data* on page 151

BRLMM and BRLMM-P Data

For BRLMM and BRLMM-P, the clustering is performed in the transformed contrast dimension.

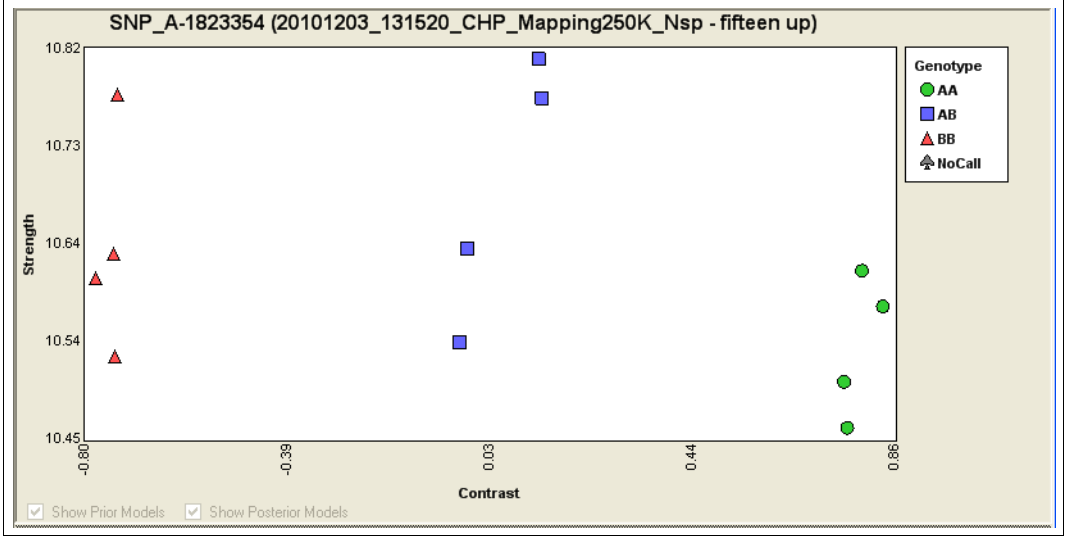
Contrast is defined as:

$$Contrast = f\left[\frac{(A - B)}{(A + B)}\right]$$

See the BRLMM-P white paper for more details on the transformation applied to the contrast.

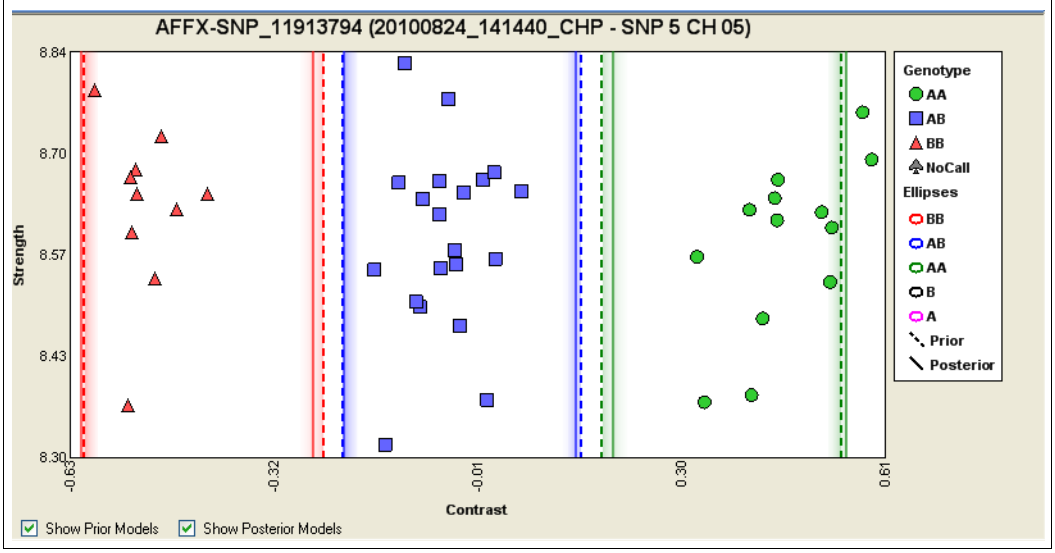
The BRLMM graph does not display the cluster location information (Figure 9.2).

Figure 9.2 SNP Cluster Graph for samples analyzed using the BRLMM algorithm



The graph for BRLMM-P graph displays the cluster location information as straight lines, using solid lines for posterior model files and dashed lines for prior model files lines (Figure 9.3).

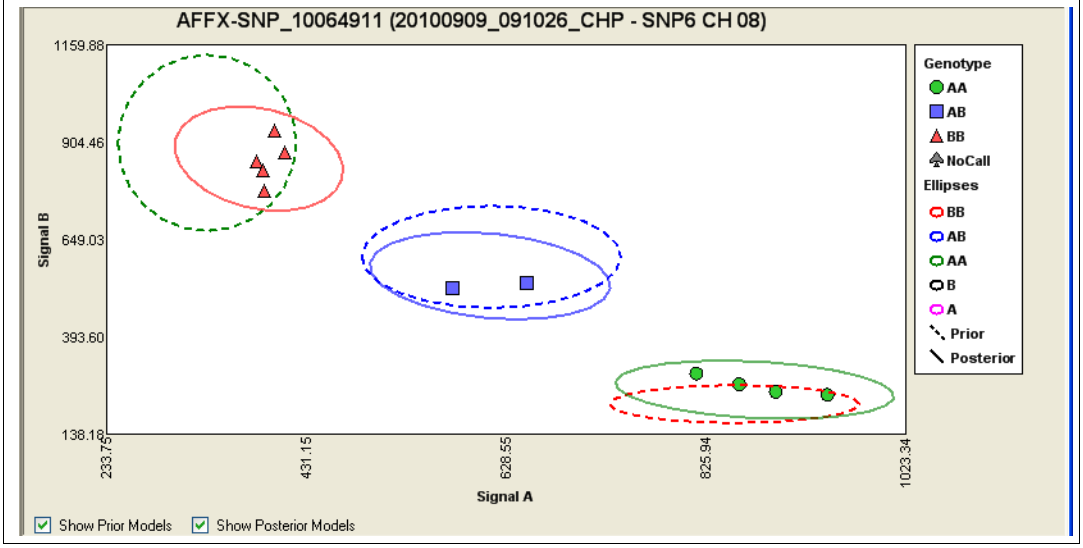
Figure 9.3 SNP Cluster Graph for samples analyzed using the BRLMM-P algorithm



Birdseed Data

For Birdseed, clustering is performed in a two dimensional A versus B space (Figure 9.4).

Figure 9.4 SNP Cluster Graph for samples analyzed using the Birdseed v2 algorithm



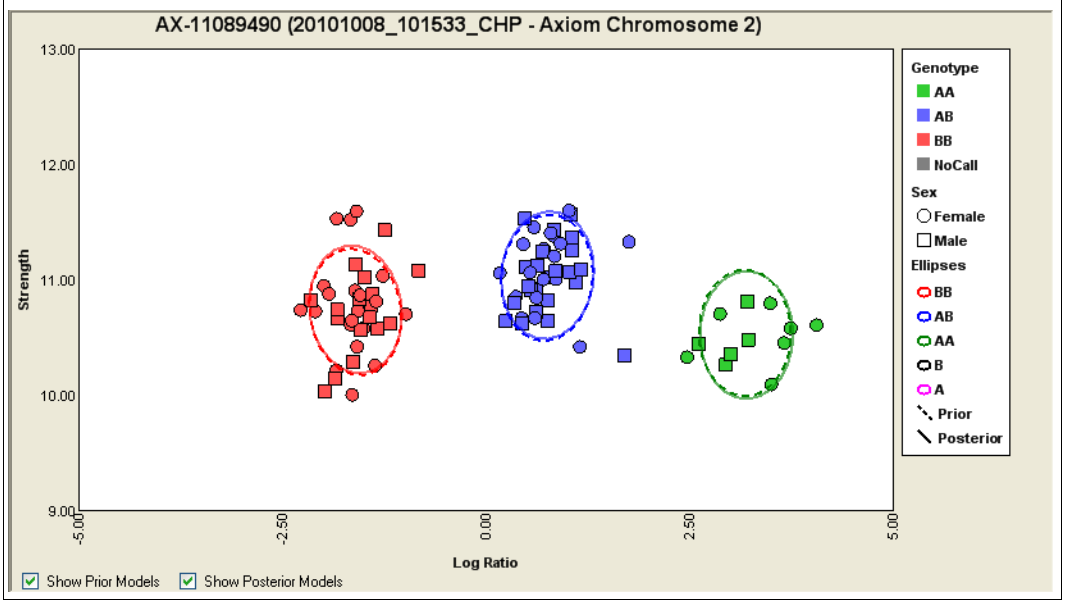
The graph displays the prior and posterior cluster location information as ellipses, using solid lines for posterior model files and dashed lines for prior model files lines (Figure 9.4).

Axiom Data

For the Axiom GT1 algorithm, clustering is performed in Log ratio versus strength space (Figure 9.5). Log ratio and strength are defined as:

- Log Ratio = $\log_2(A)-\log_2(B)$
- Strength = $(\log_2(A)+\log_2(B))/2$

Figure 9.5 SNP Cluster Graph for samples analyzed using the Axiom GT1 algorithm



The graph displays the prior and posterior cluster location information as ellipses, using solid lines for posterior model files and dashed lines for prior model files lines.

Generating SNP Cluster Graphs

Before generating a SNP cluster graph, you need:

- A set of genotyping results

  See Chapter 7, *Genotyping Analysis* on page 89.

- A SNP List for that set of results

  See *Create a SNP List* on page 125.

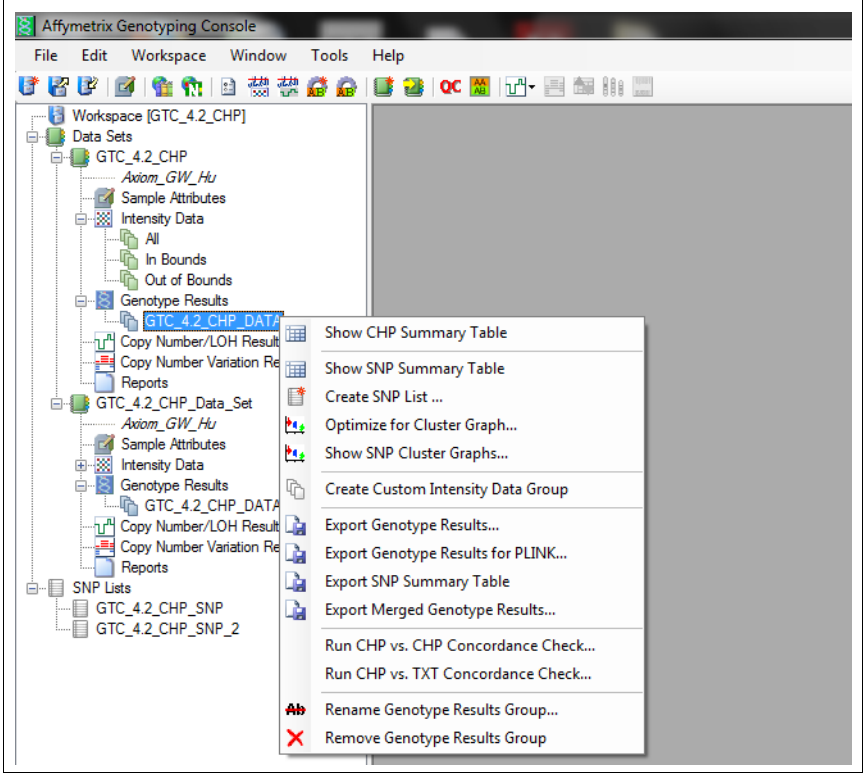
TIP: If you are using a Workspace from a previous version of GTC (4.1 or older), make sure you click on *Optimize for Cluster Graph* (Figure 9.6) before generating SNP Cluster Graphs. This feature creates a backup of the samples original calls for the samples that results in faster processing times after editing.

This action is not required for Workspaces created in GTC 4.2, as they are automatically optimized by default.

To generate SNP cluster graphs:

1. Right-click a Genotyping Results batch and select **Show SNP Cluster Graphs** on the shortcut menu (Figure 9.6).

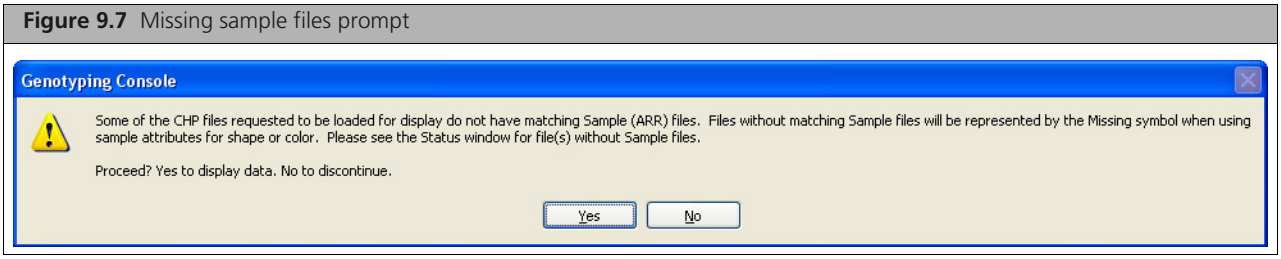
Figure 9.6 Genotype Results menu, Show SNP Cluster Graphs selected



If none of the CHP files have matching sample files (for example, if the files were generated by GCOS), or if all of the CHP files have matching sample files, no warning appears and the cluster graph is generated.

If some of the CHP files are missing matching sample files (ARR), the following warning appears (Figure 9.7):

Figure 9.7 Missing sample files prompt



If you click:

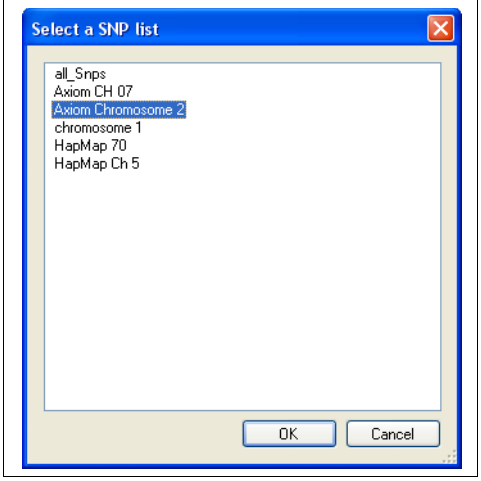
- Yes – The cluster graph will displays a gray spade (♠) for samples without the attributes selected from the Color or Shape drop-down lists.

  The Status window lists the files with missing sample data. No user attributes are available for these CEL files. Only the physical array attributes (scanner ID or fluidics information, if available) can be selected from the Color and Shape drop-down lists.

  Files with sample data available will be displayed normally.
- No – The cluster graph is not created.

The Select a SNP list dialog box appears (Figure 9.8).

Figure 9.8 Select a SNP List dialog box



If no SNP List is available, you must first generate one.

For more details, see:

- *Create a SNP List* on page 125.
- *Import a Custom SNP List* on page 130.

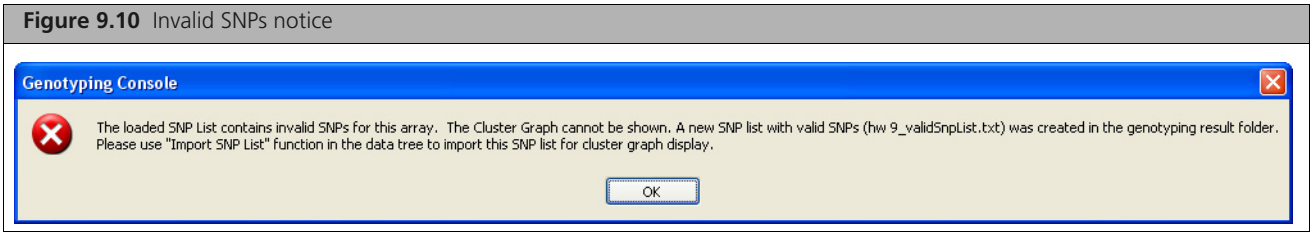
2. Select a SNP List and click **OK**.

   If there are no common SNPs in the selected SNP list and the array probes, the following notice appears (Figure 9.9).

Figure 9.9 No valid SNPs notice

If some SNP lists are in common, the following notice appears (Figure 9.10):

Figure 9.10 Invalid SNPs notice



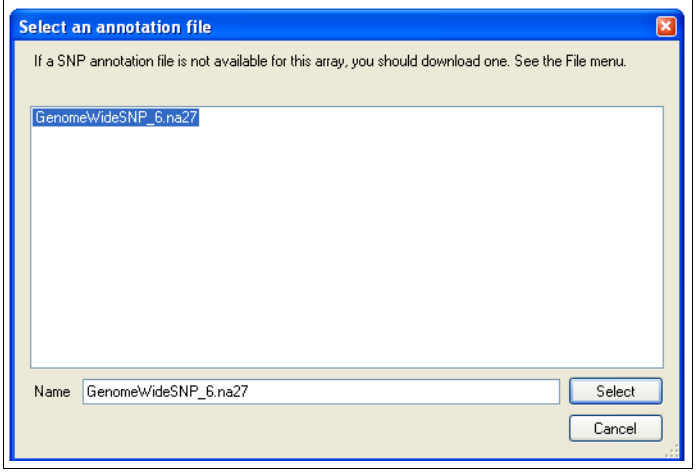
A new SNP list with the SNPs common to both the original SNP list and the array probe set is created in the genotyping results folder. It can be imported and used for the SNP Cluster graph.

NOTE: Depending on the number of CHP files in the Genotyping Results batch and the number of SNPs in the SNP List, generating the SNP Cluster Graph can take several minutes.

If the SNP list has no invalid SNPs, the Select an Annotation File dialog box (Figure 9.11) opens if an annotation file has not already been selected.

If an annotation file is not available on the computer, you are prompted to download one.

Figure 9.11 Select an annotation file for SNP Cluster Graph



3. (optional) Select an annotation file and click **OK**.

   The SNP Cluster graph is displayed (Figure 9.12).

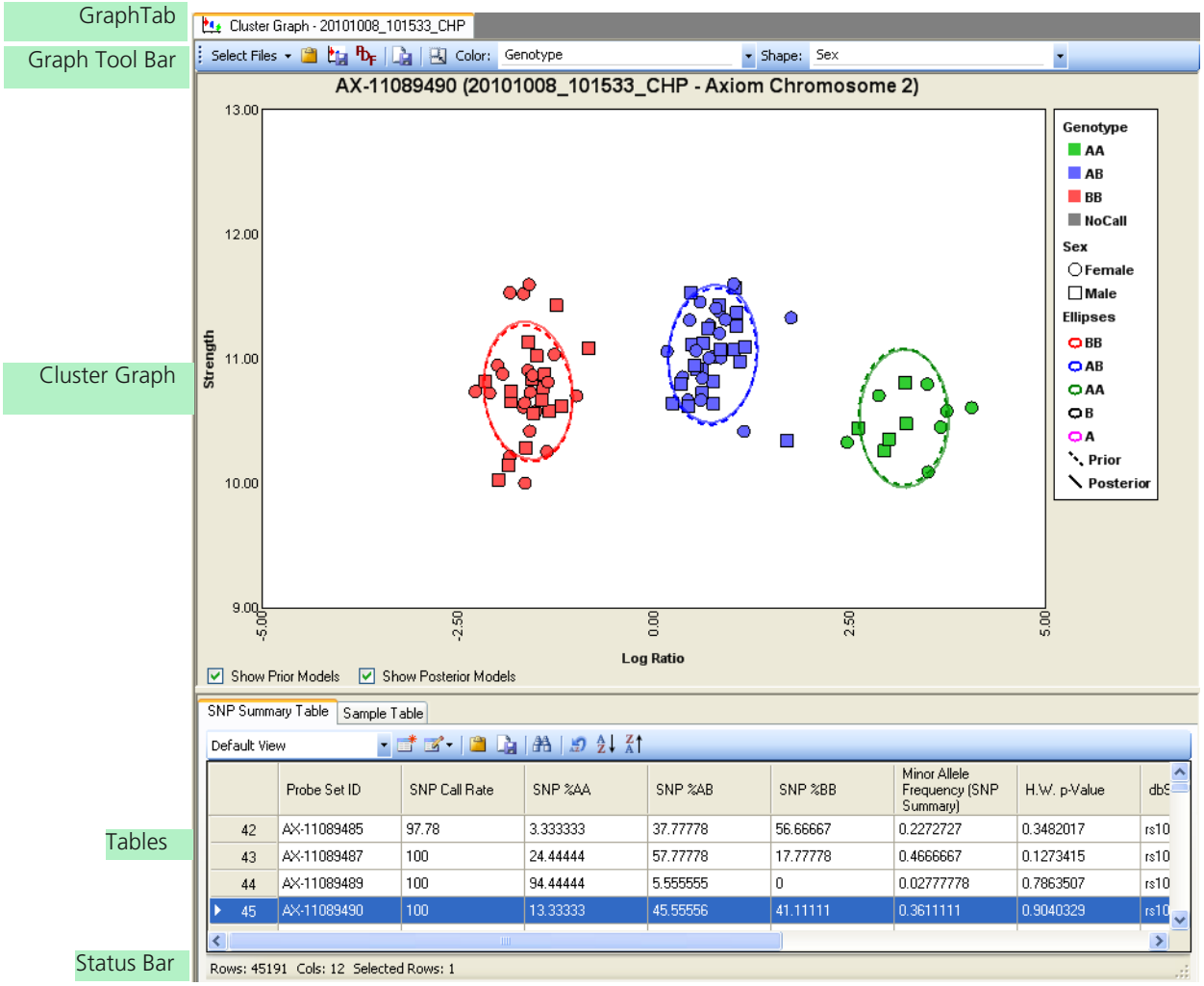
   See *Parts of the SNP Cluster Graph* on page 155 for more information.

   The values of the graph axes are different, depending upon the type of array data displayed:

   - *BRLMM and BRLMM-P Data* on page 149
   - *Birdseed Data* on page 150
   - *Axiom Data* on page 151

Parts of the SNP Cluster Graph

Figure 9.12 Parts of the SNP Cluster Graph



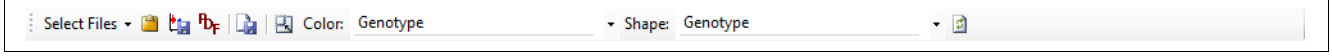
The SNP Cluster Graph has the following components:

- Graph Tab: displays name of genotype results set
- *Cluster Graph Tool Bar* on page 156
- *Cluster Graph* on page 156
- Tables
  - *SNP Summary Table* on page 161
  - *Sample Table* on page 162
- Status Bar: Displays:
  - Number of SNPs in table
  - Number of Samples in Sample Table
  - Missing ARR files.

Cluster Graph Tool Bar

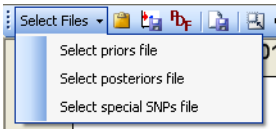
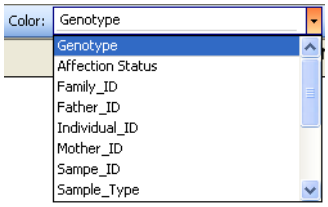
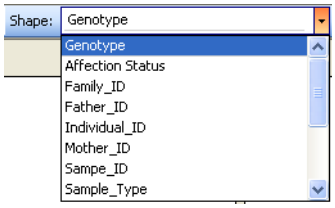
The Cluster Graph Tool bar allows quick access to the functions of the graph.

Figure 9.13 Cluster Graph Tool bar



See the table below (Table 9.1) for more information.

Table 9.1 Cluster Graph Tool bar functions

Button	Function
	Select Files: Select different model and special SNPs list files See: *Selecting Model Display Options* on page 169 *Selecting the Special SNPs File* on page 172
	*Copy Image to Clipboard* on page 174
	*Save Image to File* on page 175
	*Save All SNP Cluster Graphs to PDF File* on page 176
	*Save SNP Data from the SNP Cluster Graph* on page 179
	*Set axis on graph* (page 173)
	*Select attributes for the color of SNPs* (page 168)
	*Select attributes for the shape of SNPs* (page 168)
	Revert button: Revert back all modified Calls in each sample file back to original values.

Cluster Graph

The cluster graph displays the SNP calls for each sample in the results group.

Each sample is plotted on the axes appropriate for the analysis type:

- *BRLMM and BRLMM-P Data* on page 149
- *Birdseed Data* on page 150
- *Axiom Data* on page 151

The components of the cluster graph are described in *Parts of the SNP Cluster Graph* on page 157.

The default view shows samples colored by the genotype call.

See *Selecting Colors and Shapes for Attributes* on page 168 for information on changing the attributes indicated by different shapes and colors.

The software warns you if some of the CHP files do not have matching sample files (ARR).

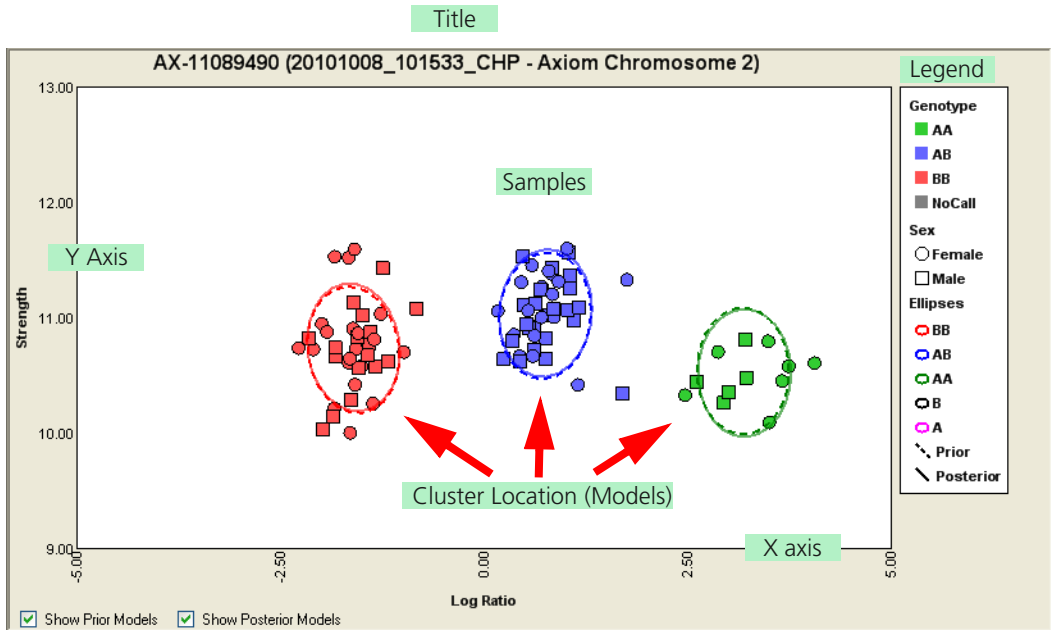
The SNP cluster graph can display up to 10 different colors and up to 10 different shapes.

If the attributes selected for display have more than 10 categories, categories 1 through 9 will be displayed normally, but categories 10 and higher will be grouped together.

See *Selecting Colors and Shapes for Attributes* on page 168 for more information.

Parts of the SNP Cluster Graph

Figure 9.14 SNP Cluster Graph with samples analyzed using Axiom GT1 algorithm



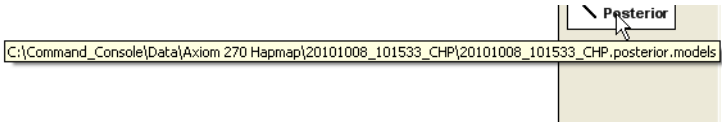
- Title: displays the ID of displayed SNP in the following format:
  SNP ID (Genotyping Results batch name - SNP list name)
- X and Y axes

  See *Change the Scale of the SNP Cluster Graph Axes* on page 173
- Legend box: Displays the legend for the graph, including information on:
  - Use of colors and shapes for displaying calls

    See *Selecting Colors and Shapes for Attributes* on page 168 for information on changing the attributes indicated by different shapes and colors.
  - Colors of cluster location ellipses
  - Use of dashed or solid lines to display cluster location (model) information

    You can mouse over the Posterior or Prior legend to display information on the name and location of the displayed model file (Figure 9.15).

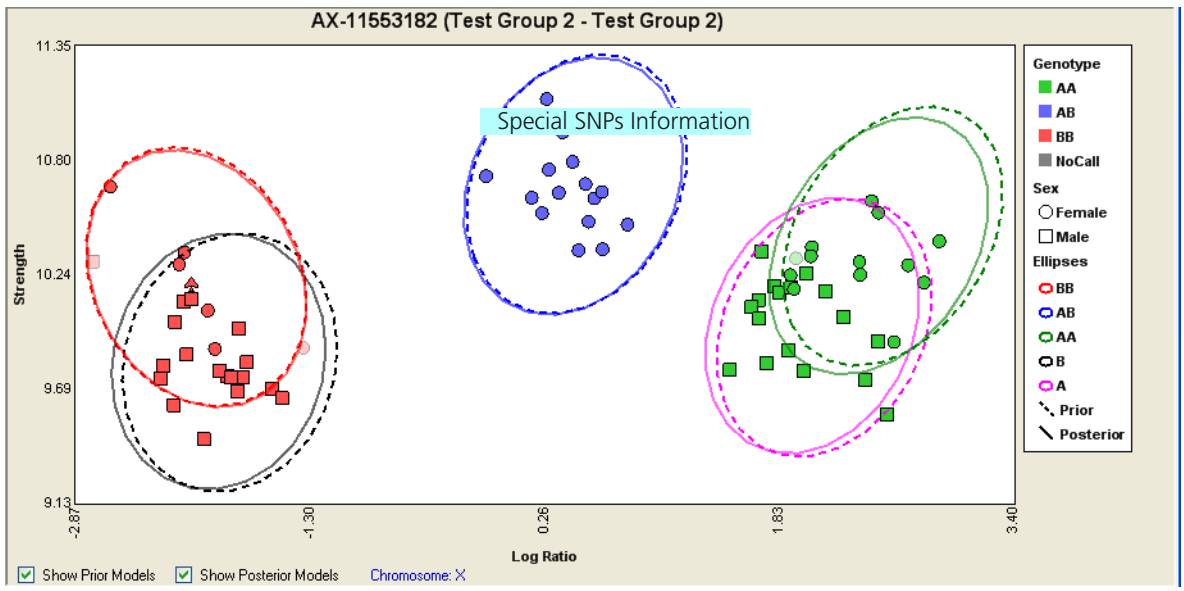
Figure 9.15 Displaying model file name and location



- Show Models checkboxes: toggle the display of the model ellipses on and off.
  See *Selecting Model Display Options* on page 169
- Special SNP information (see Figure 9.16):
  Provides info on SNPs on the following types of chromosomes:
  - Mitochondrial
  - X
  - Y
  - PAR (PseudoAutosomal Region)

  See *Selecting the Special SNPs File* on page 172.

Figure 9.16 SNP Cluster Graph with Special SNPs data



You can change the SNP cluster graph being displayed by toggling through the data displayed in the SNP Summary table.

To display the SNP cluster graph that corresponds to a particular SNP:

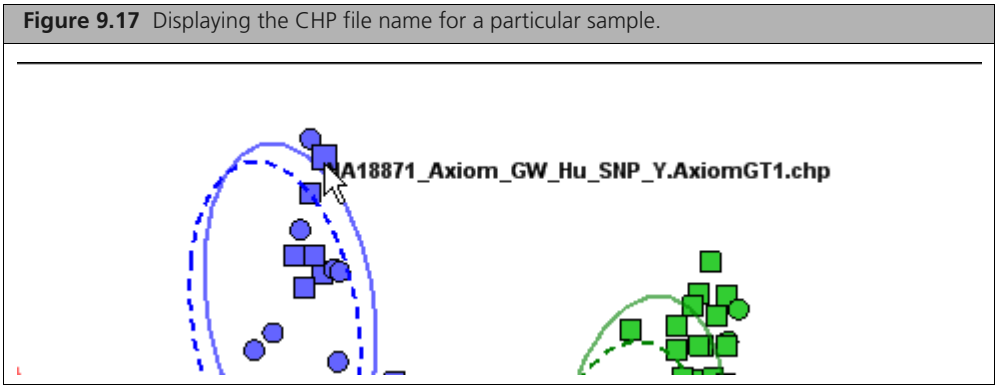
- Click on the corresponding row in the SNP summary table.
  You can use the arrow keys on the keyboard to toggle through the list. The SNP Cluster graph updates to display the data for the SNP.

In the graphical portion of the window, you can copy the current image to the Clipboard, or save the current image to file (*.png format).

To learn more about a particular sample:

- Place the cursor over the sample symbol (Figure 9.17).
  The CHP file name of that particular sample is displayed.

Figure 9.17 Displaying the CHP file name for a particular sample.

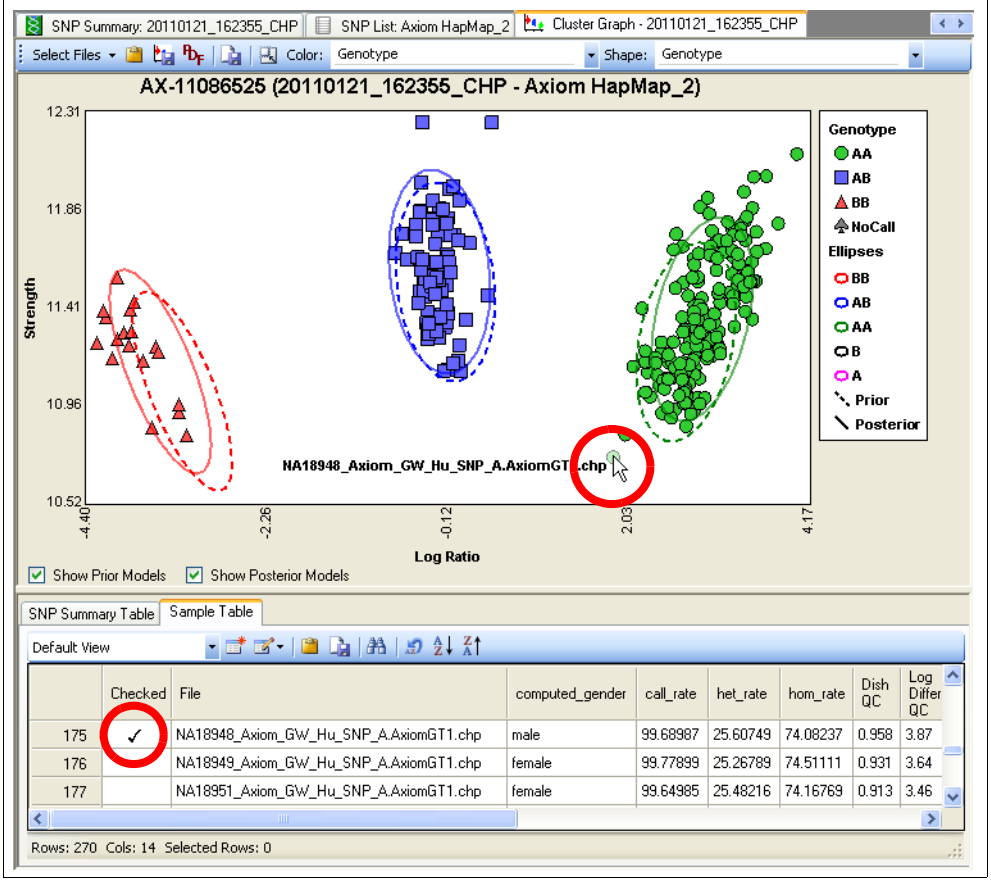


To select a single sample:

- Click on the data point in the SNP Cluster Graph (Figure 9.18, A).

  The CHP file corresponding to the selected sample is checked in the Sample Table.

Figure 9.18 Selecting a single sample



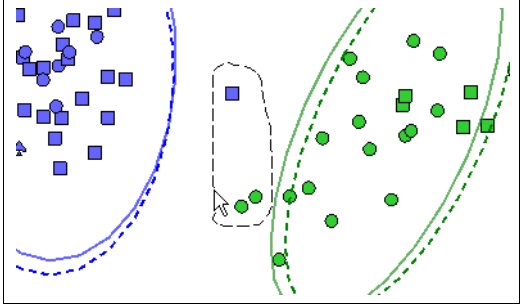
The CHP file corresponding to the selected sample is checked in the Checked column (Figure 9.18, B).

To select multiple samples (aka symbols):

1. Drag the cursor around the group of samples to draw a closed shape around them (Figure 9.19).

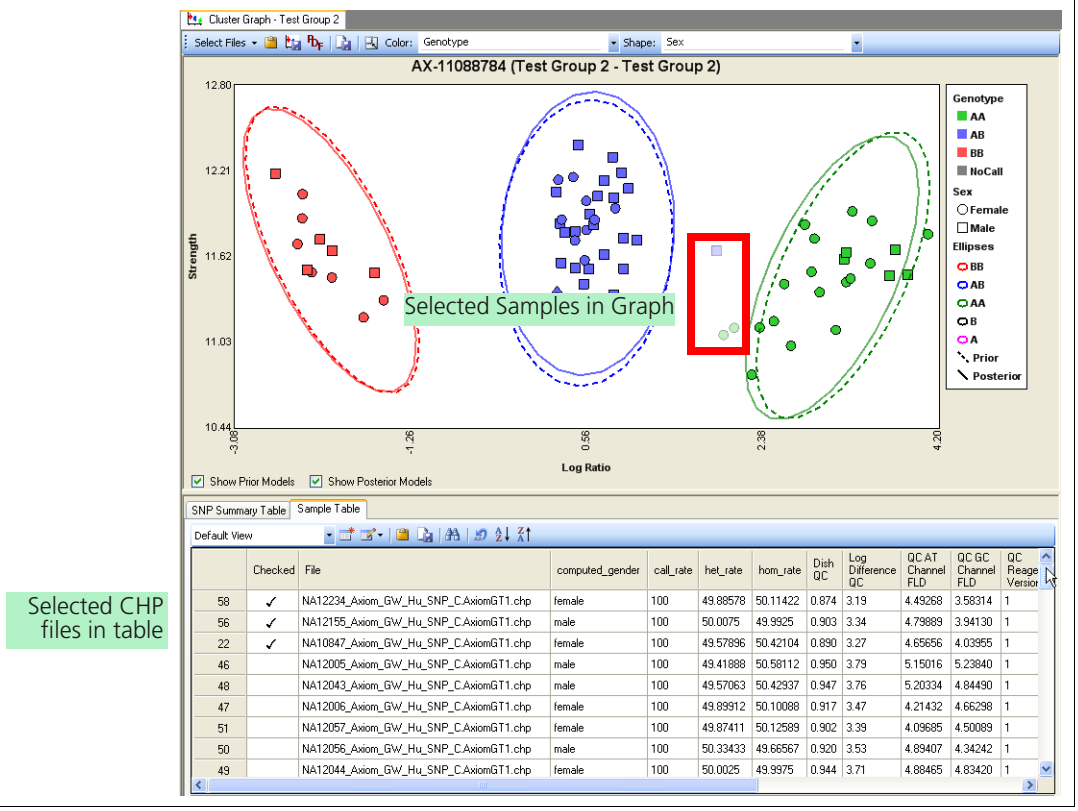
   The lasso function automatically draws a straight line to the starting point if you release the mouse button.

Figure 9.19 Selecting multiple samples



The samples in the group and their associated CHP files in the Sample table are selected when you release the button (Figure 9.20).

Figure 9.20 Selected samples and CHP files with multiple samples selected



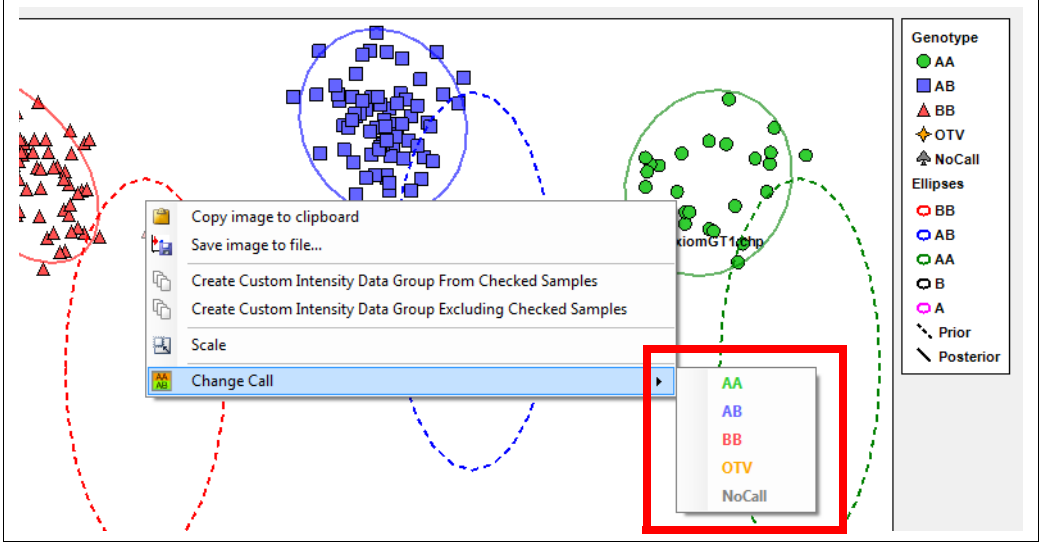
You may need to sort the Checked column in the Sample table to locate the selected files.

Changing a SNP's Call

To change a SNP's call:

1. Select the samples for which you want to modify the call, then right-click.
   A menu appears.
2. Click **Change Call**, then move your cursor to the right and click to select a different call, an OTV (Off Target Variant), or NoCall. (Figure 9.21)

Figure 9.21 Change Call menu



The SNP call is now changed and its new Call is reflected within the SNP Cluster Graph.

TIP: You can always revert your modified Calls back to their original values by clicking on the Cluster Graph's *Revert* button. See Table 9.1 on page 156.

SNP Summary Table

The SNP Summary Table (Figure 9.22) displays information on all the SNPs in the selected SNP List.

NOTE: SNP filtering uses the full precision of stored metrics. The displayed precision in tables is less than this for readability.

Figure 9.22 SNP Summary Table in the SNP Cluster Graph

SNP Summary Table | Sample Table

Default View

	Probe Set ID	SNP Call Rate	SNP %AA	SNP %AB	SNP %BB	Minor Allele Frequency (SNP Summary)	H.W. p-Value	dbSNP RS ID	Chr
1	AX-11086564	100	25.71428	67.14286	7.142857	0.4071428	0.001075898	rs1000057	5
2	AX-11086581	100	58.57143	32.85714	8.571428	0.25	0.3002652	rs1000081	5
3	AX-11086599	98.57	0	2.857143	95.71429	0.01449275	0.9027753	rs1000113	5
4	AX-11086764	100	14.28571	38.57143	47.14286	0.3357143	0.257953	rs1000381	5
5	AX-11086782	100	85.71429	14.28571	0	0.07142857	0.5198448	rs1000412	5
6	AX-11086885	100	7.142857	47.14286	45.71429	0.3071429	0.3677754	rs1000590	5
7	AX-11087218	100	38.57143	48.57143	12.85714	0.3714286	0.7365547	rs1001114	5
8	AX-11087405	100	51.42857	44.28571	4.285714	0.2642857	0.2454967	rs1001420	5
9	AX-11087537	98.57	0	10	88.57143	0.05072464	0.6571399	rs1001636	5
10	AX-11087862	100	88.57143	11.42857	0	0.05714286	0.6121081	rs1002176	5
11	AX-11088014	100	90	8.571428	1.428571	0.05714286	0.08701651	rs1002429	5
12	AX-11088085	100	58.57143	35.71429	5.714286	0.2357143	0.9414452	rs1002548	5

Rows: 35540 Cols: 12 Selected Rows: 0

The table displays the same information and has the same functions as the SNP Summary Table that is displayed after genotyping. See *SNP Summary Table* on page 131 for more information.

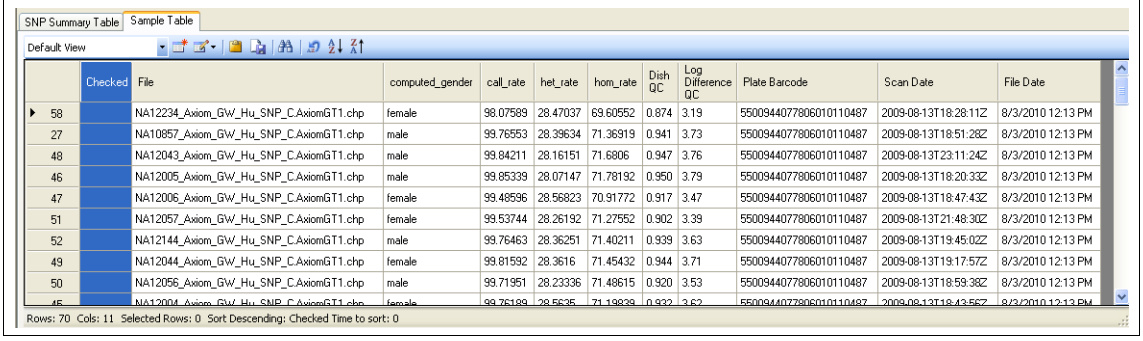
Step through the SNPs in the table by clicking on a line, or by pressing the down arrow button. The Cluster Graph will automatically update to the selected SNP.

Within the table, you can also add and remove SNPs from SNP lists. See *Adding and Removing SNPs from SNP Lists* on page 166 for more information.

Sample Table

The Sample Table (Figure 9.23) displays information on the samples from which the displayed SNPs are derived.

Figure 9.23 Sample Table in the SNP Cluster Graph



The SNP Cluster Graph Sample table has the same labels as the *CHP Summary Table* on page 110.

You can:

- Select a sample in the table and have it highlighted in the Cluster graph.
- Select samples in the cluster graph and have the corresponding CHP files checked in the table.

The highlighting of selected samples persists as you move from SNP to SNP.

The Checked column in the Sample Table enables you to select individual CHP files for additional analyses and manipulation, such as creating custom intensity groups (below).

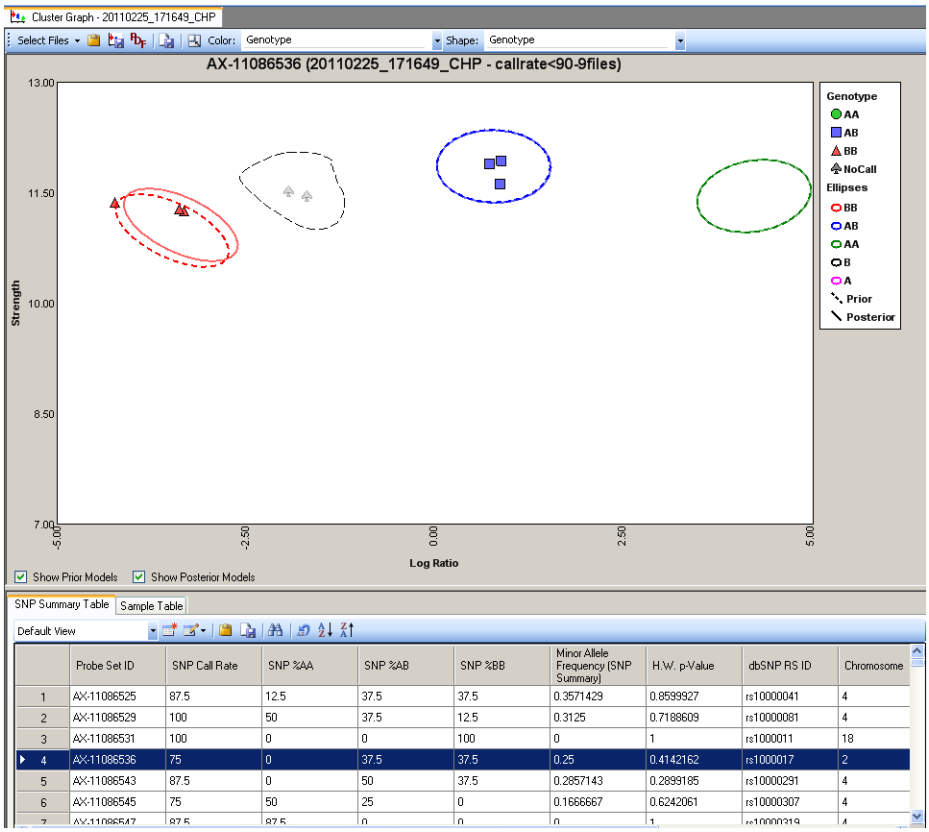
Creating Custom Intensity Data Groups Using the SNP Cluster Graph

The lasso option allows you to select a group of samples of interest in the cluster graph.

You can use this option as the basis for either the inclusion or exclusion of certain samples and create a custom intensity group for a second round of genotyping to obtain optimum genotyping performance.

The following section describes how to create a custom intensity data group based on information displayed in the SNP Cluster Graph. In this example, a custom intensity data group is created to specifically exclude the outlier samples (Figure 9.24).

Figure 9.24 SNP Cluster Graph, two samples have NoCalls



	Probe Set ID	SNP Call Rate	SNP %AA	SNP %AB	SNP %BB	Minor Allele Frequency (SNP Summary)	H.W. p-Value	dbSNP RS ID	Chromosome
1	AX-11086525	87.5	12.5	37.5	37.5	0.3571429	0.8599927	rs10000041	4
2	AX-11086529	100	50	37.5	12.5	0.3125	0.7188609	rs10000081	4
3	AX-11086531	100	0	0	100	0	1	rs1000011	18
4	AX-11086536	75	0	37.5	37.5	0.25	0.4142162	rs1000017	2
5	AX-11086543	87.5	0	50	37.5	0.2857143	0.2899185	rs10000291	4
6	AX-11086545	75	50	25	0	0.1666667	0.6242061	rs10000307	4

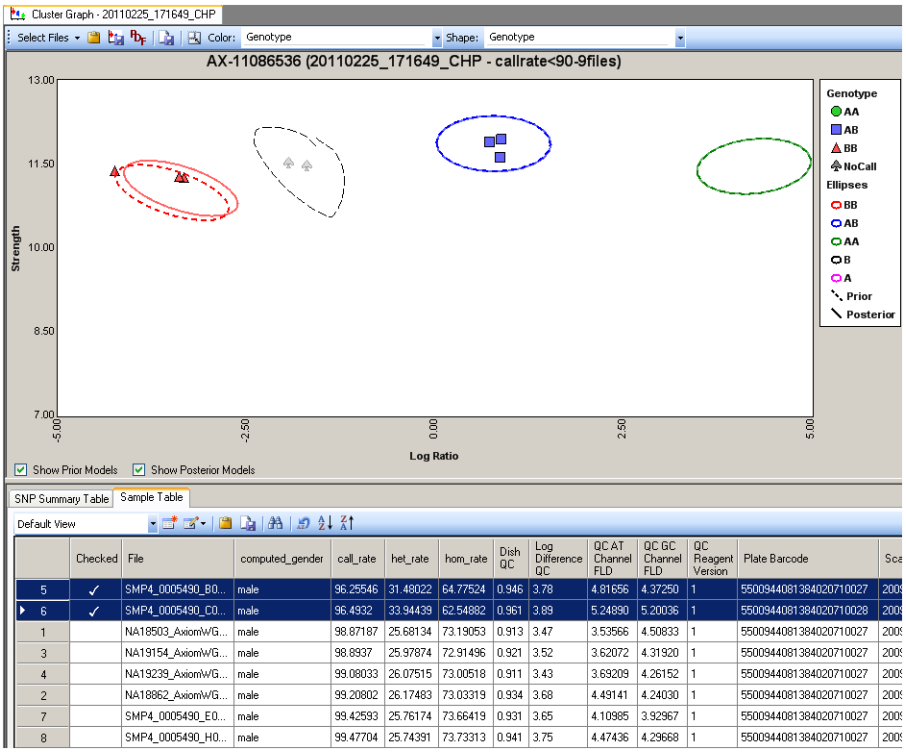
In this example, two samples have no calls for the SNP AX-11086536 (gray spade). Two samples lie outside the cluster ellipses defined by dashed lasso lines.

To create a Custom Intensity Data Group:

1. Draw a closed shaped around the samples you want to select by dragging the cursor around the samples (Figure 9.25).

   The lasso function automatically draws a straight line to the starting point if you release the mouse button.

Figure 9.25 Selecting samples for exclusion



	Checked	File	computed_gender	call_rate	het_rate	hom_rate	Dish QC	Log Difference QC	QC AT Channel FLD	QC GC Channel FLD	QC Reagent Version	Plate Barcode	Sca
5	✓	SMP4_0005490_B0...	male	96.25546	31.48022	64.77524	0.946	3.78	4.81656	4.37250	1	55009440813840207100 27	2009
6	✓	SMP4_0005490_C0...	male	96.4932	33.94439	62.54882	0.961	3.89	5.24890	5.20036	1	55009440813840207100 28	2009
1		NA18503_AxiomWG...	male	98.87187	25.68134	73.19053	0.913	3.47	3.53566	4.50833	1	55009440813840207100 27	2009
3		NA19154_AxiomWG...	male	98.8937	25.97874	72.91496	0.921	3.52	3.62072	4.31920	1	55009440813840207100 27	2009
4		NA19239_AxiomWG...	male	99.08033	26.07515	73.00518	0.911	3.43	3.69209	4.26152	1	55009440813840207100 27	2009
2		NA18862_AxiomWG...	male	99.20802	26.17483	73.03319	0.934	3.68	4.49141	4.24030	1	55009440813840207100 27	2009
7		SMP4_0005490_E0...	male	99.42593	25.76174	73.66419	0.931	3.65	4.10985	3.92967	1	55009440813840207100 27	2009
8		SMP4_0005490_H0...	male	99.47704	25.74391	73.73313	0.941	3.75	4.47436	4.29668	1	55009440813840207100 27	2009

Samples corresponding to those selected symbols will have the Checked cell marked in the Sample table.

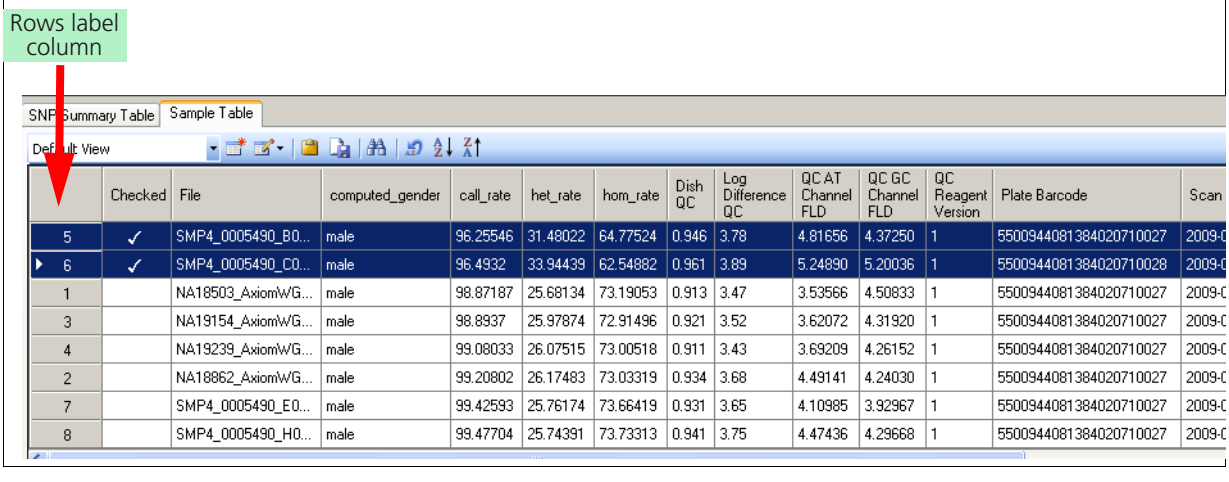
2. Select the Checked column header and select the **Sort Ascending** or **Sort Descending** buttons in the Sample Table tool bar.

   The CHP files will be grouped by their "checked" column status.

3. Select the rows in the table that you wish to include or exclude.

   You must select the rows by clicking in the rows label column (Figure 9.26).

Figure 9.26 Selecting CHP files



	Checked	File	computed_gender	call_rate	het_rate	hom_rate	Dish QC	Log Difference QC	QC AT Channel FLD	QC GC Channel FLD	QC Reagent Version	Plate Barcode	Scan
5	✓	SMP4_0005490_B0...	male	96.25546	31.48022	64.77524	0.946	3.78	4.81656	4.37250	1	55009440813840207100 27	2009-0
6	✓	SMP4_0005490_C0...	male	96.4932	33.94439	62.54882	0.961	3.89	5.24890	5.20036	1	55009440813840207100 28	2009-0
1		NA18503_AxiomWG...	male	98.87187	25.68134	73.19053	0.913	3.47	3.53566	4.50833	1	55009440813840207100 27	2009-0
3		NA19154_AxiomWG...	male	98.8937	25.97874	72.91496	0.921	3.52	3.62072	4.31920	1	55009440813840207100 27	2009-0
4		NA19239_AxiomWG...	male	99.08033	26.07515	73.00518	0.911	3.43	3.69209	4.26152	1	55009440813840207100 27	2009-0
2		NA18862_AxiomWG...	male	99.20802	26.17483	73.03319	0.934	3.68	4.49141	4.24030	1	55009440813840207100 27	2009-0
7		SMP4_0005490_E0...	male	99.42593	25.76174	73.66419	0.931	3.65	4.10985	3.92967	1	55009440813840207100 27	2009-0
8		SMP4_0005490_H0...	male	99.47704	25.74391	73.73313	0.941	3.75	4.47436	4.29668	1	55009440813840207100 27	2009-0

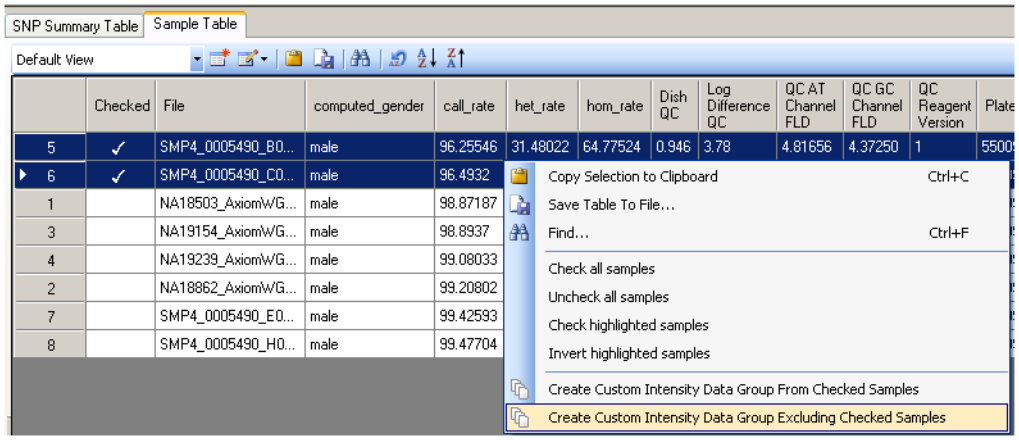
Select contiguous rows by clicking in the top and bottom rows while holding down the Shift key. Select multiple non-contiguous roles by clicking in the rows while holding down the CTRL key.

You can also select samples using the Sample table shortcut menu (right-click the Sample table):

- **Check all samples** – Puts a check mark next to all samples.
- **Uncheck all samples** – Removes all check marks.
- **Check highlighted samples** – Puts a check mark next to user-selected rows.
- **Invert highlighted samples** – Removes check marks from user-selected rows or adds check marks to user-selected rows.

4. Right-click on the Sample Table and select **Create Custom Intensity Data Group Excluding Checked Samples** (Figure 9.27).

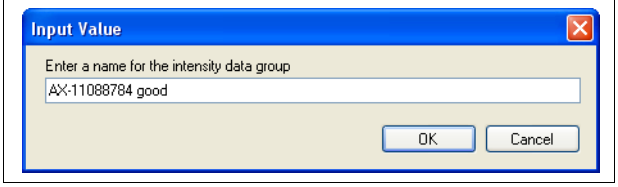
Figure 9.27 Sample Table with shortcut menu



You can also use **Create Custom Intensity Data Group Including Checked Samples**, depending upon whether you want to include or exclude the selected samples from the custom intensity data group.

The Enter a name for the intensity data group dialog box appears (Figure 9.28)

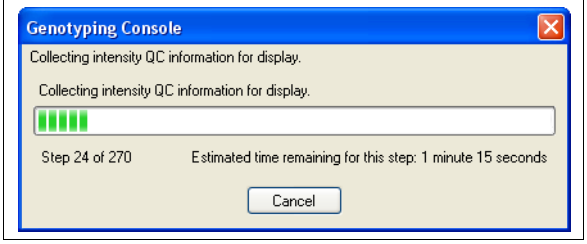
Figure 9.28 Input Value dialog box



5. Enter a name for the intensity data group and click **OK**.

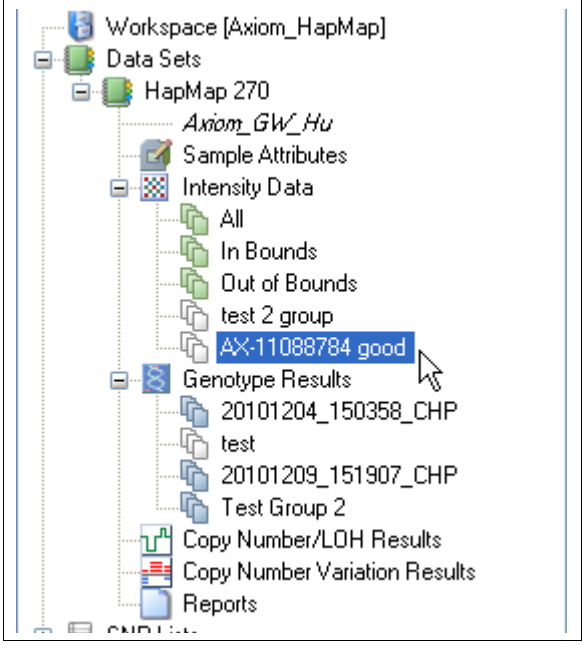
   A progress bar appears (Figure 9.29).

Figure 9.29 Progress bar for custom intensity data group creation



The new intensity data group (white icon) is displayed in the Data Tree (Figure 9.30) and the CEL files are listed in the Intensity QC Data Table.

Figure 9.30 GTC data tree with the custom intensity data group



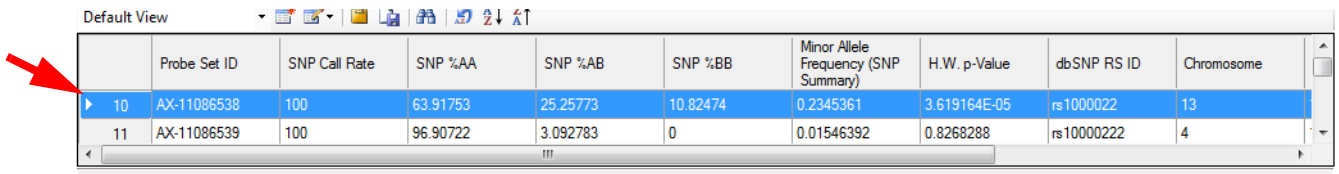
After the outliers have identified and/or removed, you can perform a second round of genotyping to get the optimal call rate.

Adding and Removing SNPs from SNP Lists

To add a SNP to a SNP list:

1. Single-click on row's entry number (farthest left column). (Figure 9.31)
   The row is now highlighted.

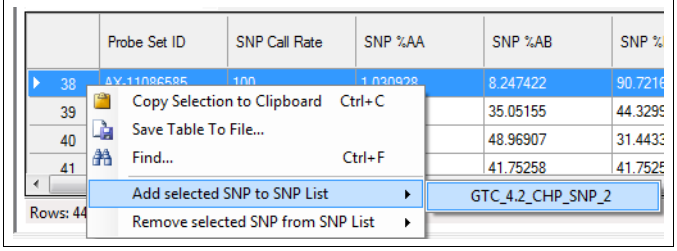
Figure 9.31 Highlighting a row



2. Right-click on the highlighted row.

A menu appears. (Figure 9.32)

Figure 9.32 Add selected SNP to SNP List



3. Click on **Add selected SNP to SNP List**, then move your cursor to the right and click onto the SNP list.

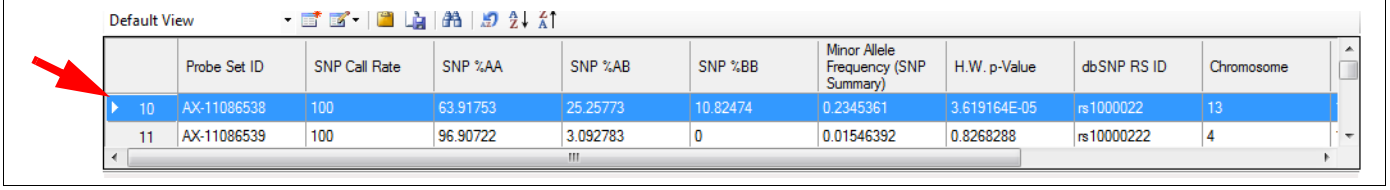
   Your selected SNP is now added to the SNP list.

To remove a SNP from a SNP list:

1. Single-click on row's entry number (farthest left column). (Figure 9.31)

   The row is now highlighted.

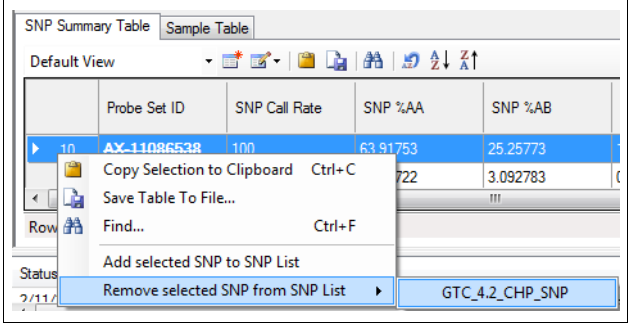
Figure 9.33 Highlighting a row



2. Right-click on the highlighted row.

   A menu appears. (Figure 9.32)

Figure 9.34 Remove selected SNP from a SNP List



3. Click on **Remove selected SNP from SNP List**, then move your cursor to the right and click on the SNP list you want to remove the highlighted SNP from.

NOTE: You cannot remove a SNP from a SNP List that is currently displayed in the Cluster Graph.

Changing the Display

Users can change the appearance of the SNP Cluster Graph using the following options:

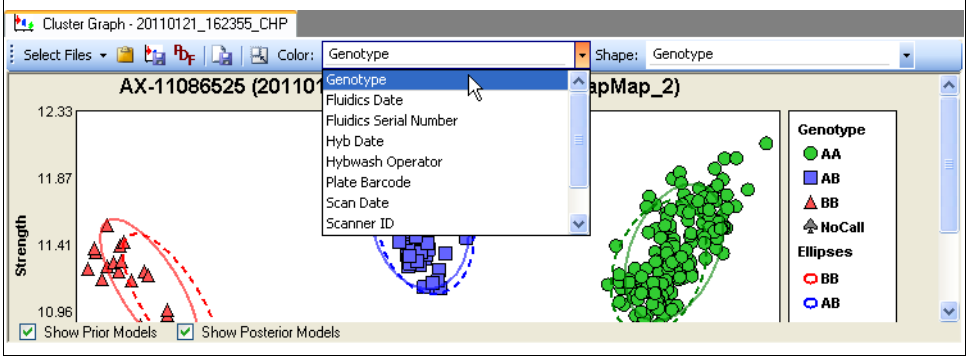
Selecting Colors and Shapes for Attributes

In the cluster graph, user-selected colors and shapes can be assigned to:

- Genotype and gender call data
- all user attributes
- array plate information
- fluidics instrument information
- scanner information (if available)

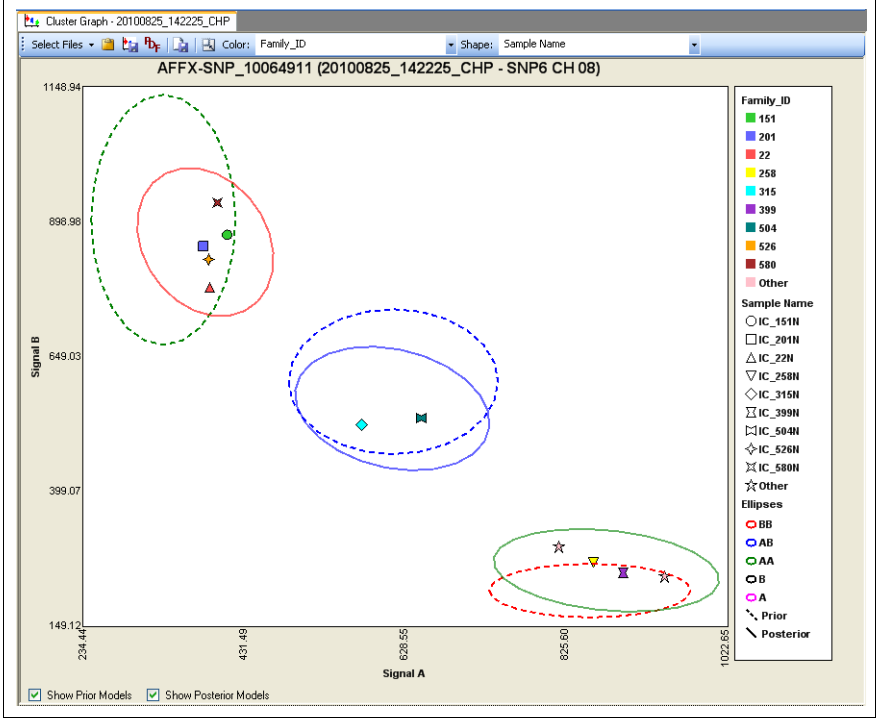
To change the color or shape assigned to an attribute, make selections from the Color or Shape drop-down lists (Figure 9.35).

Figure 9.35 Selecting Genotype call as the color attribute



The figure below (Figure 9.36) shows a SNP Cluster Graph where the user has customized the display to include information on the Family ID and sample name attributes.

Figure 9.36 Multiple Attributes displayed by color and shape



Some attributes are provided by default by the GTC software and the CHP file. Other attributes are derived from ARR files.

The ARR file attributes are only available in the drop-down lists if the files are available for the sample data. If an attribute is missing in a particular file, the SNP cluster Graph assigns shapes and colors as shown in the table below (Table 9.2).

Table 9.2 Color and shape assignments with attribute information missing

	Shape attribute available	**Shape attribute missing**
Color attribute available	SNP call marked by shape and color	colored spade shape
Color attribute missing	Gray attribute shape	Gray spade shape

If an attribute has more than 10 values:

- When the attribute value is text, the software takes the first nine values and assigns each a color or shape. The remaining values are put into a bin called “Other”. All values in the Other bin have the same color or shape.
- When the attribute value is a date or number, the software divides the range of data into 10 equal bins and assigns a color or shape to each bin. If the data includes one or more outliers, it is possible to have one value in a particular bin and all other values in another bin.

Selecting Model Display Options

GTC 4.2 uses model files for the following genotyping analyses:

- BRLMM-P
- Birdseed v1 and V2
- Axiom GT1

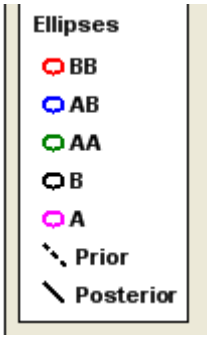
NOTE: The SNP Cluster Graph does not display model files for BRLMM (100K and 500K) analyses.

These model files contain cluster location information that is used in generating genotyping calls.

Borders of clusters can be displayed for individual SNPs in the SNP Cluster Graph.

The colors and lines used for different models are displayed in the Legend box of the SNP Cluster Graph (Figure 9.37).

Figure 9.37 Model Ellipses legends as shown in the SNP Cluster Graph

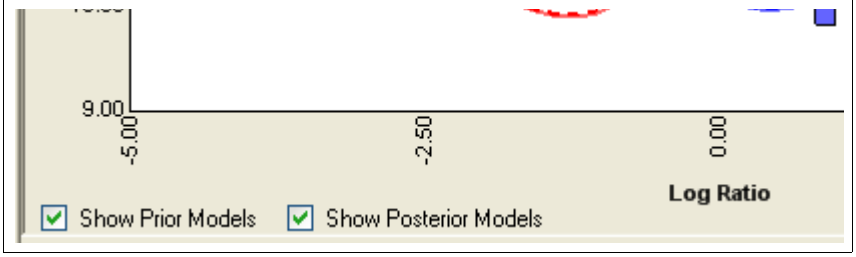


See *Model Files Options* on page 104 for more information about the use of model files.

To display or conceal the cluster model data:

- Select or deselect the Show Prior Models and/or Show Posterior Models checkboxes in the graph (Figure 9.38).

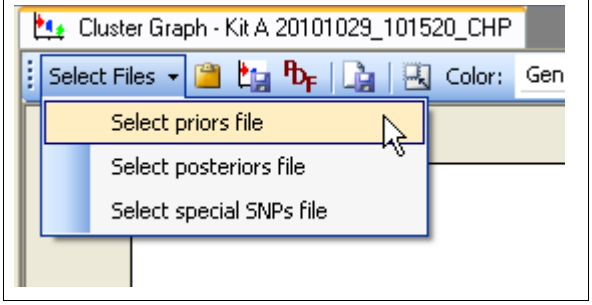
Figure 9.38 Model Ellipses checkboxes for displaying models



To select prior model files for display:

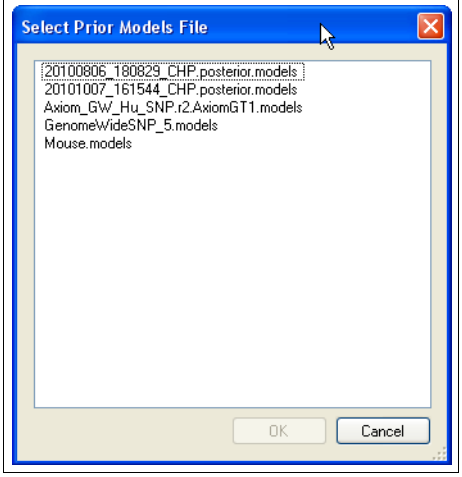
1. Make sure the model files are in the GTC Library folder.
2. Choose **Select priors file** from the Select Files menu (Figure 9.39).

Figure 9.39 Select Files menu



3. The Select Prior Model File dialog box opens (Figure 9.40).

Figure 9.40 Select Prior Models File dialog box

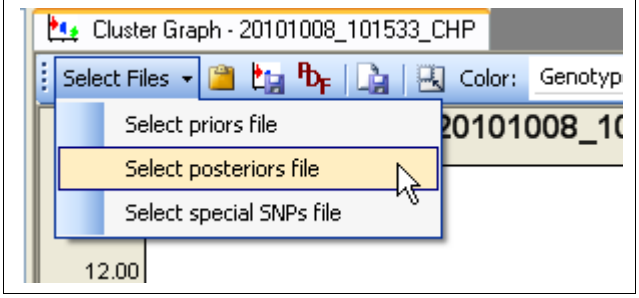


4. Select the desired file from the dialog box and click **OK**.
   The selected model file information is displayed in the Cluster graph.

To select posterior model files for display:

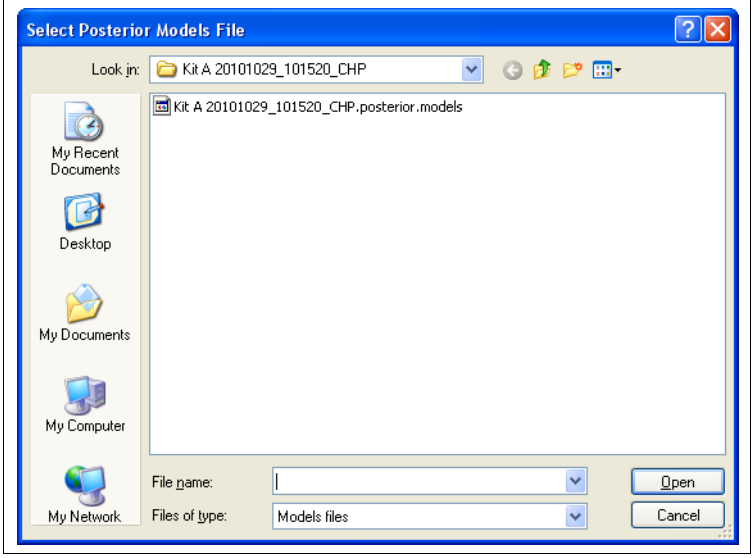
1. Choose **Select posteriors file** from the Select Files menu (Figure 9.41).

Figure 9.41 Select Files menu



2. The Select Posterior Model File dialog box opens (Figure 9.42).

Figure 9.42 Select Posterior Models File dialog box



NOTE: The Select Posterior Models File dialog box displays the contents of the Genotyping Results Group folder for the displayed results group.

3. Select the desired file from the dialog box and click **OK**.
   The selected model file information is displayed in the Cluster graph.

Selecting the Special SNPs File

The Special SNPs File provides a notice when SNPs from the following types of chromosomes and regions are displayed:

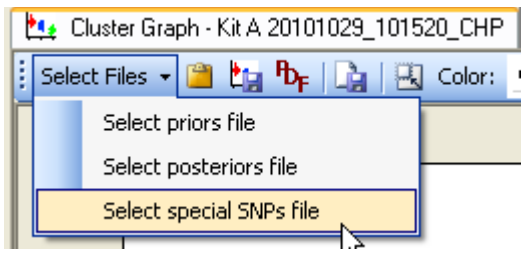
- Mitochondrial
- X
- Y
- PAR (PseudoAutosomal Region)

A default Special SNP file is loaded while creating the SNP Cluster Graph. You may select a different special SNP file using the Select Files drop-down menu.

To select Special SNP files for display:

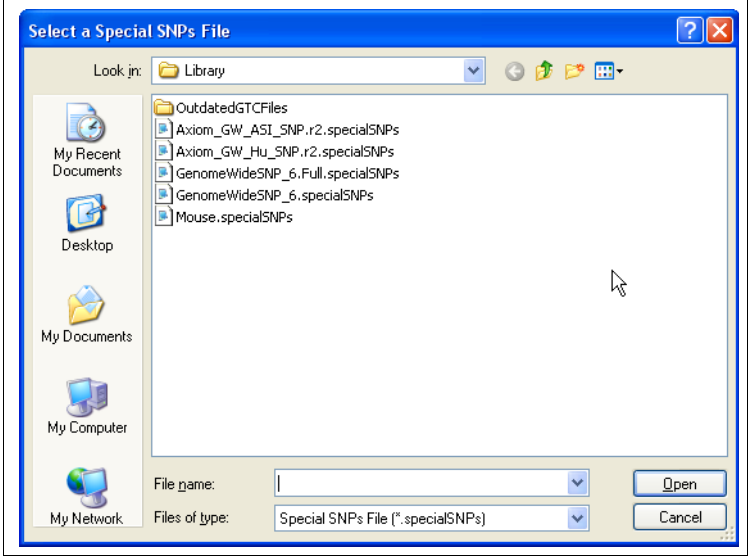
1. Choose **Select special SNPs file** from the Select Files menu (Figure 9.43).

Figure 9.43 Select Files menu



2. The Select a Special SNPs File dialog box opens (Figure 9.44).

Figure 9.44 Select a Special SNPs File dialog box



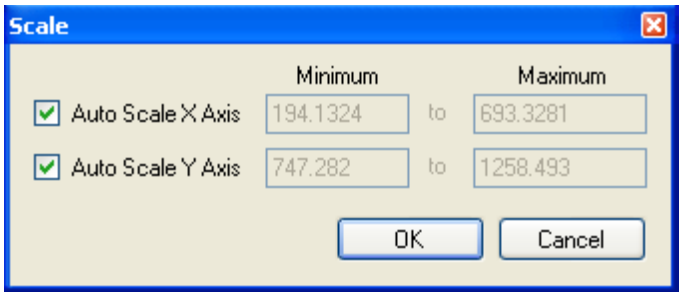
3. Select the desired file from the dialog box and click **OK**.
   Additional information is displayed when a SNP in the Special SNP file is selected for display.

Change the Scale of the SNP Cluster Graph Axes

To change the scale of the graph axes:

1. Click the **Set Axis Scale** shortcut on the graph tool bar. Alternately, right-click the graph and select Scale on the shortcut menu.
2. In the Scale dialog box that appears (), enter values for the x and y-axis minimum and maximum.
3. To automatically scale the axes, choose the Auto Scale X Axis and Auto Scale Y Axis options.
   Auto-scaling sets the graph width to include all sample symbols (Figure 9.45).

Figure 9.45 Scale dialog box



Saving Cluster Graph Information

You can save:

- *The actual SNP Cluster Graph image for use as an illustration* (below)
- *SNP data as a tab-delimited text file* (page 179)

NOTE: You can also save data from the SNP Summary Table and Sample Table (see ***Table Features*** on page 198).

Saving the SNP Cluster Graph Image

You can use the following options to save the SNP Cluster Graph image:

- *Copy Image to Clipboard* on page 174
- *Save Image to File* on page 175
- *Save All SNP Cluster Graphs to PDF File* on page 176

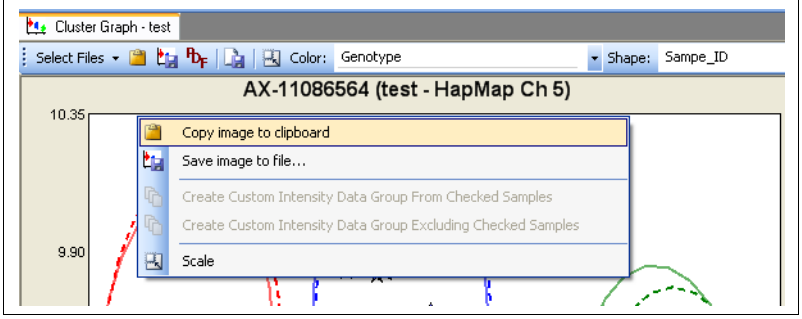
Copy Image to Clipboard

You can save the SNP Cluster Graph image to the Clipboard and then paste it into a graphics program such as Paint for use in a document.

To save the image to the Clipboard:

- Click the Save to Clipboard button on the SNP Cluster Graph tool bar; or
  Right click on the SNP Cluster Graph and select Copy image to Clipboard (Figure 9.46).

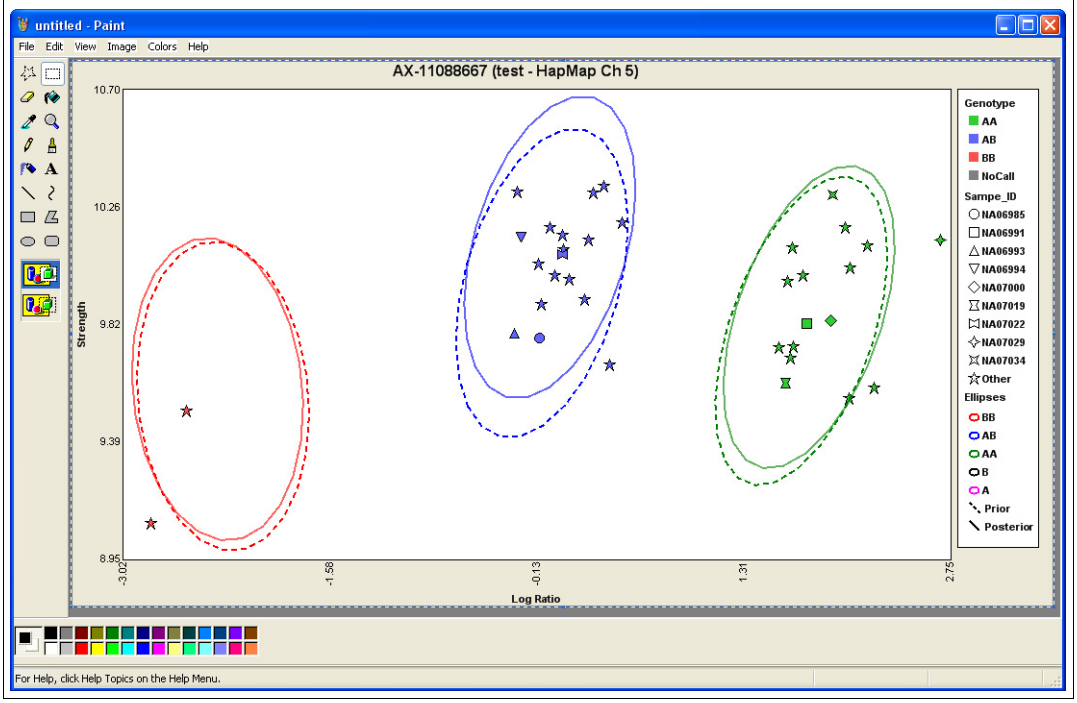
Figure 9.46 SNP Cluster Graph right-click menu



NOTE: If you right-click in one of the SNP Cluster Graph tables you will access a different set of functions.

The image of the SNP Cluster Graph is copied to the Clipboard and can be pasted into a graphics program such as Paint (Figure 9.47).

Figure 9.47 Paint software with image pasted in



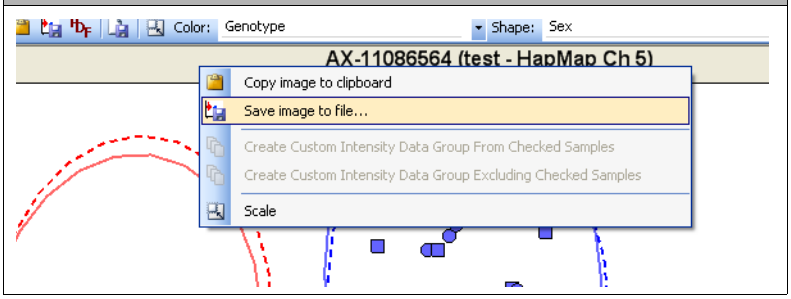
Save Image to File

You can save the SNP Cluster Graph image as a PNG file for use in other documentation.

To save the image as a graphics file:

1. Click the Save Image to File button on the SNP Cluster Graph tool bar; or Right-click the SNP Cluster Graph and select Save image to file (Figure 9.48).

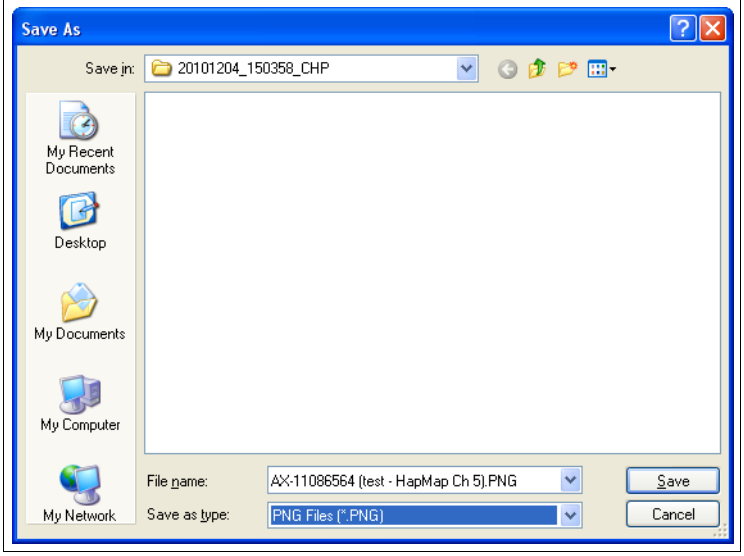
Figure 9.48 SNP Cluster Graph right-click menu



NOTE: If you right-click in one of the SNP Cluster Graph tables you will access a different set of functions.

The Save As dialog box opens (Figure 9.49).

Figure 9.49 Save As dialog box for Cluster Graph



The dialog box automatically opens to the folder for the genotyping results and with the default name of the file displayed using the following format:

`SNP ID (Genotyping Results batch name - SNP list name)`

You can change the file location and name in the dialog box.

2. Click **Save**.

   The file is saved in the selected folder.

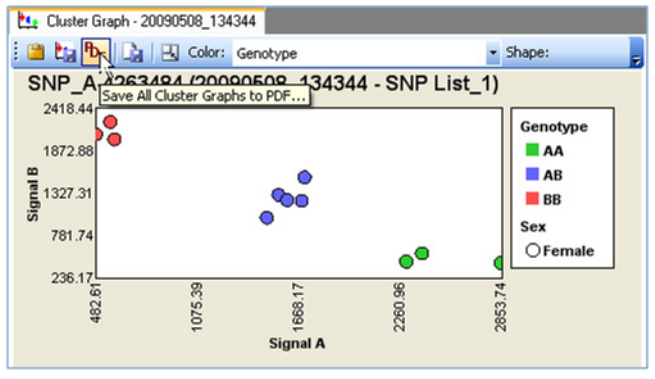
Save All SNP Cluster Graphs to PDF File

You can save the cluster graph visualizations for all SNPs in a SNP List to a single PDF file.

To save to a PDF:

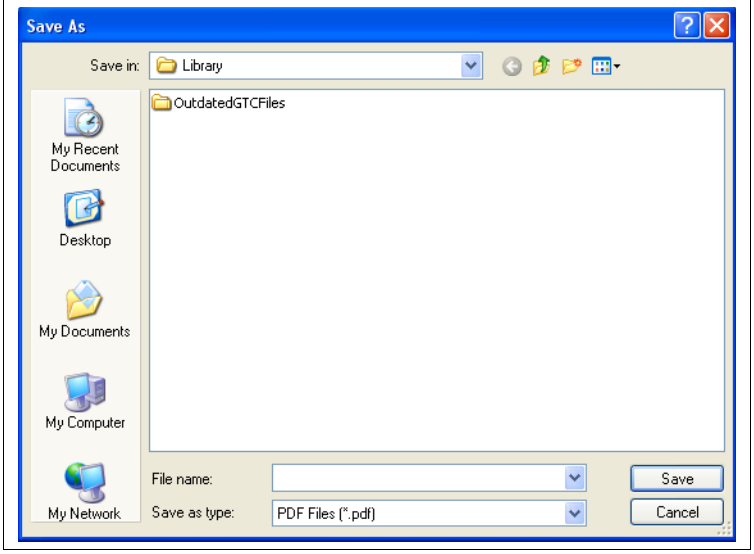
1. Click on the **Save All Cluster Graphs to PDF** shortcut on the SNP Cluster Graph tool bar (Figure 9.50).

Figure 9.50 Save to PDF button in graph tool bar



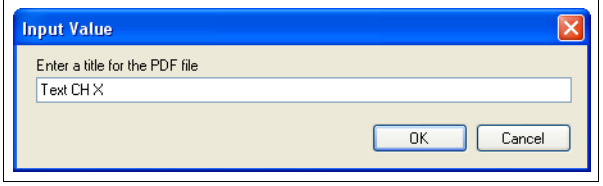
The Save As dialog box opens (Figure 9.51).

Figure 9.51 Save As dialog box



2. Select a location to save this file and enter a name for the file.
3. Click **Save**.
   The Enter a title for the PDF file dialog box opens (Figure 9.52).

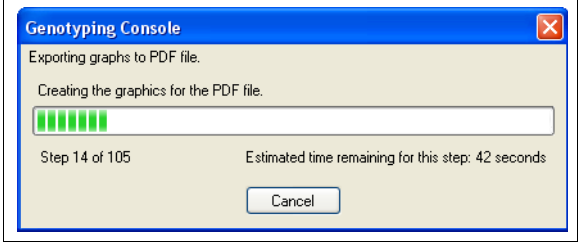
Figure 9.52 Enter a title for the PDF file dialog box



NOTE: The PDF title has a 55 character limit.

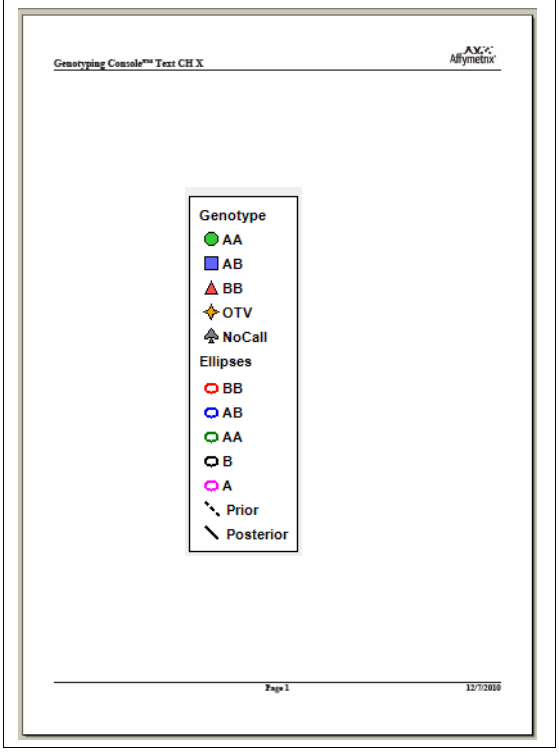
4. Enter a title for the PDF file. This title will be displayed at the top of every page in the PDF document.
5. Click **OK** in the Enter a title for the PDF file dialog box.
   A progress bar displays the progress of the export (Figure 9.53).

Figure 9.53 Progress dialog box



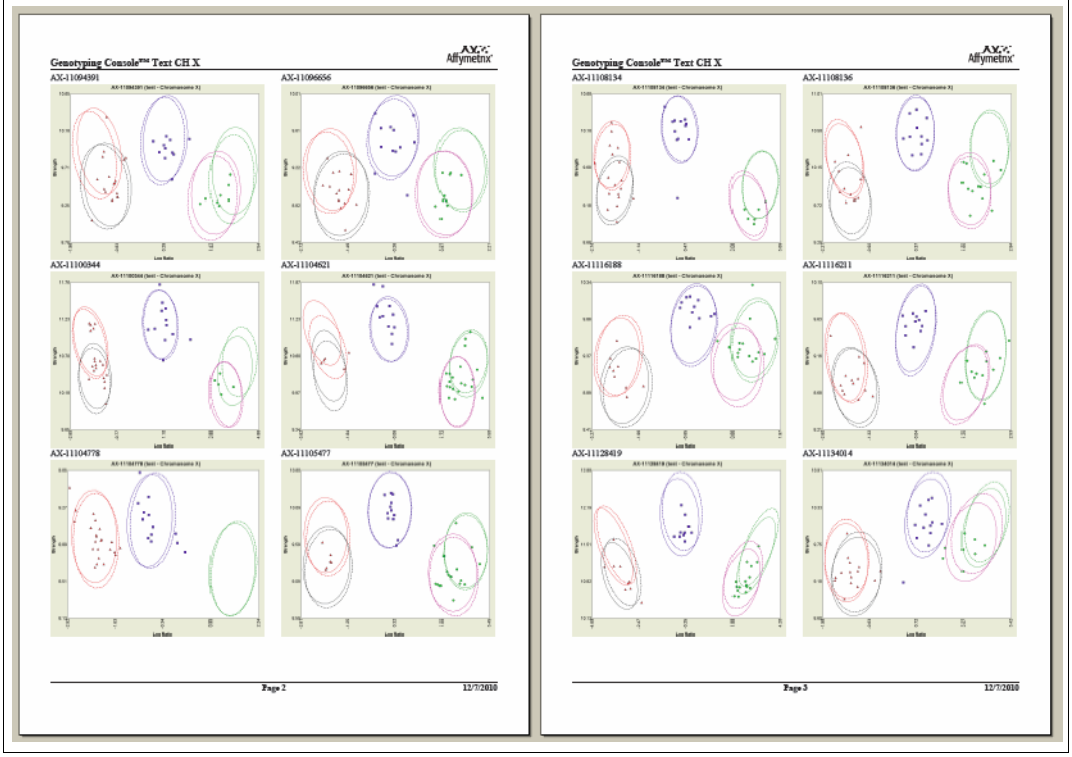
The first page of the PDF displays the Legend for the SNP Cluster Graph (Figure 9.54).

Figure 9.54 Legend information for SNP Cluster Graph



The remaining pages display six graphs per page (Figure 9.55).

Figure 9.55 Pages two and 3 of the PDF file



Save SNP Data from the SNP Cluster Graph

This feature saves SNP data for SNPs displayed in the SNP Cluster Graph in a tab-delimited text file (Figure 9.56) with the following column headers:

- SNP ID
- CHP Name
- Genotype
- The X and Y axes values plotted for the algorithm type:
  - *BRLMM and BRLMM-P Data* on page 149
  - *Birdseed Data* on page 150
  - *Axiom Data* on page 151

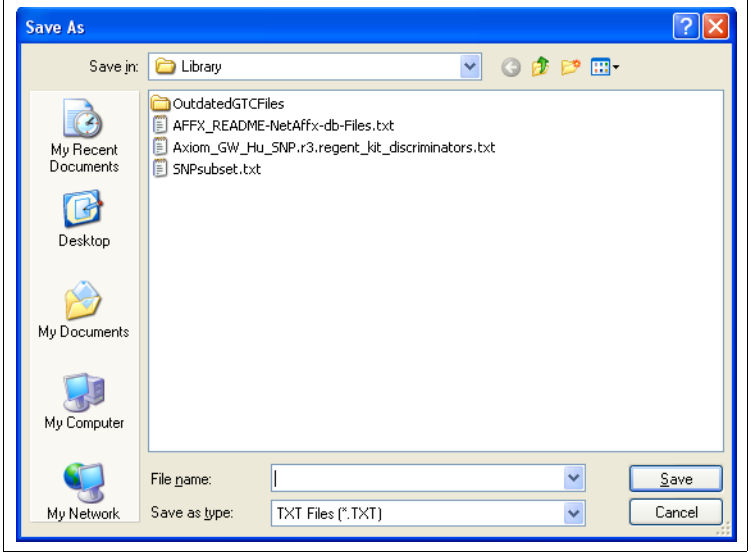
Figure 9.56 .TXT file open in Microsoft Excel

	A	B	C	D	E
1	SNP ID	CHP Name	Log Ratio	Strength	Genotype
2	AX-11094391	NA11839_Axiom_GW_Hu_SNP_C.AxiomGT1.chp	-1.115201	9.405842	BB
3	AX-11094391	NA11832_Axiom_GW_Hu_SNP_C.AxiomGT1.chp	-0.6714983	9.81592	BB
4	AX-11094391	NA11831_Axiom_GW_Hu_SNP_C.AxiomGT1.chp	-0.8731976	9.544272	BB
5	AX-11094391	NA11830_Axiom_GW_Hu_SNP_C.AxiomGT1.chp	0.497673	9.902536	AB
6	AX-11094391	NA11829_Axiom_GW_Hu_SNP_C.AxiomGT1.chp	-1.4378	9.055407	BB
7	AX-11094391	NA10863_Axiom_GW_Hu_SNP_C.AxiomGT1.chp	-1.08704	10.35621	BB
8	AX-11094391	NA10861_Axiom_GW_Hu_SNP_C.AxiomGT1.chp	0.3205709	9.833857	AB
9	AX-11094391	NA10860_Axiom_GW_Hu_SNP_C.AxiomGT1.chp	1.625024	9.360554	AA
10	AX-11094391	NA10859_Axiom_GW_Hu_SNP_C.AxiomGT1.chp	-1.310368	9.562704	BB
11	AX-11094391	NA10857_Axiom_GW_Hu_SNP_C.AxiomGT1.chp	-0.9525709	9.304726	BB
12	AX-11094391	NA10856_Axiom_GW_Hu_SNP_C.AxiomGT1.chp	1.861628	9.62349	AA
13	AX-11094391	NA10855_Axiom_GW_Hu_SNP_C.AxiomGT1.chp	-1.150822	9.917782	BB
14	AX-11094391	NA10854_Axiom_GW_Hu_SNP_C.AxiomGT1.chp	-0.8145752	9.336191	BB
15	AX-11094391	NA10851_Axiom_GW_Hu_SNP_C.AxiomGT1.chp	-0.8901691	9.513539	BB
16	AX-11094391	NA10847_Axiom_GW_Hu_SNP_C.AxiomGT1.chp	-0.6174002	9.837465	BB
17	AX-11094391	NA10846_Axiom_GW_Hu_SNP_C.AxiomGT1.chp	1.592852	9.236629	AA
18	AX-11094391	NA10839_Axiom_GW_Hu_SNP_C.AxiomGT1.chp	0.3258762	9.957435	AB
19	AX-11094391	NA10838_Axiom_GW_Hu_SNP_C.AxiomGT1.chp	-0.8368301	9.386869	BB
20	AX-11094391	NA10835_Axiom_GW_Hu_SNP_C.AxiomGT1.chp	-1.024253	9.596769	BB
21	AX-11094391	NA10831_Axiom_GW_Hu_SNP_C.AxiomGT1.chp	-1.183387	9.851353	BB

To save SNP data to a .TXT file:

1. Click on the **Save Data to File** button  on the SNP Cluster Graph tool bar. The Save As dialog box opens (Figure 9.57).

Figure 9.57 Save As dialog box



2. Select a location and enter a name for the text file.
3. Click **Save**.

   The text file is saved in the designated location.

   The text file can be opened with text editing or spreadsheet software.

You can also export data from the tables using the table functions (see *Table Features* on page 198).

Exporting Genotype Results

You can export genotype results in the following ways:

- *Export genotypes to TXT format* on page 181
- *Export the Combined Results of an Array Set* on page 187
- *Export Genotype Results for PLINK* on page 190

Export genotypes to TXT format

Genotyping Results can be exported into a tab-delimited text file or a set of files.

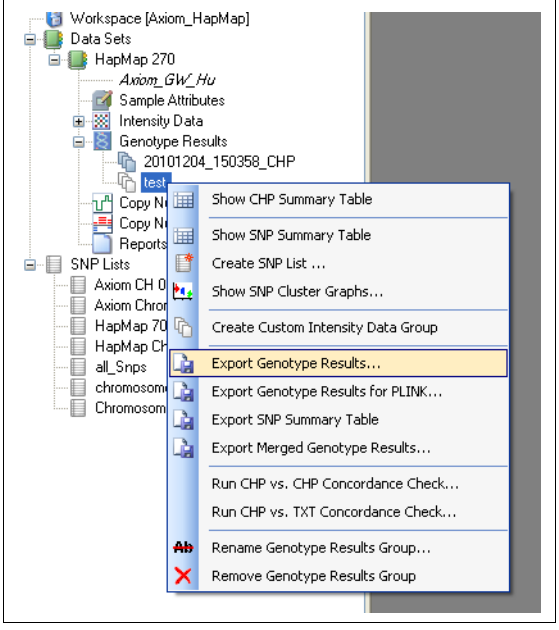
The contents of the files vary depending upon:

- The array and algorithm used to collect the data
- The options selected for the export

To export genotype results to TXT format:

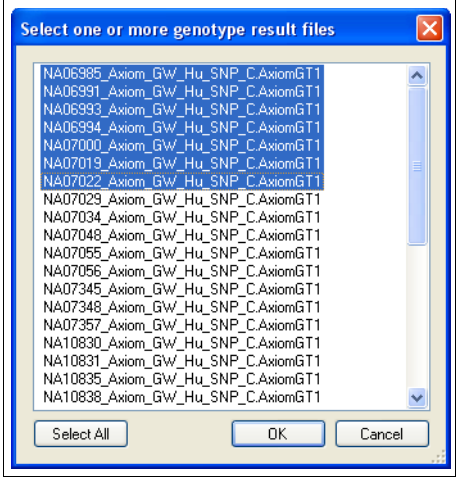
1. Do one of the following:
   - Right-click a Genotype Results group.
     1) Select **Export Genotype Results on the shortcut menu** (Figure 10.1).

Figure 10.1 Genotype results shortcut menu



The Select one or more genotype result files dialog box opens (Figure 10.2).

Figure 10.2 Select one or more genotype result files

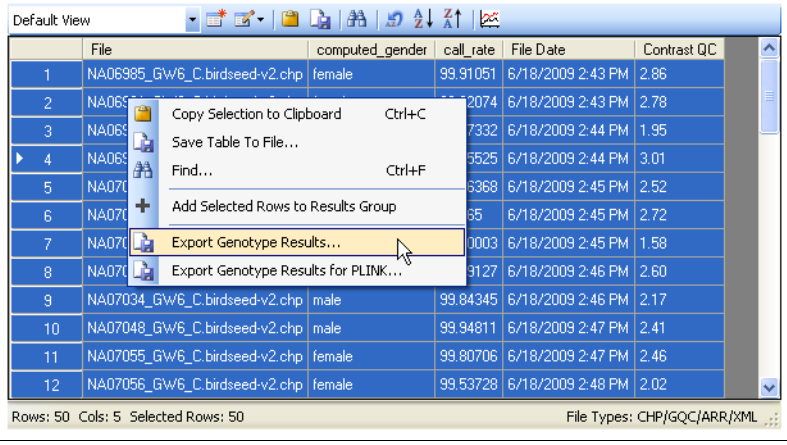


2) Select files in the dialog box and click **OK**.
   Select All selects all files.

Or

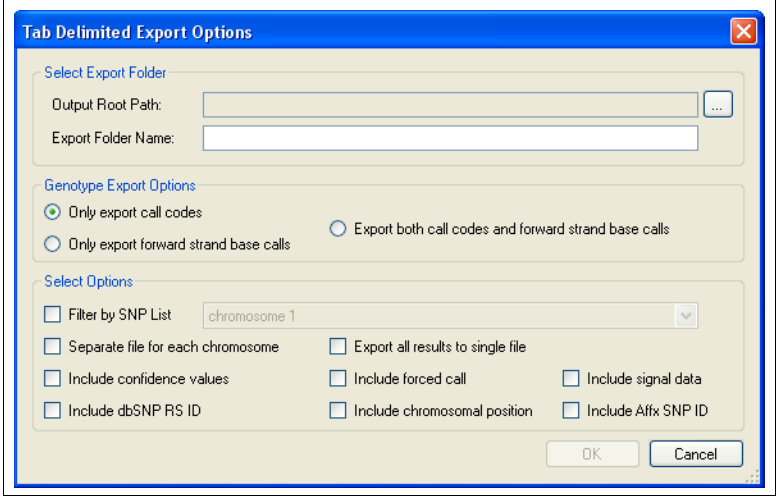
- select results (rows) in the CHP Summary table (Figure 10.3), right-click the selection, and choose Export Genotype Results on the shortcut menu.

Figure 10.3 Selecting files for export from the CHP Summary table



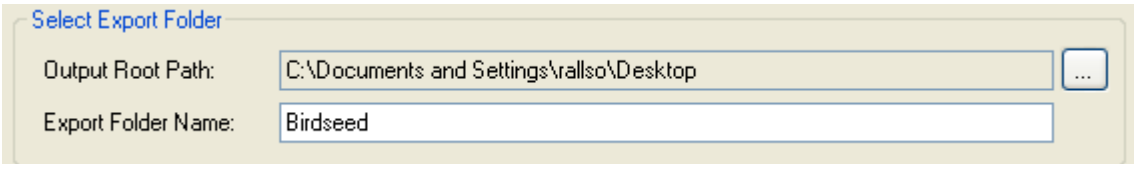
The Tab Delimited Export Options dialog box opens (Figure 10.4).

Figure 10.4 Export Options dialog box



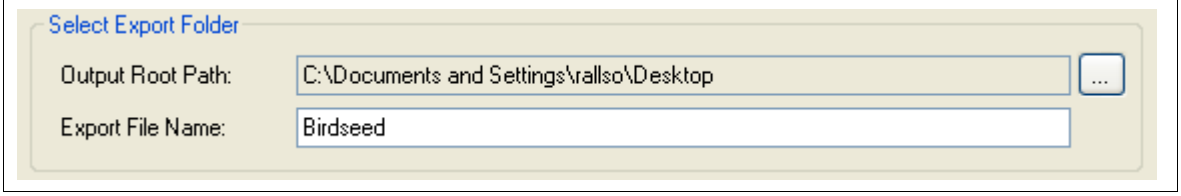
2. Click the Browse button ... to select the output directory.
3. Enter a name for the file or folder:
   - Export *Folder* Name (Figure 10.5) if **Export all results to single file** is *not* selected.

Figure 10.5 Select Export Folder Name



   - Export *File* Name (Figure 10.6) if **Export all results to single file** *is* selected.

Figure 10.6 Select Export File Name



4. Choose the Genotype Export options (Figure 10.7), as described in the table below (Table 10.1).

Figure 10.7 Genotype Export Options

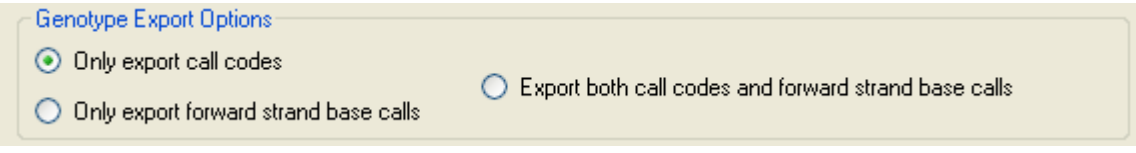


Table 10.1 Genotype Export Options for Tab Delimited Export

Genotype Export Options	Description
Only export call codes	Choose this option to include only the allele call codes (AA, AB, or BB) in the text file.
Only export forward strand base calls	Choose this option to include only the forward strand base calls (AT, CG, AG, TC, --, etc) in the text file.
Export both call codes and forward strand base calls	Choose this option to include both the allele call codes and the forward strand base calls in the text file. For more details on forward strand base call translation, see Appendix B, *Forward Strand Translation* on page 346.

5. Select the Select options (Figure 10.8), as described in the table below (Table 10.2).

Figure 10.8 Select Options

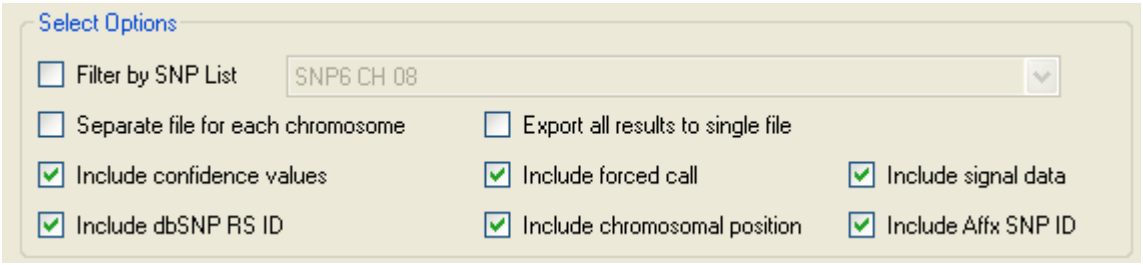


Table 10.2 Select Options for Tab Delimited Export

Select Options	Description
Filter by SNP List	Exports only the SNPs in a user-specified SNP list.
Separate file for each chromosome	Generates 26 text files (one for each chromosome, plus files containing SNPs on chromosome X, Y, or MT and a file containing SNPs that do not have chromosome information) instead of one text file for each CHP file. This option is not available if you select the "Export all results to single file" option.
Include confidence values	Choose this option to include the confidence value for each call in the exported results.
Include dbSNP RS ID	Choose this option to include the dbSNP RS ID that corresponds to the SNP probe set. The dbSNP at the National Center for Biotechnology Information (NCBI) attempts to maintain a unified and comprehensive view of known SNPs, small scale insertions/deletions, polymorphic repetitive elements, and microsatellites from the SNP consortium (TSC) and other sources. The dbSNP database is updated periodically, and the dbSNP version used for mapping is given in the dbSNP version field. For more information, please see *http://www.ncbi.nlm.nih.gov/SNP/*.
Export all results to single file	Generates a single text file. If this option is not chosen, one text file is generated for each CHP file. For more information, see:
Include forced call	Calls that do not meet the confidence score threshold specified by the configuration file are normally reported as "No Call". If the "Include forced call" option is selected, the genotype results include what the call would be if "No Calls" are not allowed.

Table 10.2 Select Options for Tab Delimited Export

Select Options	Description
Include chromosomal position	The chromosome and chromosomal position for the probe set
Include signal data	The software uses the signal data to generate the SNP cluster graphs. The specific signal data types vary depending upon the type of array and analysis used: For more information, see Chapter 9, *Using the SNP Cluster Graph* on page 148.
Include Affymetrix SNP ID	Include the Affymetrix unique identifier for the set of probes used to detect a particular SNP. See Note Below.

NOTE: If you select "Include Affymetrix SNP ID" for export when using NA29, NA30, NA31 annotation files the export column will be blank except for the column header because the annot.db files do not have that column.

6. Click **OK** in the Tab Delimited Export Options dialog box.

   The data is exported to one or more text files, depending upon the options selected.

   These options are described in more detail in:

   - *Export Each CHP file to a Separate Text File* on page 185
   - *Export All Data to One File* on page 186

NOTE: An export that generates "NoChromosome.txt" indicates an invalid SNP list (for example, retired SNPs that are no longer annotated.)

Export Each CHP file to a Separate Text File

If the "Export all results to single file" option is not selected, a separate text file or set of text files will be generated for each exported genotyping results file.

If you have not chosen to generate separate files for each chromosome, the text file name uses the following format: CHP file name.algorithm name.txt

If you have chosen to generate separate files for each chromosome, the text file name uses the following format: CHP file name.algorithm name.chromosome number.txt, where chromosome number can be:

- The number of the chromosome where the SNP was located
- X
- Y
- MT
- NoCh: SNPs with no chromosome location information.
- Contig ID – The number of the contig ID where the SNP was located (Axiom™ Genome-Wide BOS 1 array).

The header of the text file includes the following information: [show the different files differently]

- source CHP file location and name
- the execution GUID (a globally unique identifier for the genotyping batch run during which this CHP file was generated)
- SNP List (if chosen)
- Annotation versions
- Column headers for SNP data

The headers depend upon the array and algorithm type and options selected. For more information, see:

- Table 10.1, *Genotype Export Options for Tab Delimited Export,* on page 184
- Table 10.2, *Select Options for Tab Delimited Export,* on page 184

The SNP calls and information are displayed in rows below the file header.

If the confidence values, forced call, and/or signal data were selected for export, they will be included in the text file.

NOTE: Three dashes (---) represent a missing value. For Axiom™ results, two dashes (--) represent deletion in both alleles. One dash (-) represents deletion in one allele.

Figure 10.9 Exported genotype results, Genome-Wide SNP 6.0 array (tab-delimited .txt)

File Header

SNP Calls and Information

#CHP File=C:\Command_Console\Data\SNP 6 Data\20101108_112304_CHP\IC_201N.birdseed-v2.chp
#Exec GUID=0000490f-7bd6-41b1-52fe-003e2b000853
#GenomeWideSNP_6.na31.annot.db
#%genome-version-ucsc=hg19
#%genome-version-ncbi=GRCh37

Probe Set ID	Call Codes	Confidence	Forced Call Codes	Signal A	Signal B	dbSNP RS ID	Chromosome	Chromosomal Position	Affy ID
SNP_A-2131660	BB	0.004	BB	744.76	3254.096	rs2887286	1	1156131	
SNP_A-1967418	AB	0.005	AB	427.884	561.07	rs1496555	1	2234251	
SNP_A-1969580	BB	0.004	BB	954.487	3466.106	rs41477744	1	2329564	
SNP_A-4263484	AB	0.003	AB	1551.128	1296.376	rs3890745	1	2553624	
SNP_A-1978185	AA	0.002	AA	1268.171	332.748	rs10492936	1	2936870	
SNP_A-4264431	AB	0.003	AB	1173.486	874.303	rs10489588	1	2951834	
SNP_A-1980898	BB	0.01	BB	437.641	813.481	rs2376495	1	3095126	

In the figure above (Figure 10.9), the SNP data for all chromosomes has been exported into a single file and the exported SNP information includes chromosome number.

Figure 10.10 Exported genotype results, Axiom™ Genome-Wide Human Array (tab-delimited .txt)

File Header

SNP Calls and Information

#CHP File=C:\Command_Console\Data\Axiom 270 Hapmap\20101204_150358_CHP\NA11839_Axiom_GW_Hu_SNP_C.AxiomGT1.chp
#Exec GUID=0000396a-3bda-4d5d-3fcf-003f390019ae
#Axiom_GW_Hu_SNP.r2.na31.annot.db
#%genome-version-ucsc=hg19
#%genome-version-ncbi=GRCh37

Probe Set ID	Call Codes	Confidence	Forced Call Codes	Log Ratio	Strength	dbSNP RS ID	Chromosomal Position	Affy ID
AX-11700393	AA	0	AA	2.786	9.28	rs9681823	62309	
AX-11238231	BB	0	BB	-2.023	10.744	rs13072188	63411	
AX-11700430	BB	0	BB	-0.813	11.902	rs9683305	66866	
AX-11270917	AA	0	AA	2.579	10.273	rs1516320	76317	
AX-11274305	BB	0	BB	-3.373	10.087	rs1548188	77037	
AX-11270918	AA	0	AA	2.685	10.65	rs1516321	82010	

In the figure above (Figure 10.10), the data for each chromosome has been exported into a separate file, and the SNP information does not include Chromosome number.

Export All Data to One File

If the “Export all results to single file” option is selected, the data for all CHP files will be exported to a single .TXT file (Figure 10.11). The file name is the one entered in Export File Name box.

Figure 10.11 Export data to a single file option selected

File Header

SNP Calls for all Selected CHP files

#Axiom_GW_Hu_SNP.r2.na31.annot.db			
#%genome-version-ucsc=hg19			
#%genome-version-ncbi=GRCh37			
Probe Set ID	NA06985_Axiom_GW_Hu_SNP_C.AxiomGT1.chp Call Codes	NA06985_Axiom_GW_Hu_SNP_C.AxiomGT1.chp Confidence	NA06985_Axi
AX-11086525	AA	0	AA
AX-11086526	BB	0	BB
AX-11086527	BB	0	BB
AX-11086528	BB	0	BB
AX-11086529	AA	0	AA
AX-11086530	BB	0	BB
AX-11086531	AA	0	AA
AX-11086532	BB	0	BB
AX-11086534	AA	0	AA

The header of the text file includes the following information:

- Annotation versions, if available
- SNP List (if chosen)
- Column headers for SNP data

  The headers depend upon the array and algorithm type and options selected. For more information, see:

  - Table 10.1, *Genotype Export Options for Tab Delimited Export,* on page 184
  - Table 10.2, *Select Options for Tab Delimited Export,* on page 184

If the following options are selected, a column will be created for the data for each CHP file exported:

- Call Code
- Only export forward strand call codes
- Include confidence values
- Include forced calls
- Include signal data

If the following options are selected, a single column will be created in the .TXT file:

- dbSNP RS ID
- Include Chromosomal Position (displays chromosome number and position in separate columns)
- Include Affx ID

Export the Combined Results of an Array Set

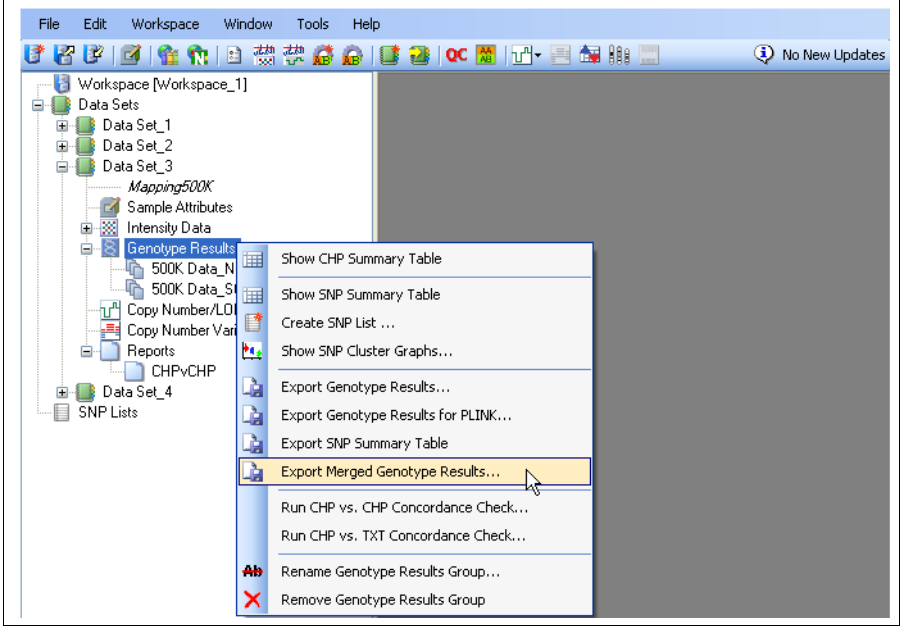
The genotype results from the arrays of an array set (for example, Human Mapping 250K Nsp and 250K Sty results) can be combined and exported to one text file.

The Sample (ARR) files for the for each paired array in the array set must have an attribute in common that can be used to match the files for merging.

NOTE: Sample files (ARR) are required for the genotype results that you want to combine and export.

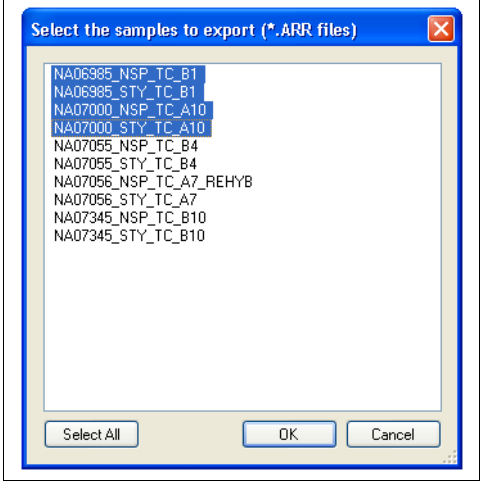
1. Right-click a Genotype Results group and select **Export Merged Genotype Results** on the shortcut menu (Figure 10.12).

Figure 10.12 Select genotype results to merge for export



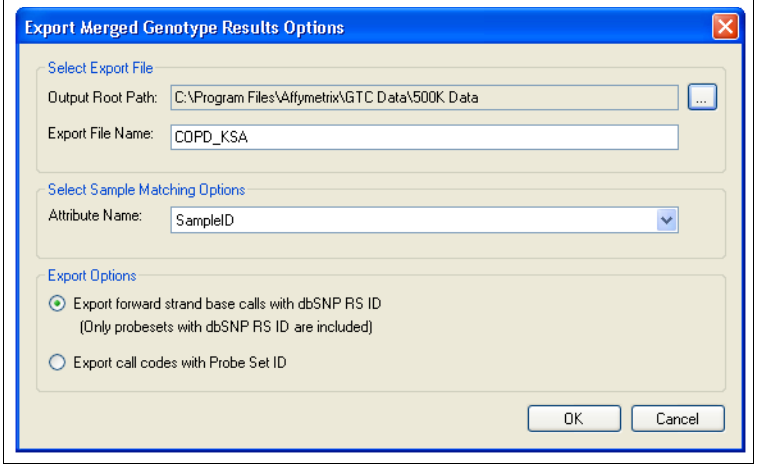
The Select the samples to export (*.ARR files) dialog box opens (Figure 10.13).

Figure 10.13 Select samples for merged export



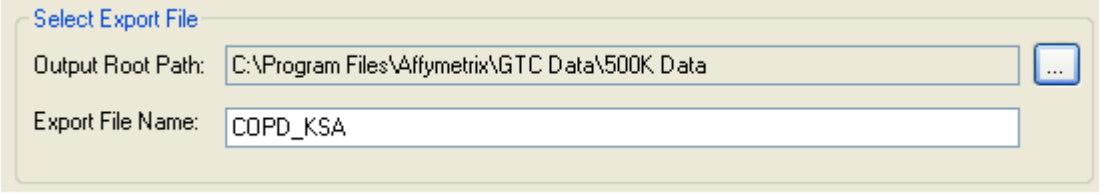
2. Select the samples to export and click **OK**.
   The Export Merged Genotype Results Options dialog box opens (Figure 10.14).

Figure 10.14 Export Merged Genotype Results Options dialog box



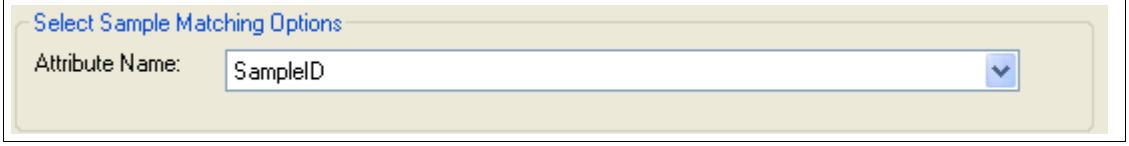
3. Select a destination directory and enter a name for the results file (Figure 10.15).

Figure 10.15 Select Export File options



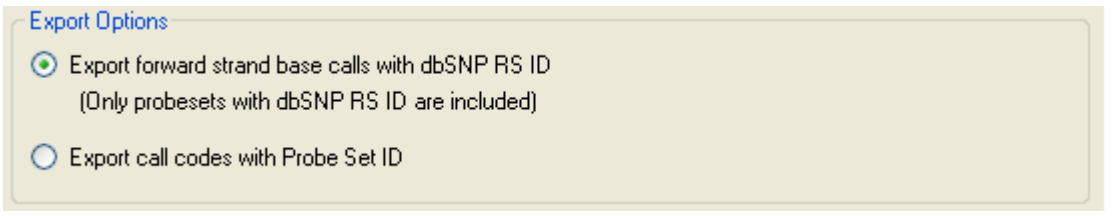
4. Select a sample matching option using one user attribute from ARR files for these samples (Figure 10.16).

Figure 10.16 Select Export File options



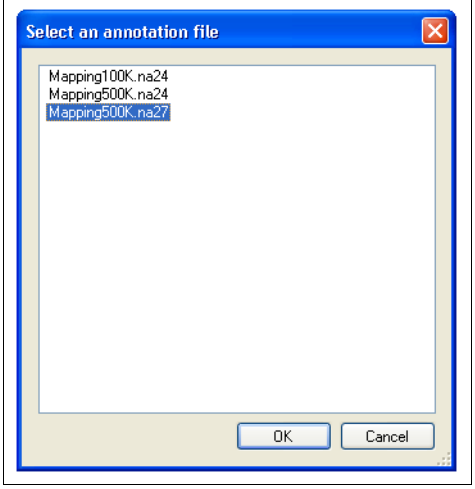
5. Select an export option (Figure 10.17):
   - Export forward strand base calls with dbSNP RS ID – Choose this option to include the forward strand base calls (AT, CG, AG, TC, --, etc.) in the text file. Only probe sets with dbSNP RS ID are included.
   - Export call codes with Probe Set ID - Choose this option to include the Affymetrix call codes (AA, AB, or BB) in the text file.

Figure 10.17 Export Options



6. Click **OK** in the Export Merged Genotype Results Options dialog box.
   The Select an annotation file dialog box opens (Figure 10.18).

Figure 10.18 Select an annotation file



7. Select an annotation file and click **OK**.
   The merged genotyping calls are exported to a .TXT file.
   The figure below (Figure 10.19) shows an example of merged results.

Figure 10.19 Example merged results with one sample per row

	A	B	C	D	E	F
1	#AnnotationFile=Mapping500K.na27.annot.db					
2	#ArraySet=Mapping500K					
3	#ArrayTypes=Mapping250K_Nsp,Mapping250K_Sty					
4	#BatchNames=500K Data_Nsp,500K Data_Sty					
5	#ProbeSetType=db SNP Id					
6	#CallType=Real Allele Call					
7	#SnpList=					
8	SampleID	Sample files	rs16994928 Call	rs16994928 Confidence	rs233978 Call	rs233978 Conf
9	ref_1	NA06985_NSP_TC_B1.brlmm.chp, NA06985_STY_TC_B1.brlmm.chp	GG	0.012	GG	0.01
10	ref_2	NA07000_NSP_TC_A10.brlmm.chp, NA07000_STY_TC_A10.brlmm.chp	GG	0.024	AA	0.093

COPD_KSA

Export Genotype Results for PLINK

Genotype results can be exported to a file format that is compatible with PLINK software. To export files for PLINK, the genotype CHP result files must have matched sample attribute files (ARR) created with the Pedigree template (available in the Affymetrix® AGCC software) and the corresponding information for each sample. If the ARR files were created without this template or are missing data for some of the samples, update the ARR files using the Pedigree template before you attempt to export the data using this option.

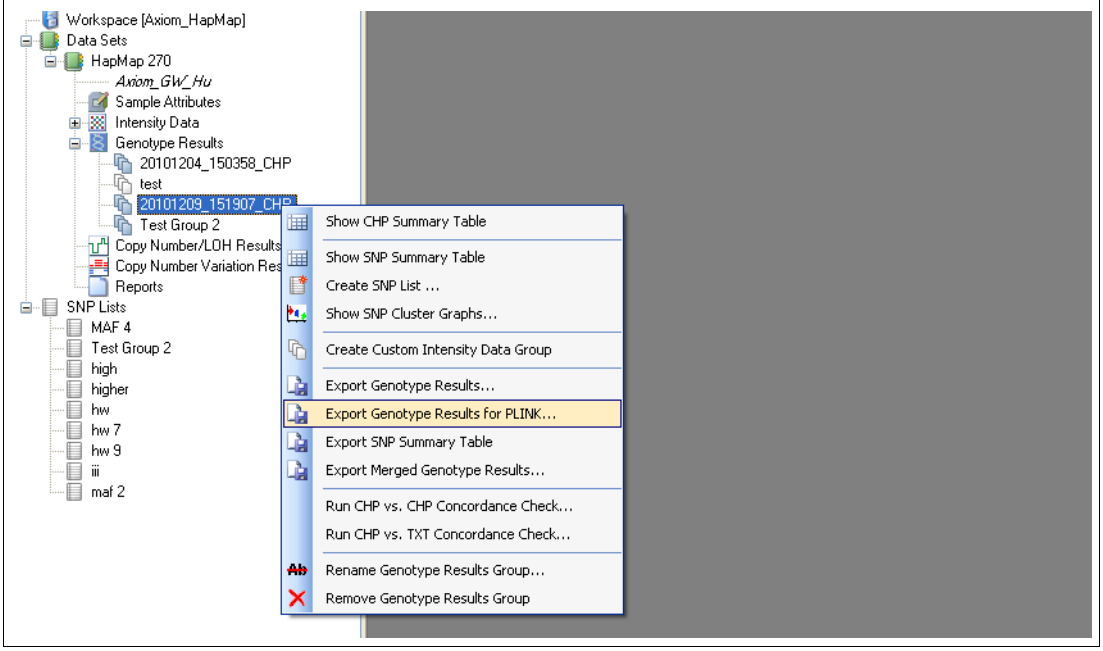
NOTE: PLINK export is not available for non-human arrays in GTC 4.2. For more information on exporting non-human arrays in PLINK compatible format, see ***Exporting Non-Human Genotype Results for PLINK* on page 196**.

Exporting Human Genotype Results in PLINK Format

1. Do one of the following:

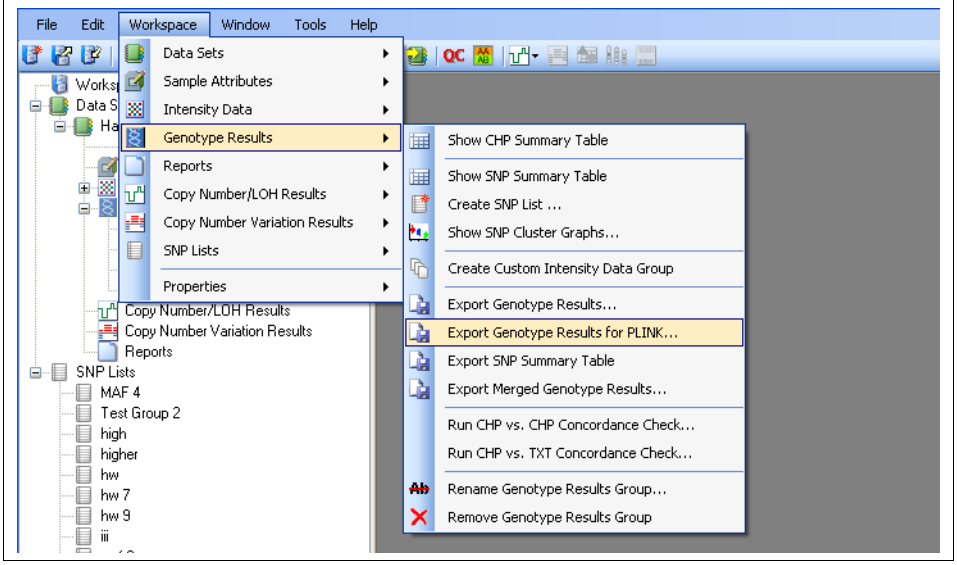
   Right-click a Genotype Results group and select **Export Genotype Results for PLINK on the shortcut menu** (Figure 10.20).

Figure 10.20 Select Genotype Results group in data tree for PLINK export



Select **Workspace > Genotype Results > Export Genotype Results** for PLINK on the menu bar (Figure 10.21).

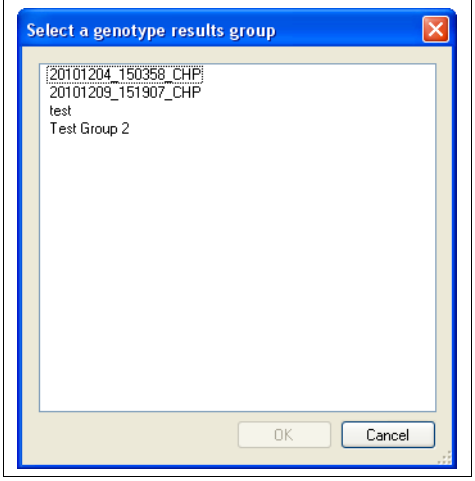
Figure 10.21 Export Genotype Results for PLINK from Workspace menu



Select results (rows) in the CHP Summary table. Right-click the selection, and choose **Export Genotype Results for PLINK** on the shortcut menu (Figure 10.21).

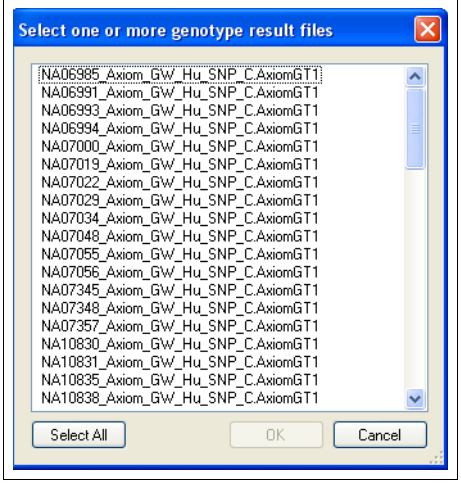
If you have not picked a specific results group, the Select a genotype results group dialog box opens (Figure 10.22).

Figure 10.22 Select a genotype results group dialog box



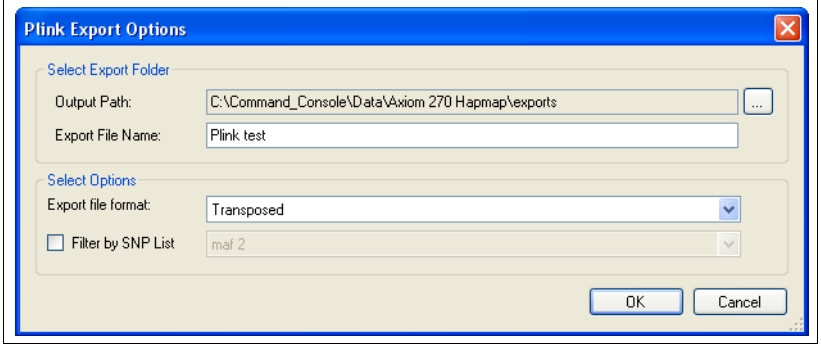
2. Select a group and click the **OK** button in the Select a genotype results group dialog box. The Select one or more genotype results files dialog box appears (Figure 10.23).

Figure 10.23 Select one or more genotype results file dialog box



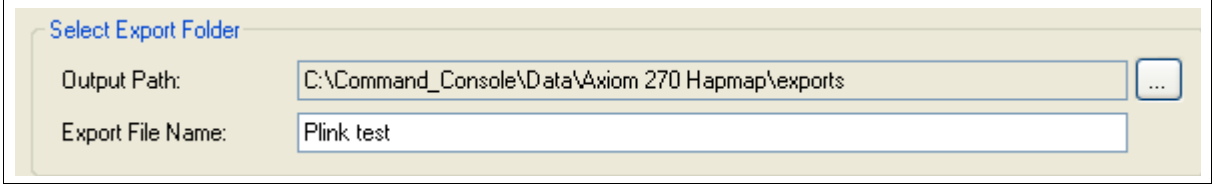
3. Select the results to export and click **OK** in the Select one or more Genotype results files dialog box. The Plink Export Options dialog box opens (Figure 10.24).

Figure 10.24 PLINK export options dialog box



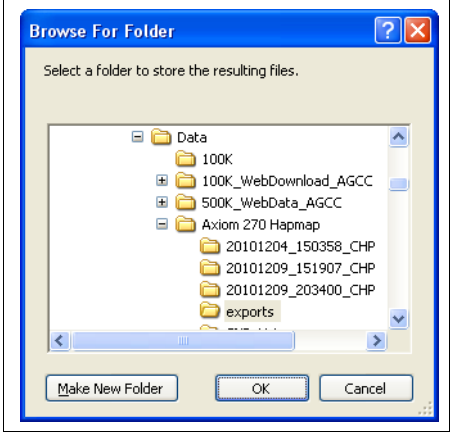
4. Click the **Browse** button ... to select the output directory (Figure 10.25).

Figure 10.25 Select Export Folder options



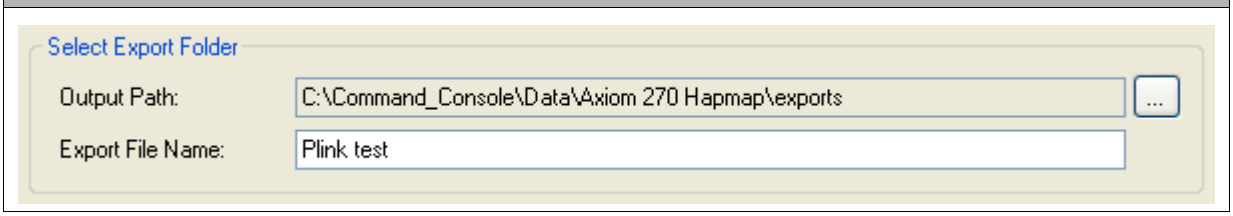
The Browse for Folder dialog box opens (Figure 10.26).

Figure 10.26 Browse for Folder dialog box



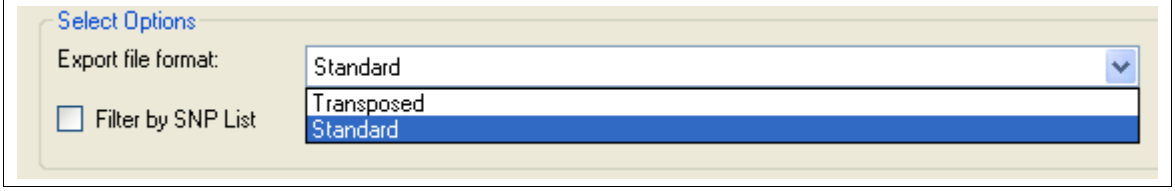
5. Navigate to the folder and click OK.
6. Enter a name for the Export file (Figure 10.27).

Figure 10.27 Select Export Folder options



7. Select an export file format (Figure 10.28).

Figure 10.28 Selecting export file format



You can select from the following options:

- Transposed – Generates three files: .tped, .tfam, and .map (Table 10.3)

Table 10.3 Example PLINK transposed format

SNP	Patient 1	Patient 2	Patient 3
SNP 1	Call	Call	Call
SNP 2	Call	Call	Call
SNP 3	Call	Call	Call
SNP 4	Call	Call	Call

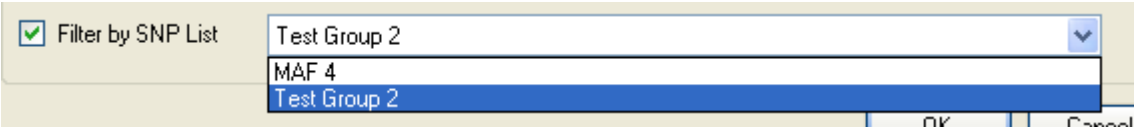
- Standard – Generates two files: .map and .ped (Table 10.4)

Table 10.4 Example PLINK standard format

Patient	SNP 1	SNP 2	SNP 3
Patient 1	Call	Call	Call
Patient 2	Call	Call	Call
Patient 3	Call	Call	Call
Patient 4	Call	Call	Call

8. Select the Filter by SNP List option (Figure 10.29)
   Choose this option to export only the SNPs specified in a user-selected SNP list.

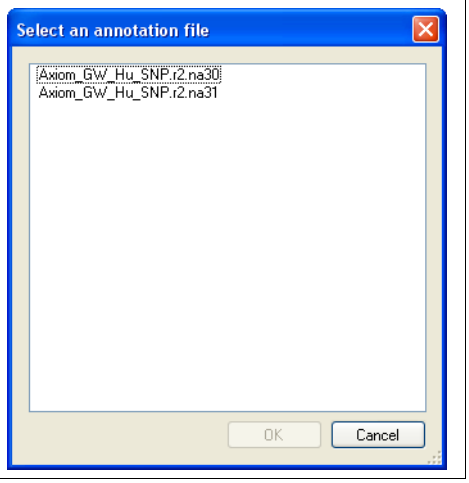
Figure 10.29 Selecting Filter by SNP List option



Click the checkbox and select a SNP list from the dropdown list.

9. Click **OK** in the Plink Export Options dialog box.
   If you do not have an annotation file selected for the array type, the Select an Annotation file dialog box opens (Figure 10.30).

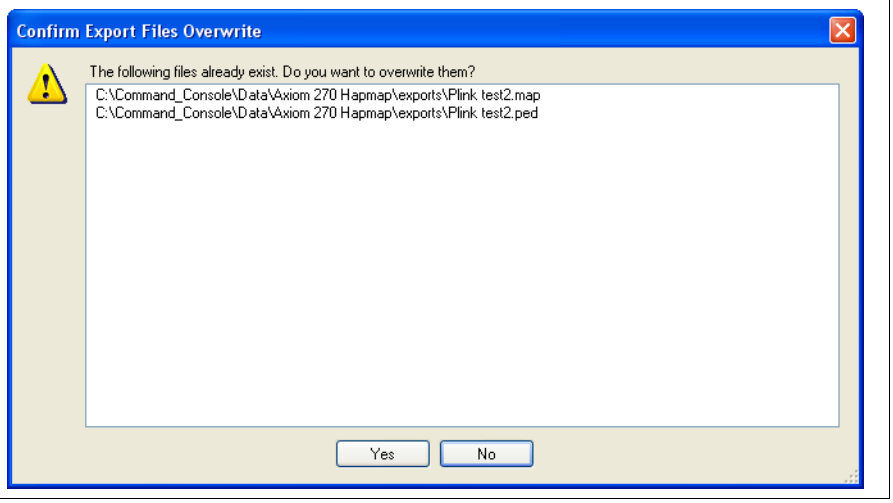
Figure 10.30 Select an annotation file dialog box



10. Select an annotation file and click **OK** in the Select an annotation file dialog box.

If you are overwriting previously exported files, the Confirm Export Files Overwrite dialog box opens (Figure 10.31).

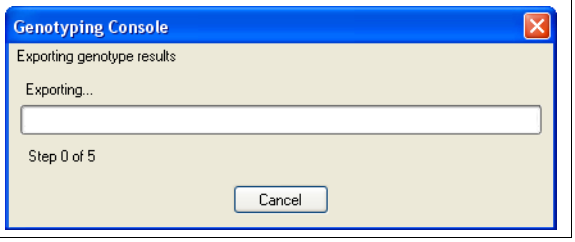
Figure 10.31 Confirm Export Files Overwrite dialog box



- Click **Yes** to overwrite the files
- Click **No** to return to the Plink export options dialog box (Figure 10.24).

The Exporting genotype results progress bar appears (Figure 10.32)

Figure 10.32 Exporting genotype results progress bar

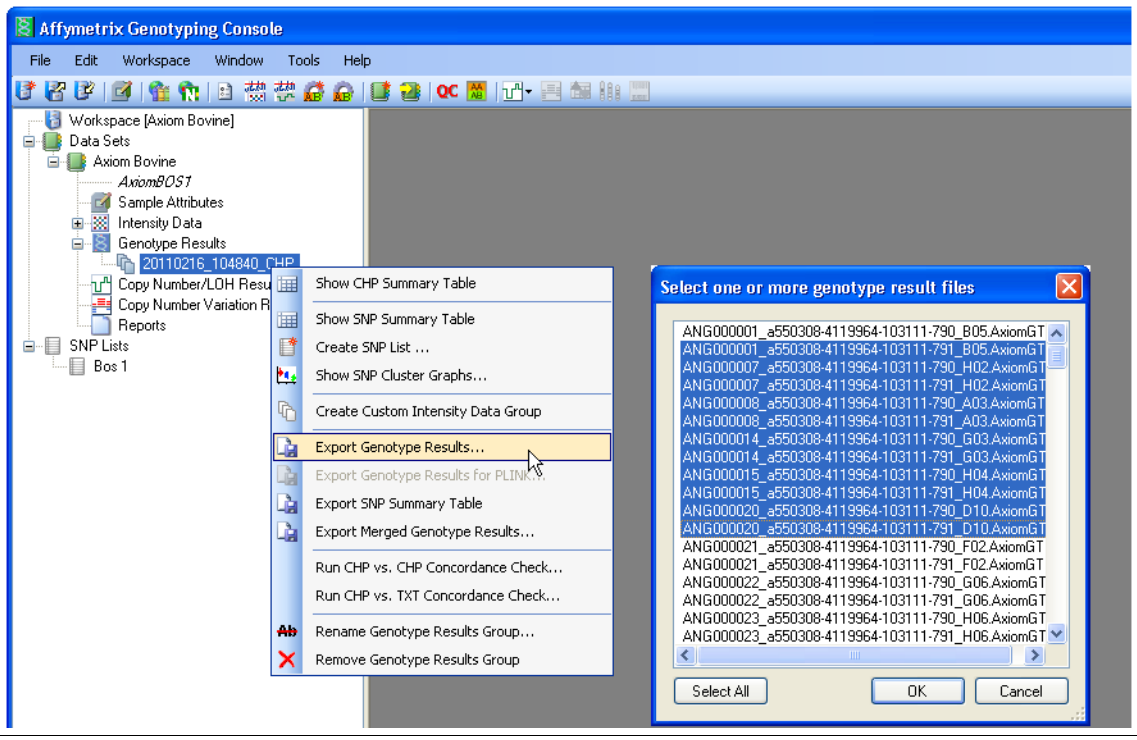


The exported files are placed in the location you chose.

Exporting Non-Human Genotype Results for PLINK

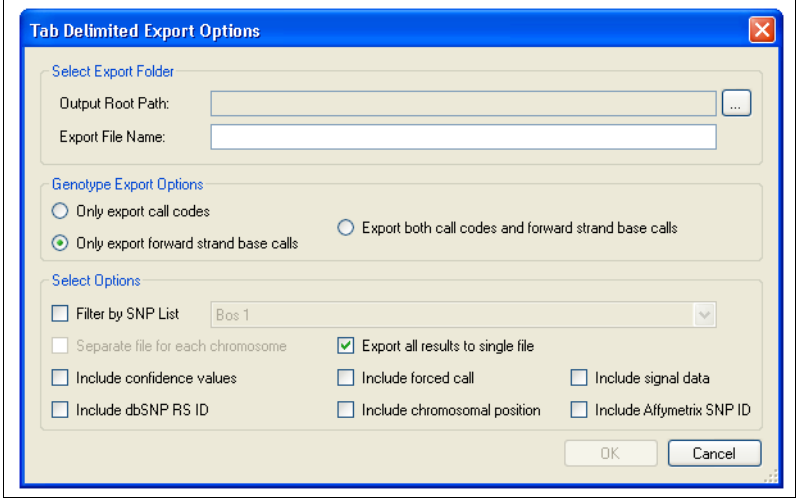
1. Right-click the genotype results and select **Export Genotype Results** on the shortcut menu (Figure 10.33).

Figure 10.33 Select genotype results for export



2. In the dialog box that appears, select the genotype result to export and click **OK**.
3. In the Tab Delimited Export Options dialog box that appears (Figure 10.34), set the output root path and enter the export file name. Choose the following export options:
   - "Only export forward strand base calls"
   - "Export all results to single file"

Figure 10.34 Export options



4. Click **OK**.
5. Modify the exported text file to a PLINK compatible format.

Table & Graph Features

In Genotyping Console, there are several properties which are common to all tables and graphs. The following sections describe:

NOTE: The use of the GTC Copy Number Browser to view Copy Number/LOH data is described in the *GTC Browser Manual.*

Table Features

In Genotyping Console, the tables used to display data share several common features:

All common table functions are accessible through the shortcuts on the table tool bar (Figure 11.1, Table 11.1).

Figure 11.1 Table tool bar

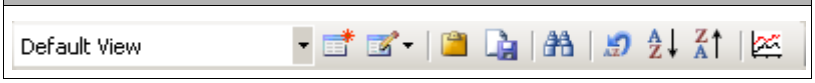
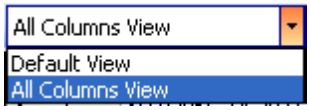


Table 11.1 Table Tool bar functions

Table Function	Tool bar
Table Views	
New Views	
Edit Views	
Copy to Clipboard	
Save Table to File	
Find	
Reset Sort Order	
Sort Ascending	
Sort Descending	
Show Line Graph	

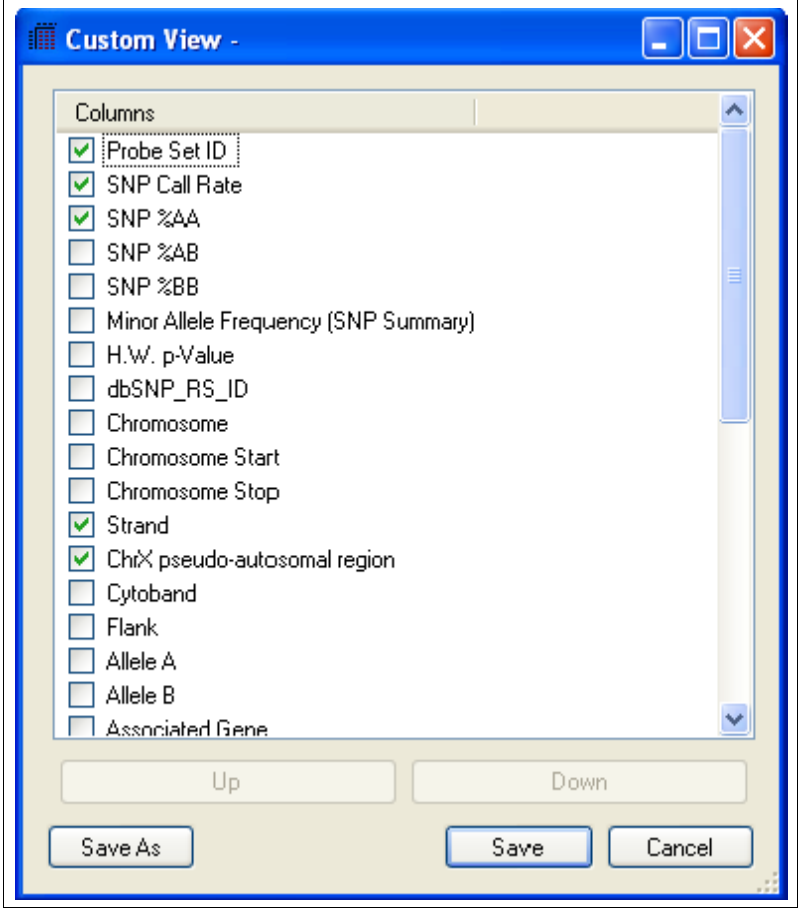
Custom View Features

Each table in Genotyping Console has a default set of displayed columns. The features described below enable you to change these columns.

To create custom views:

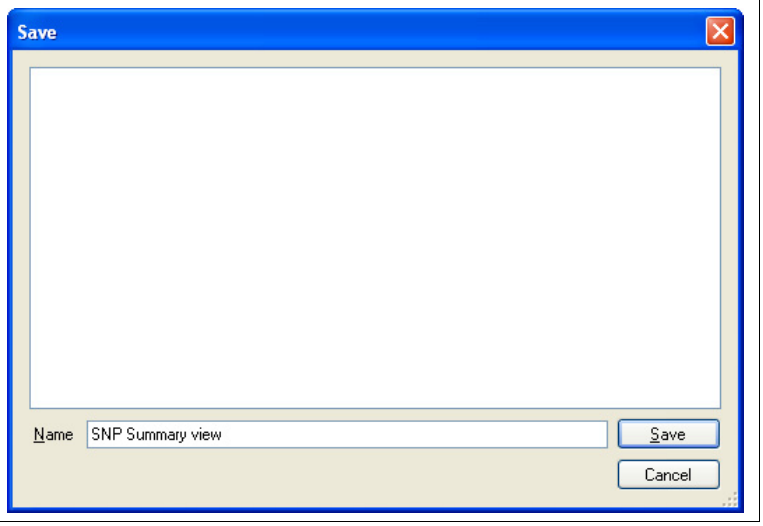
1. Select the **New View** shortcut .
   The Custom View dialog box opens (Figure 11.2).

Figure 11.2 Custom View dialog box



2. Select the columns to be displayed. To re-order the columns in the table, click the column name and use the **Up** and **Down** buttons.
3. Click **Save** and enter a name for this view.
   The Save dialog box opens (Figure 11.3).

Figure 11.3 Save dialog box

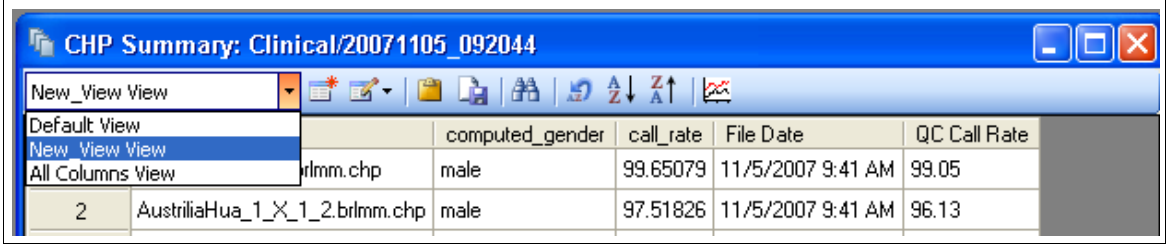


4. Enter a name for the view and click **Save** in the Save dialog box.
   Use the drop down menu to display this custom view.

To select a previously generated view:

- Select the view from the drop-down menu (Figure 11.4).

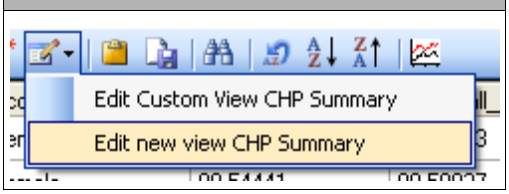
Figure 11.4 Selecting view



To edit a previously generated custom view:

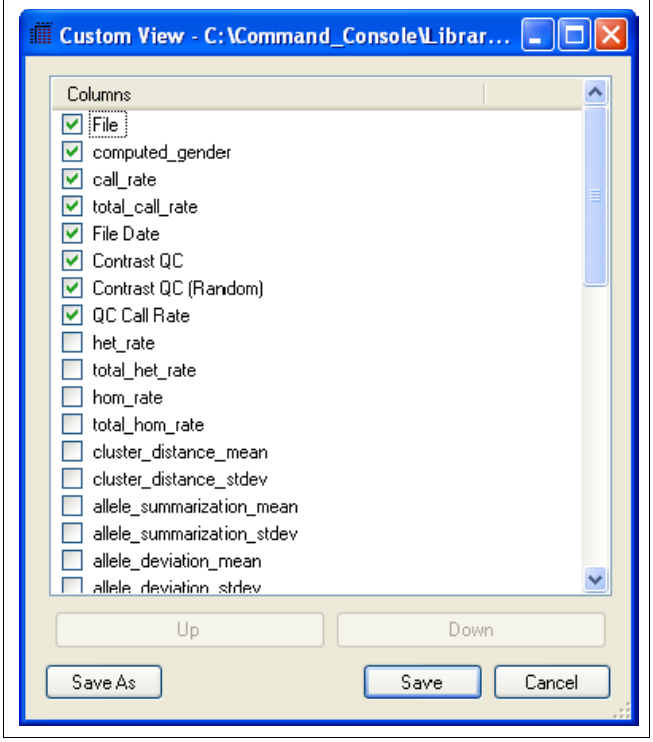
1. Click on the **Edit View** shortcut
   The dropdown menu displays a list of user-generated views (Figure 11.5).

Figure 11.5 Selecting view for editing



2. Select the View to edit.
   The Custom View dialog box opens (Figure 11.6).

Figure 11.6 Custom View dialog box



3. Make the desired changes and save the view
   Click **Save As** to save the changes with a new view name.

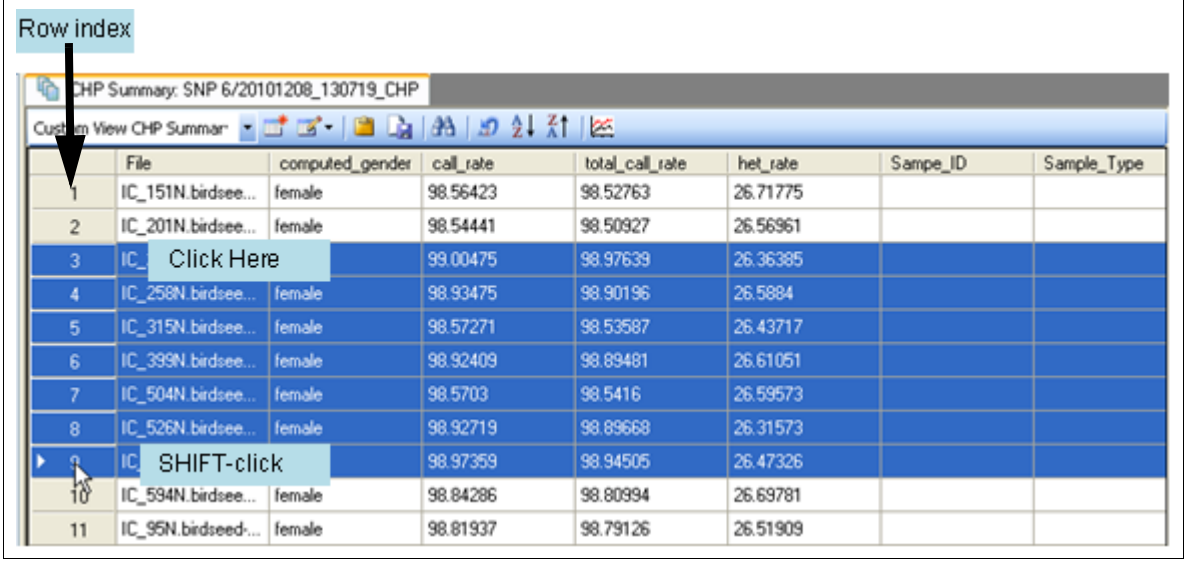
Other Table Features

You can select one or many cells, rows, or columns.

To quickly select a range of rows:

- Click the first row index, and then SHIFT-click the last row index (Figure 11.7).

Figure 11.7 Selecting a range of rows



To select multiple rows that are not adjacent:

- CTRL-click on each row.

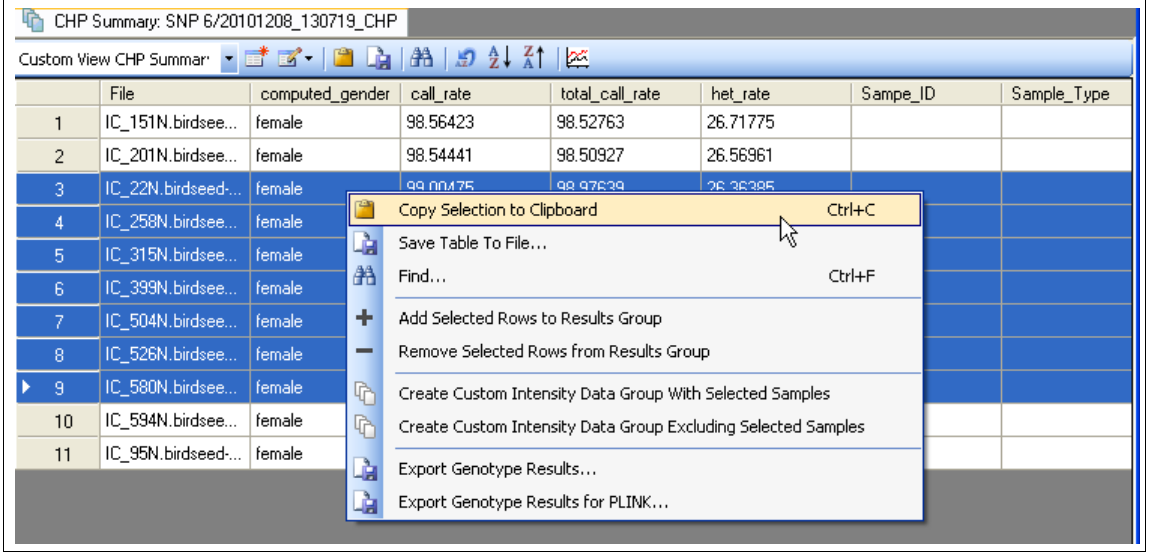
  These options are available for columns and cells as well.

To copy a selection to the Clipboard:

1. Select the desired cells, rows, columns.
2. Click on the **Copy to Clipboard** shortcut on the tool bar, or

   Right-click on the selected items and select the Copy Selection to Clipboard from the right-click menu (Figure 11.8).

Figure 11.8 Right-click menu



NOTE: Copy to Clipboard may fail if too much data is copied (for example, copying the entire SNP Summary Table). Affymetrix recommends that you Save Table To File if you wish to transfer table information to another application.

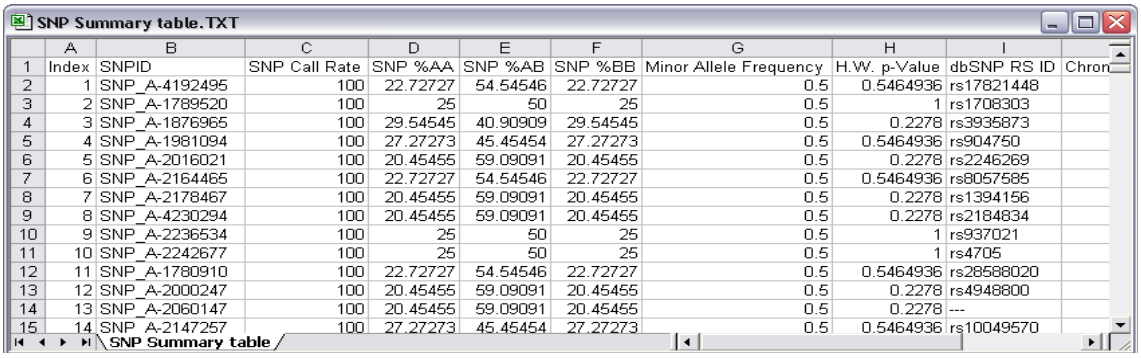
To save all of the data in the open table to a text file:

1. Select the **Save Table to File** shortcut from the tool bar, or

   Right-click and select the same command from the right-click menu.
2. Enter a name for the file and select **Save**.

   All displayed data will be written to the text file (Figure 11.9).

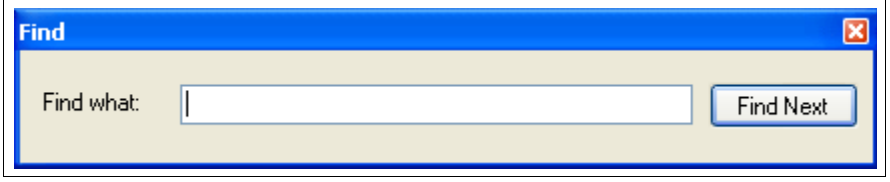
Figure 11.9 Text file



To find data in the table:

- Select the **Find** shortcut and enter the value to search on in the Find dialog box (Figure 11.10).

Figure 11.10 Find value dialog box



The Find Next button will continue to search the table for additional instances of the search criteria. When the end of the document is reached, it will restart the search from the top of the table.

NOTE: The Find function does not utilize wildcards.

To return the table to the default sort order:

- Select the **Reset Sort Order** shortcut.

To sort the table:

- Select a column header and select the **Sort Ascending** or **Sort Descending** shortcuts.

In the Intensity QC and CHP Summary tables, a line graph can be displayed.

To invoke the line graph

- Click on the **Show Line Graph** shortcut.

  See *Graph Features* on page 203 for more information.

NOTE: Features and functions specific to a particular table type are described in the section of the manual dealing with that table and data.

Graph Features

This section describes the line graph features that can be used to display different metrics in the tables

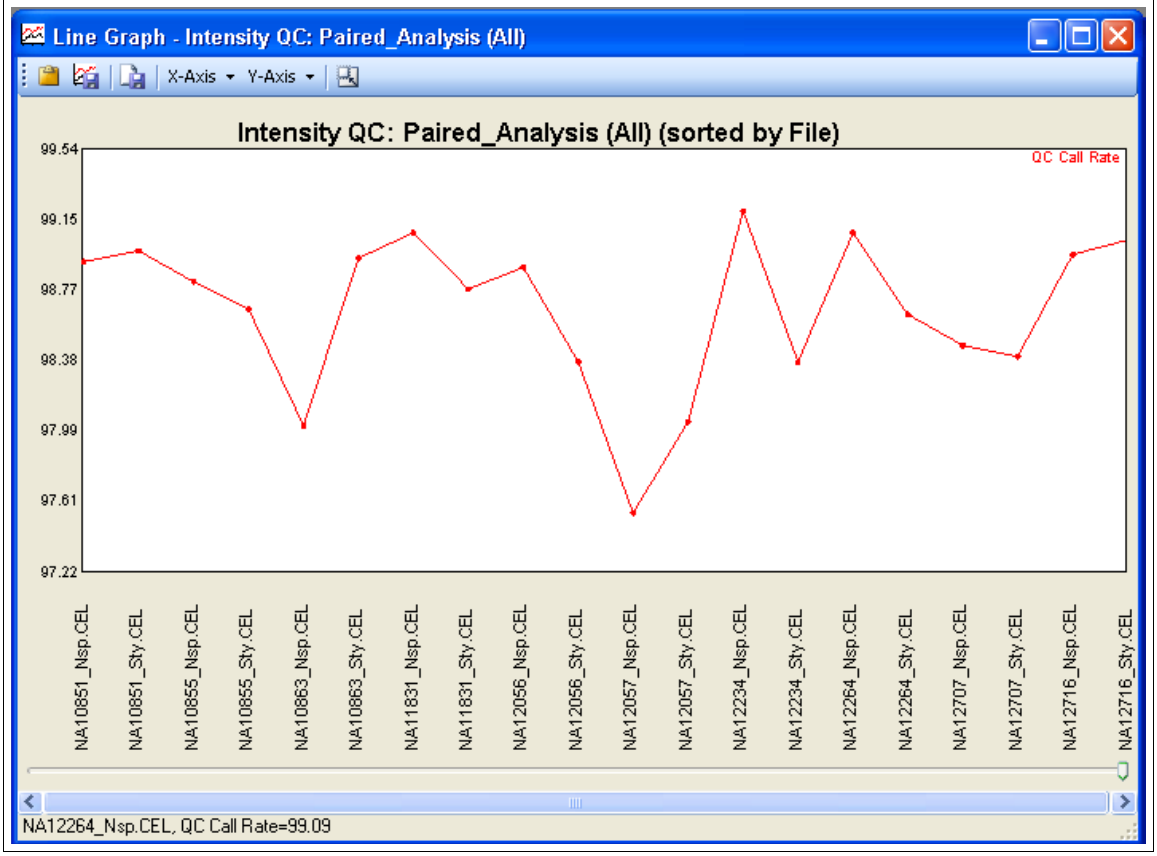
The Line Graph is not available for every table.

Line graphs can be generated for the different results.

To invoke the line graph:

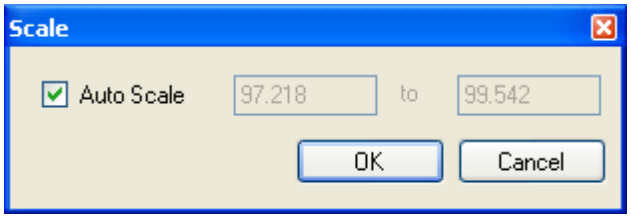
1. Click on the **Line Graph** shortcut from the table shortcut bar.

Figure 11.11 Line Graph



2. To sort the X-axis by another category (e.g. Bounds), select the category from the X-axis drop-down menu or right-click on the graph and select **Set X-axis Category**.
3. To graph additional results, right-click on the graph and select **Set Y-axis Categories** or use the Y-axis drop-down menu.
4. To set the axis scale, right-click on the graph and select **Set Axis Scale** or select the **Set Scale** shortcut  from the tool bar.

Figure 11.12 Scale dialog box



The line graph data can be copied to the Clipboard  , saved as an image file  (*.png format), or saved as a text file  (tab-delimited *.txt format).

Copy Number & LOH Analysis for Human Mapping 100K/500K Arrays

GTC can be used to perform the following analyses for Human Mapping 100K/500K arrays:

- Copy Number (CN)
- Loss of Heterozygosity (LOH)
- Copy Number Segment Reporting
- Custom Region Copy Number Segment Reporting

For information on performing CN and LOH analysis on SNP 6 data, see Chapter 13, *Copy Number & LOH Analysis for Genome-Wide Human SNP 6.0 Arrays* on page 242.

Features common to Human Mapping 100K/500K arrays and Genome-Wide Human SNP Arrays 6.0 arrays, including running the Segment Reporting Tool, are described in Chapter 14, *Common Functions for Copy Number/LOH Analyses* on page 285.

IMPORTANT: CN and LOH analyses for Human Mapping 100K and 500K platforms in GTC 4.2 are algorithmically the same as in CNAT4.0.1 software, and for more details users should refer to the Affymetrix White Paper "Copy Number and Loss of Heterozygosity Estimation Algorithms for the GeneChip Human Mapping Array Sets." Files that result from this analysis have the extension .CN4.

NOTE: GTC does not perform copy number, LOH, or Copy number region analysis on data from Genome-Wide Human SNP 5.0 or Axiom™ Genome-Wide Human arrays.

IMPORTANT: Affymetrix recommends that you perform Copy Number/LOH analysis with all files stored locally.

The basic workflow for Copy Number/LOH analysis involves:

1. Performing Copy Number/LOH analysis on a selection of CEL or CHP files.
   There are two options for this:
   - *Paired Copy Number and LOH Analysis* on page 207
   - *Unpaired Copy Number and LOH Analysis* on page 214
2. *Performing the Copy Number Segment analysis on the CN data files* (page 285).

   NOTE: Segment Reporting Analysis can be performed on Human Mapping 100K/500K data and on Genome-Wide Human SNP Array 6.0 data.
3. *Viewing QC data in table format* (page 231)
4. *Viewing the data in the GTC Browser* (page 305)
5. *Exporting data into formats that can be used by secondary analysis software* (page 307)
6. You can also:
   - *Change the QC threshold settings* (page 312)
   - *Change the algorithm parameters for 100K/500K analysis* (page 232)

Introduction to 100K/500K Analysis

This section provides a brief description of:

100K/500K Array Configuration

Human Mapping 100K/500K analyses use two arrays to provide full coverage of the genome. Analyses can also be performed using only the data from a single 50K or 250K array.

- Human Mapping 100K is a combination of data from the following arrays:
  - Mapping50K_Xba240
  - Mapping50K_Hind240
- Human Mapping 500K is a combination of data from the following arrays:
  - Mapping250K_Nsp
  - Mapping250K_Sty

The Segment Report Tool is run after Copy Number analysis.

If you wish to run CN number and/or LOH analysis on both array types at the same time, you need to have Enzyme Set attributes set up for the files. You can use Enzyme Set and Sample + Reference attributes to make sorting and pairing up the files easier. For more information on these steps see *Using Shared Attributes to Group Samples* on page 226.

CN and LOH Algorithms

CN4 performs paired and unpaired CN analysis:

- Paired CN Analysis

  Paired CN Analysis is used to compare two samples from the same individual to look for copy number differences in different types of tissues (examples of the two samples would be Tumor/Normal or Treated/Untreated samples from the same individual).

  Paired analysis requires that genotyping batch analysis be performed on the data that will be used for CN analysis.
- Unpaired CN Analysis

  Unpaired CN Analysis is used to compare sample files to a set of reference files.

  Unpaired analysis requires that genotyping batch analysis be performed on the data that will be used for CN analysis.

Copy number data is output in files with the suffix .CN4.cnchp.

LOH analysis can be run at the same time as copy number analysis or in a separate step without running the copy number analysis.

Human Mapping 100K/500K copy number and LOH data is output in separate files (CN4.cnchp files and CN4.lohchp files).

Copy number segment reports can be run on Human Mapping 100K/500K array CN data, but no gender calls are made by the Segment Reporting Tool.

Copy Number/LOH Analysis for Human Mapping 100K/500K Arrays

IMPORTANT: Affymetrix recommends that you perform Copy Number/LOH analysis with all files stored locally.

This section describes the different Copy Number/LOH workflows for Human Mapping 100K/500K arrays.

Paired Copy Number and LOH Analysis

Paired CN Analysis is used to compare two samples from the same individual to look for copy number differences in different types of tissues (Normal/Tumor, for example).

Genotyping batch analysis must be performed on the data used for CN analysis prior to the CN analysis.

Enzyme Set attributes must be assigned to the arrays to match array sets originating with the same sample. For example, you could use the "Subject ID" attribute as the Enzyme Set identifier.

Sample/Reference attributes can be useful in group arrays into either the Sample or Reference category. For example, you could use "Disease State" or "Tissue State" attributes to distinguish between reference and sample arrays for paired analysis.

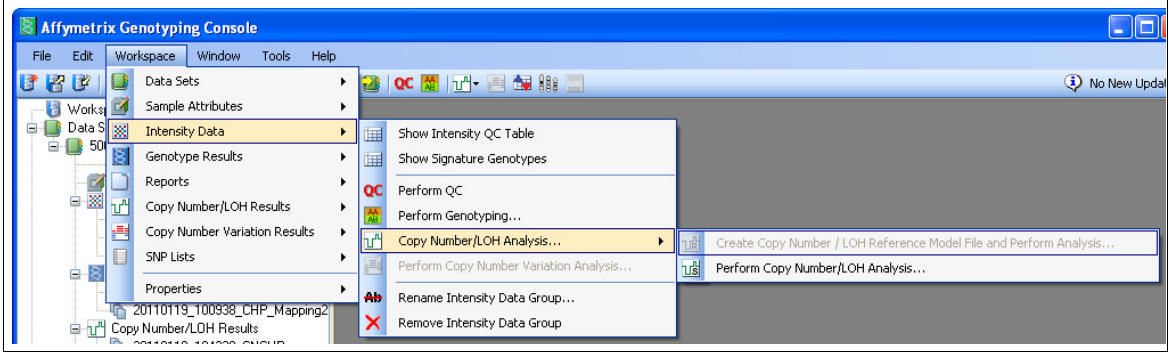
The Copy Number and LOH files resulting from combined Enzyme Set data will be named using the Enzyme Set attribute for the array set. Output files can be given a suffix.

For more information about using shared attributes to pair files by enzyme set or sample/reference group, see *Using Shared Attributes to Group Samples* on page 226.

To perform a Paired copy number and/or LOH analysis:

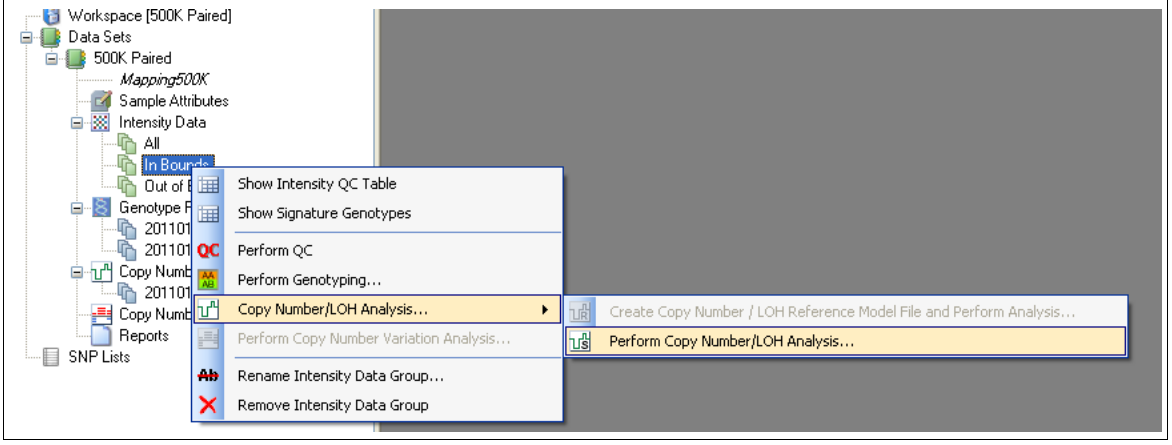
1. Open the Workspace and select the Data Set with the data for analysis.
2. Select the Intensity Data file set.
3. Do one of the following:
   - From the Workspace menu, select **Intensity Data > Copy Number/LOH Analysis > Perform Copy Number/LOH Analysis…**. (Figure 12.1).

Figure 12.1 Selecting CN/LOH analysis from Workspace menu



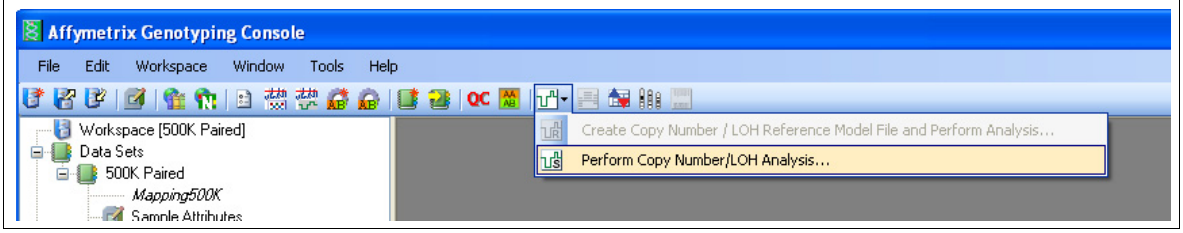
   - Right-click the Intensity Data file set and select **Perform Copy Number/LOH Analysis** from the pop-up menu (Figure 12.2).

Figure 12.2 Selecting CN/LOH analysis from the Data tree



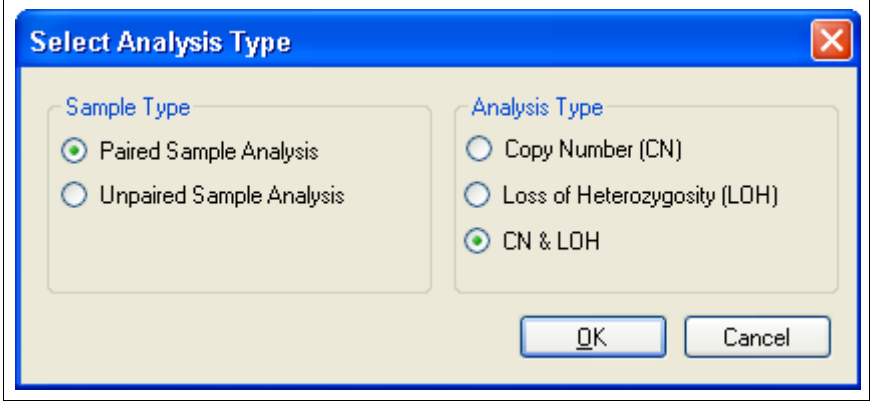
- Click the **Perform Copy Number Analysis** button in the tool bar and select **Perform Copy Number/LOH Analysis** from the dropdown list (Figure 12.3).

Figure 12.3 Selecting CN/LOH analysis from the Tool bar



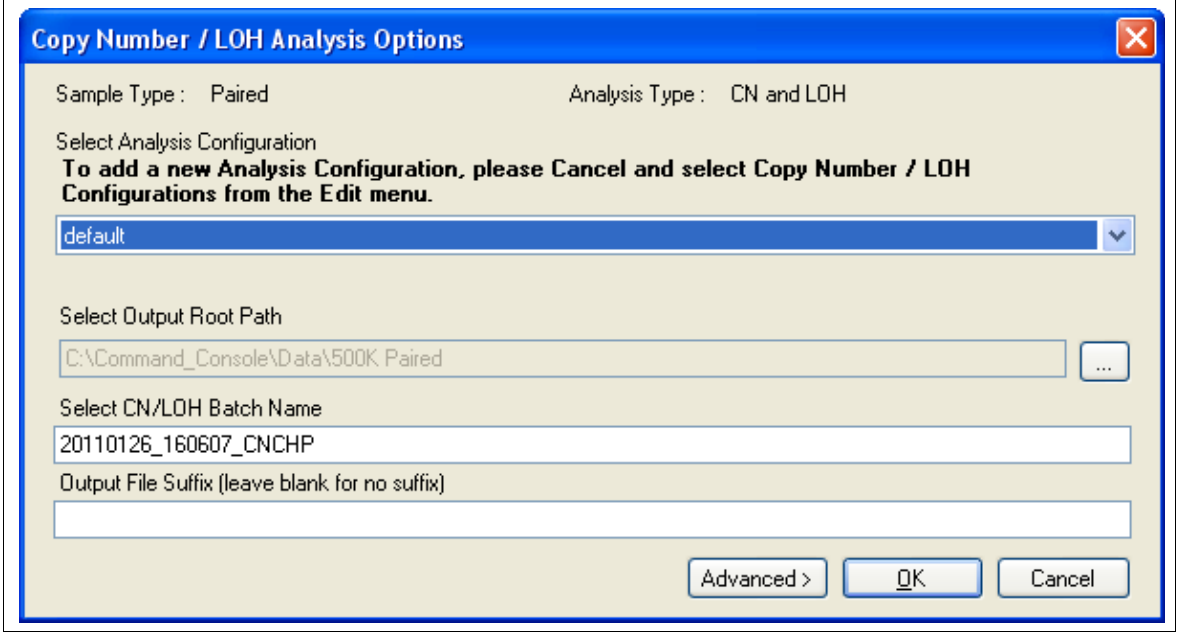
The Select Analysis Type dialog box opens (Figure 12.4).

Figure 12.4 Select Analysis Type dialog box



4. Select **Paired Sample Analysis** for Sample type.
5. Select the analysis type (CN, LOH, or both).
6. Click **OK**.
   The Copy Number Analysis Options dialog box opens (Figure 12.5).

Figure 12.5 Copy Number Analysis Options dialog box



7. Review analysis configuration parameters and select new analysis configuration if desired.

   See *Changing Algorithm Parameters for Human Mapping 100K/500K Analysis* on page 232 for more information on creating a new analysis configuration.

   Change the following if desired:

   - Output Root Path: location of the CN/LOH Results Group folder.
   - Select CN/LOH Batch Name: Name of the CN/LOH Results Group and its folder.

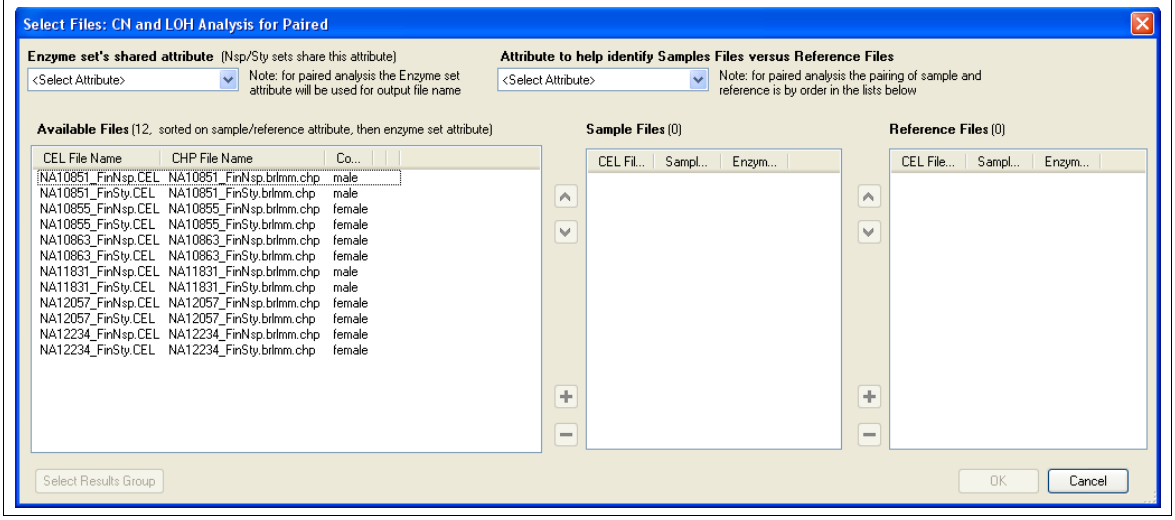
     NOTE: This folder is the location where the different Data Results files are kept. You can access the folder through Windows Explore to view report files.

   - Output File Suffix: suffix added to distinguish output file names.

8. Click **OK**.

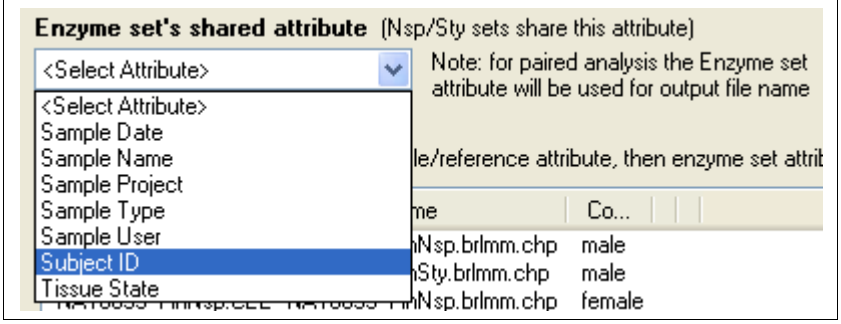
   The Select Files dialog box opens (Figure 12.6).

Figure 12.6 Select Files dialog box where files users can select files for paired sample CN and LOH analysis



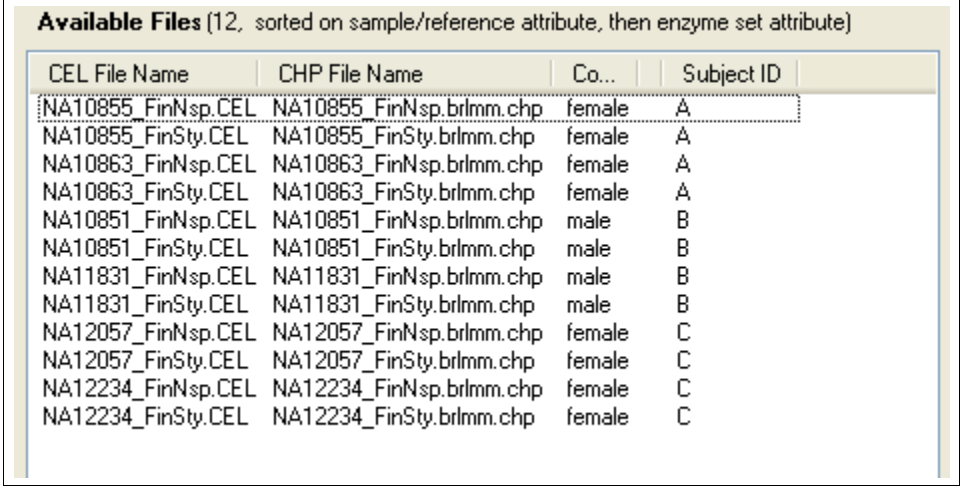
9. Select the Enzyme Set shared attribute from the Enzyme Set Shared Attribute drop-down list (Figure 12.7).

Figure 12.7 List of Enzyme Set shared attributes that can be used to find matching array sets which originate with the same sample



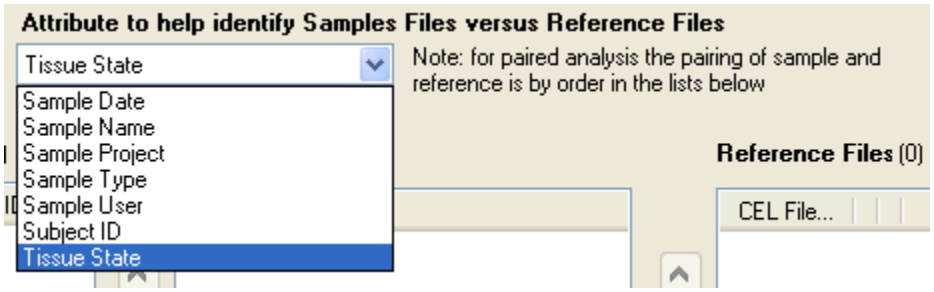
The files are sorted by Enzyme Set Attribute (Figure 12.8).

Figure 12.8 Files sorted by Subject Name Enzyme Set Attribute



10. Select the Sample vs. Reference attribute from the drop-down list (Figure 12.9).

Figure 12.9 Selecting the Sample vs. Reference attribute



The files in the Available Files box are sorted by both the Enzyme Set Shared attribute and the sample/reference attribute (Figure 12.10).

Figure 12.10 Sorted by Enzyme Set and Sample/Reference attributes

Available Files (12, sorted on sample/reference attribute, then enzyme set attribute)

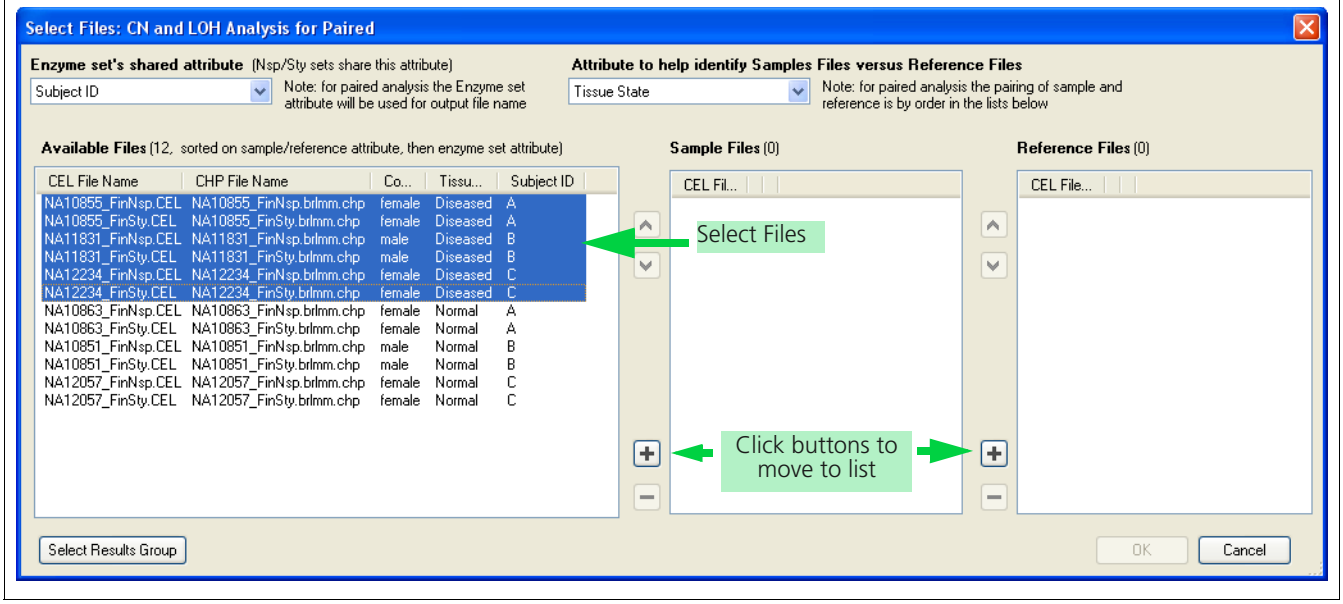
CEL File Name	CHP File Name	Co...	Tissu...	Subject ID
NA10855_FinNsp.CEL	NA10855_FinNsp.brlmm.chp	female	Diseased	A
NA10855_FinSty.CEL	NA10855_FinSty.brlmm.chp	female	Diseased	A
NA11831_FinNsp.CEL	NA11831_FinNsp.brlmm.chp	male	Diseased	B
NA11831_FinSty.CEL	NA11831_FinSty.brlmm.chp	male	Diseased	B
NA12234_FinNsp.CEL	NA12234_FinNsp.brlmm.chp	female	Diseased	C
NA12234_FinSty.CEL	NA12234_FinSty.brlmm.chp	female	Diseased	C
NA10863_FinNsp.CEL	NA10863_FinNsp.brlmm.chp	female	Normal	A
NA10863_FinSty.CEL	NA10863_FinSty.brlmm.chp	female	Normal	A
NA10851_FinNsp.CEL	NA10851_FinNsp.brlmm.chp	male	Normal	B
NA10851_FinSty.CEL	NA10851_FinSty.brlmm.chp	male	Normal	B
NA12057_FinNsp.CEL	NA12057_FinNsp.brlmm.chp	female	Normal	C
NA12057_FinSty.CEL	NA12057_FinSty.brlmm.chp	female	Normal	C

11. In the Select Files dialog box, choose files from the Available Files list and move them to the Sample or Reference Files lists (Figure 12.11).

Click the **Add** button to add data to the Sample Files list or Reference Files list.

Click the **Remove** button to remove data from a list.

Figure 12.11 Moving files in the Select Files dialog box



If the files in the Available Files list are highlighted (Figure 12.12), you will not be able to move them to the Sample or Reference lists until you have selected a results group for the file.

Figure 12.12 Highlighted file (need to select Results Group for the file).

NA12234_Sty.CEL	NA12234_Sty.brlmm.chp	female
NA12264_Nsp.CEL	<Select Results Group>	male
NA12264_Sty.CEL	NA12264_Sty.brlmm.chp	male

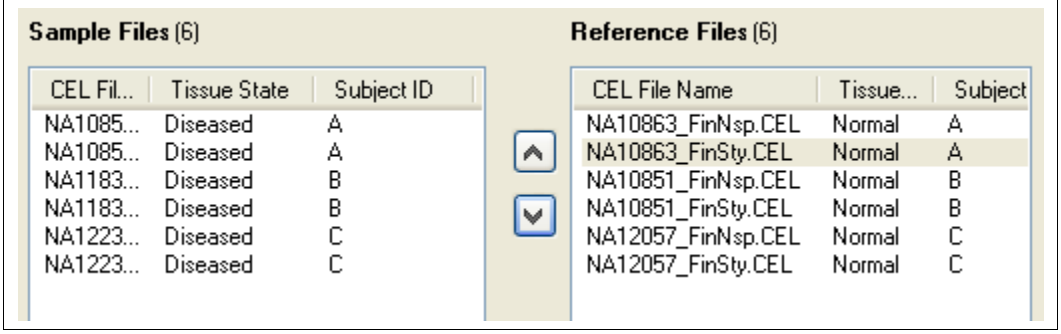
The message "<Select Results Group>" appears if a file is selected for movement to a reference or sample group without first choosing a results group as the destination for the file to be moved. See *Selecting Results Groups* on page 223 for more information.

12. Click the Up and Down buttons to change the file's position and align arrays by enzyme set and sample/reference attributes. The analysis will compare the first sample CEL+CHP in the list with the first reference CEL+CHP, the second sample CEL+CHP with the second reference CEL+CHP, and so on (Figure 12.13).

NOTE: You can also change the sort order of the Sample and Reference files list by clicking on the column headers in the list.

For more information about using shared attributes to pair files by enzyme set or sample/reference group, see *Using Shared Attributes to Group Samples* on page 226.

Figure 12.13 Files paired by enzyme set and sample/reference attributes



13. When the files are paired by enzyme set and sample/reference attributes, click **OK**.

Various error messages may appear if you do not have the samples paired properly or the attributes selected properly (Figure 12.14, Figure 12.15, Figure 12.16).

Figure 12.14 Unable to verify pairing error message

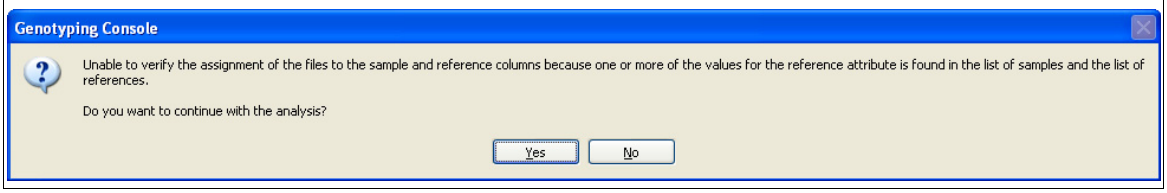
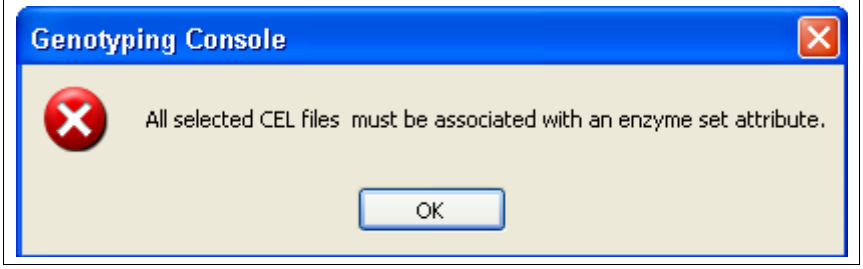


Figure 12.15 Unable to run analysis error message

You will see the following notice (Figure 12.16) if you try running a paired CN/LOH analysis without selecting an enzyme set attribute:

Figure 12.16 Warning notice



See *Using Shared Attributes to Group Samples* on page 226 for more information about using attributes.

The Copy Number and LOH use different naming conventions depending upon whether array enzyme sets are matched or not:

- If array enzyme sets are being matched in the analysis, the output files are named using the Enzyme Set attribute for the arrays.
- If array enzyme sets are not being matched in the analysis (if only CEL files processed using a single enzyme are being analyzed), the output files are named using the CEL file name for the Sample file.

Different progress windows open as the analysis proceeds.

After generating the Copy Number and/or LOH files, you can:

- View the QC data in the *Copy Number QC Summary Table for 100K/500K* on page 231
- *Generate a Segment Report* (page 285)
- *View the CN/LOH/CN Segment data in the GTC Browser* (page 305)
- *Export data to other software* (page 307)

The data file format is described in *Copy Number/LOH File Format for Human Mapping 100K/500K Array Data* on page 221.

Unpaired Copy Number and LOH Analysis

Unpaired CN Analysis is used to compare sample files to a set of reference files.

The software requires that batch genotyping analysis is performed on the data (CEL -> CHP files) before the unpaired Copy Number/LOH analysis is run.

When using a single enzyme array type (50K/250K) in an unpaired Copy Number/LOH analysis, an Enzyme Set attribute is not required.

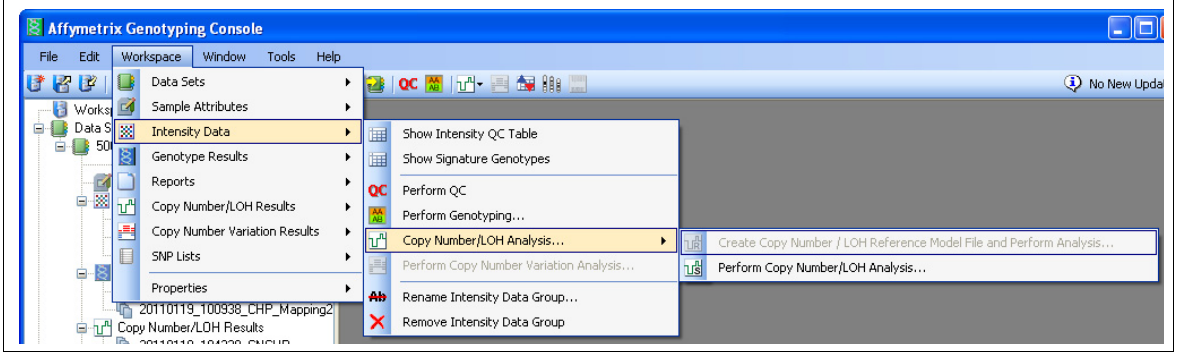
When using Enzyme Sets (100K/500K array sets) an Enzyme Set attribute shared by both members of a sample's enzyme set must be assigned and used to pair arrays in an enzyme set.

The Sample vs. Reference attribute can be helpful if entered, but is not required.

To perform unpaired copy number/LOH analysis:

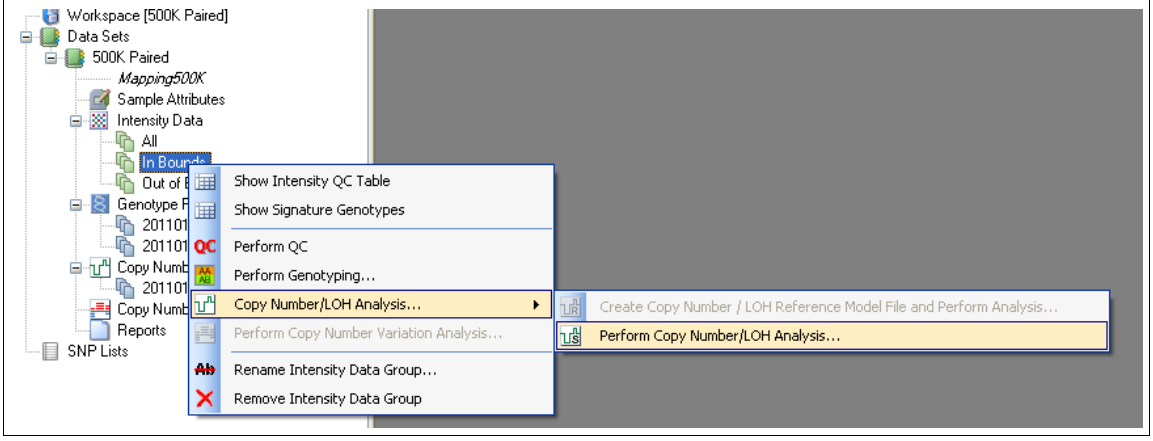
1. Open the Workspace and select the Data Set with the data for analysis.
2. Select the Intensity Data file set from the Data tree.
3. Do one of the following:
   - From the Workspace menu, select **Intensity Data > Copy Number/LOH Analysis > Perform Copy Number/LOH Analysis…** (Figure 12.17).

Figure 12.17 Selecting CN/LOH analysis from Workspace menu



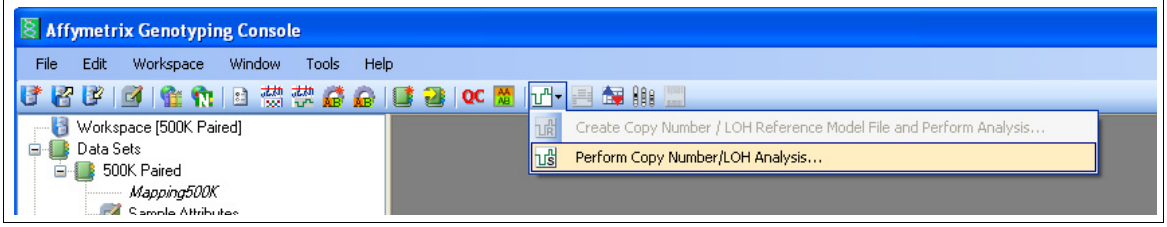
- Right-click the Intensity Data file set and select **Perform Copy Number/LOH Analysis** from the pop-up menu (Figure 12.18).

Figure 12.18 Selecting CN/LOH analysis from the Data tree



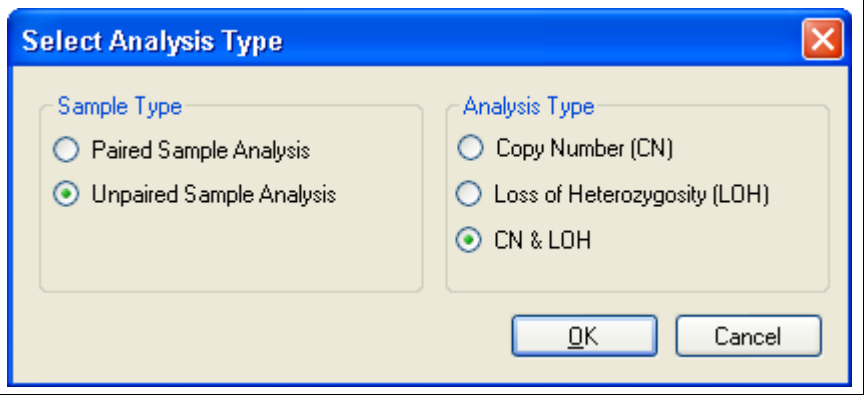
- Click the **Perform Copy Number Analysis** button in the tool bar and select **Perform Copy Number/LOH Analysis** from the dropdown list (Figure 12.19).

Figure 12.19 Selecting CN/LOH analysis from the Tool bar



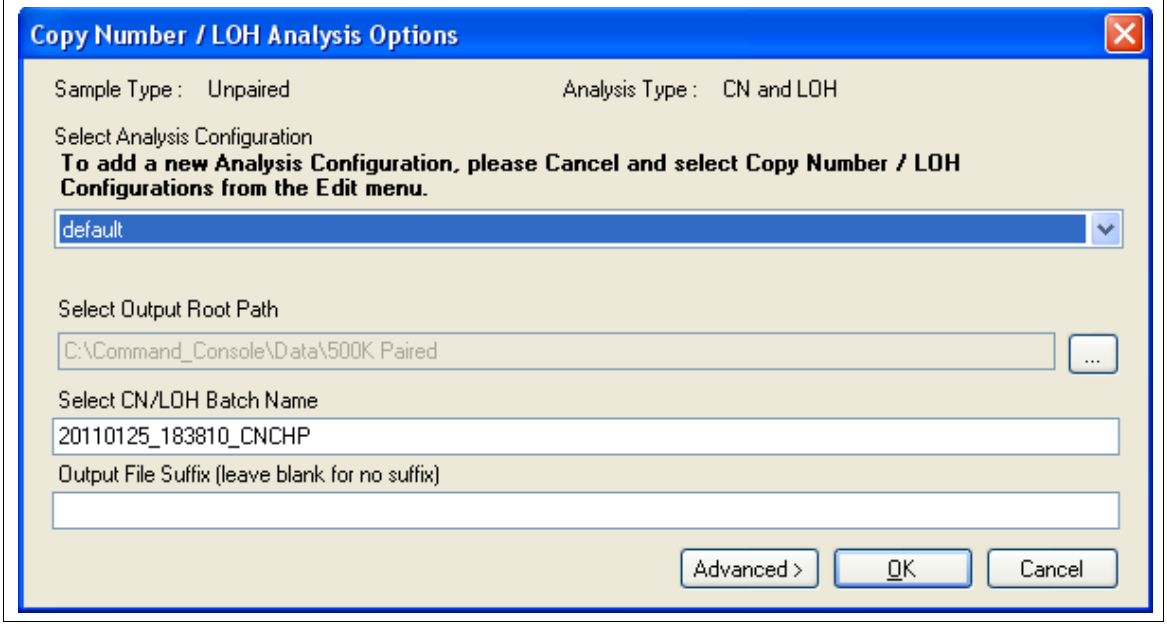
The Copy Number Analysis Options dialog box opens (Figure 12.20).

Figure 12.20 Copy Number Analysis Options dialog box for Unpaired Sample Analysis



4. Select **Un-Paired Sample Analysis** for Sample type.
5. Select the analysis type (CN, LOH, or both).
6. Click **OK**.
   The Copy Number Analysis Options dialog box opens (Figure 12.21).

Figure 12.21 Copy Number/LOH Analysis Options dialog box (unpaired analysis)

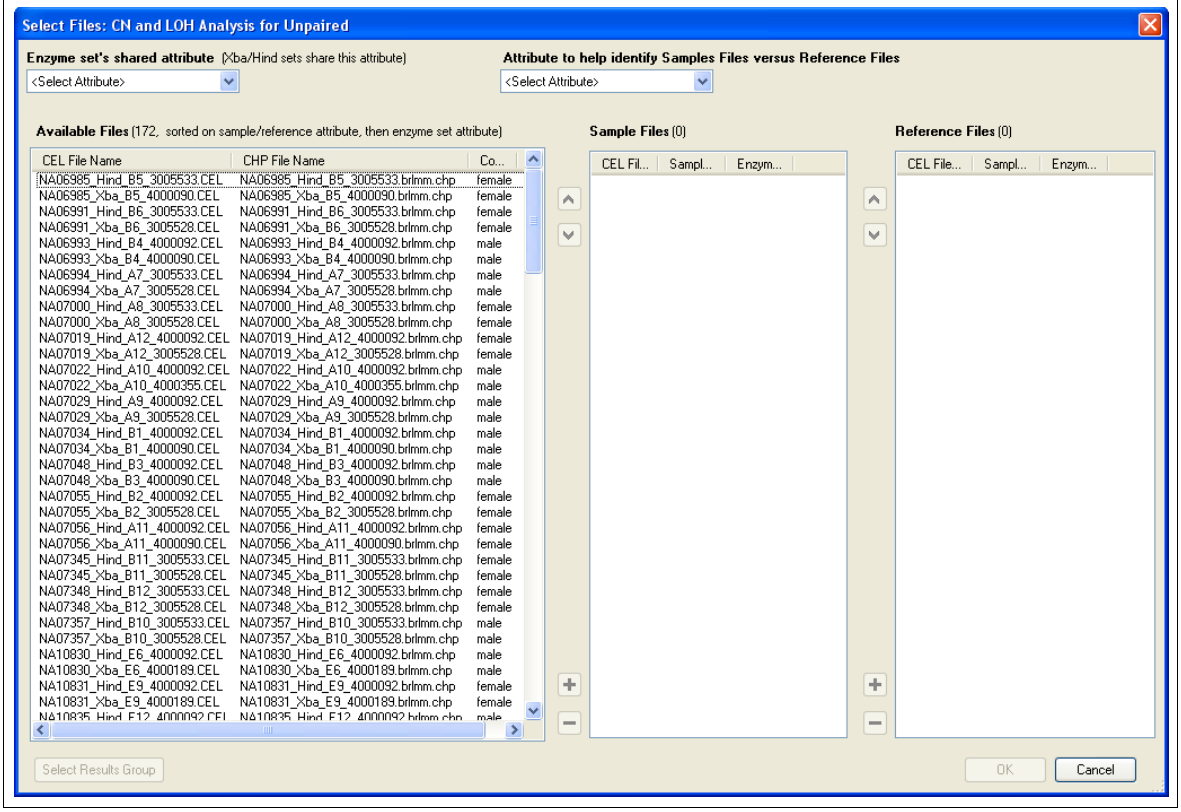


7. Review analysis configuration parameters and select new analysis configuration if desired.
   See *Changing Algorithm Parameters for Human Mapping 100K/500K Analysis* on page 232 for more information on creating a new analysis configuration.
8. Change the following if desired:
   - Output Root Path: location of the CN/LOH Results Group folder.
     Click on the Browse button [...] to search for an output path.
   - Base Batch Name: Name of the CN/LOH Results Group folder.

     NOTE: This folder is the location where the different Data Results files are kept. You can access the folder through Windows Explore to view report files.

   - Output File Suffix: suffix added to distinguish output file names.
9. Click **OK**.
   The Select Files dialog box opens (Figure 12.22).

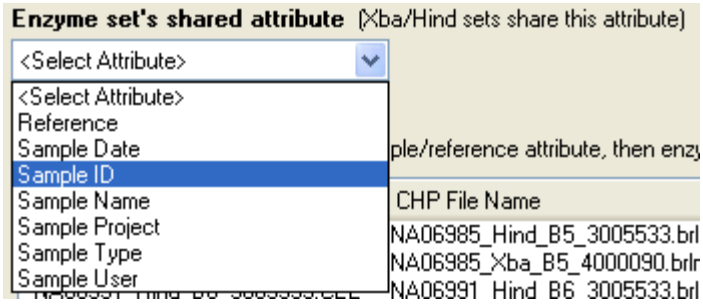
Figure 12.22 Select Files dialog box for unpaired analysis



10. Select the Enzyme Set shared attribute from the Enzyme Set Shared Attribute drop-down list (Figure 12.23).

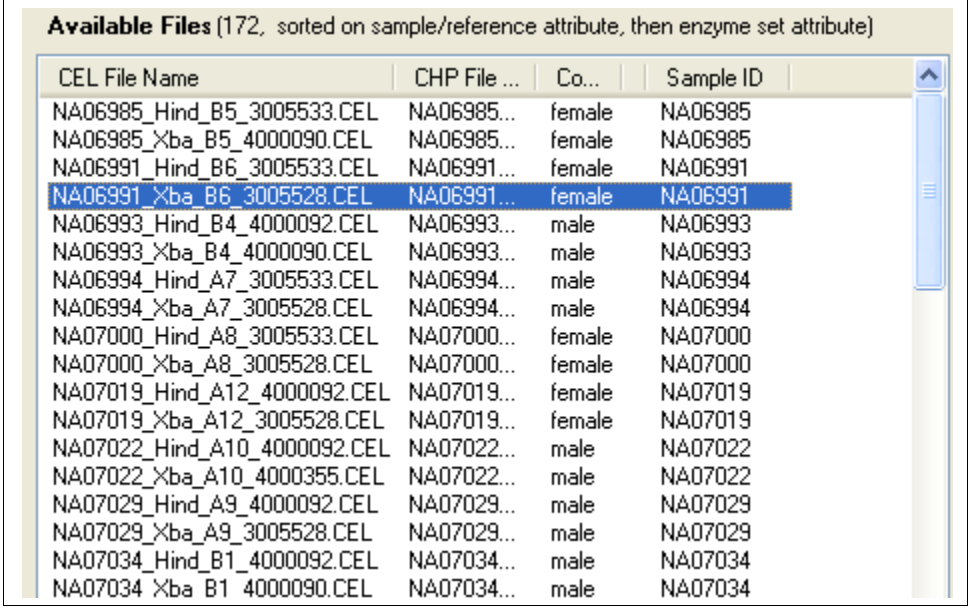
NOTE: This step is not required if you are analyzing a single Enzyme array type.

Figure 12.23 Enzyme Set Shared Attribute dropdown list



The files are sorted by Enzyme Set Attribute (Figure 12.24).

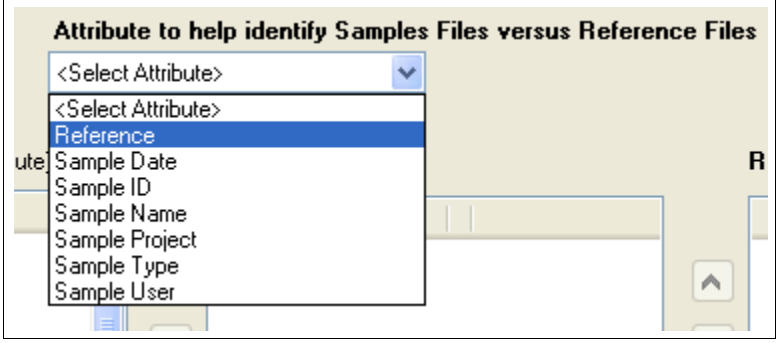
Figure 12.24 Files sorted by Enzyme Set



11. Select the Sample vs. Reference attribute from the drop-down list (Figure 12.25).

NOTE: This step is not required but may be useful if you have assigned attributes to samples you wish to use as Samples and References.

Figure 12.25 Selecting the Sample/Reference Attribute



The files are sorted by the sample/reference attribute (Figure 12.26).

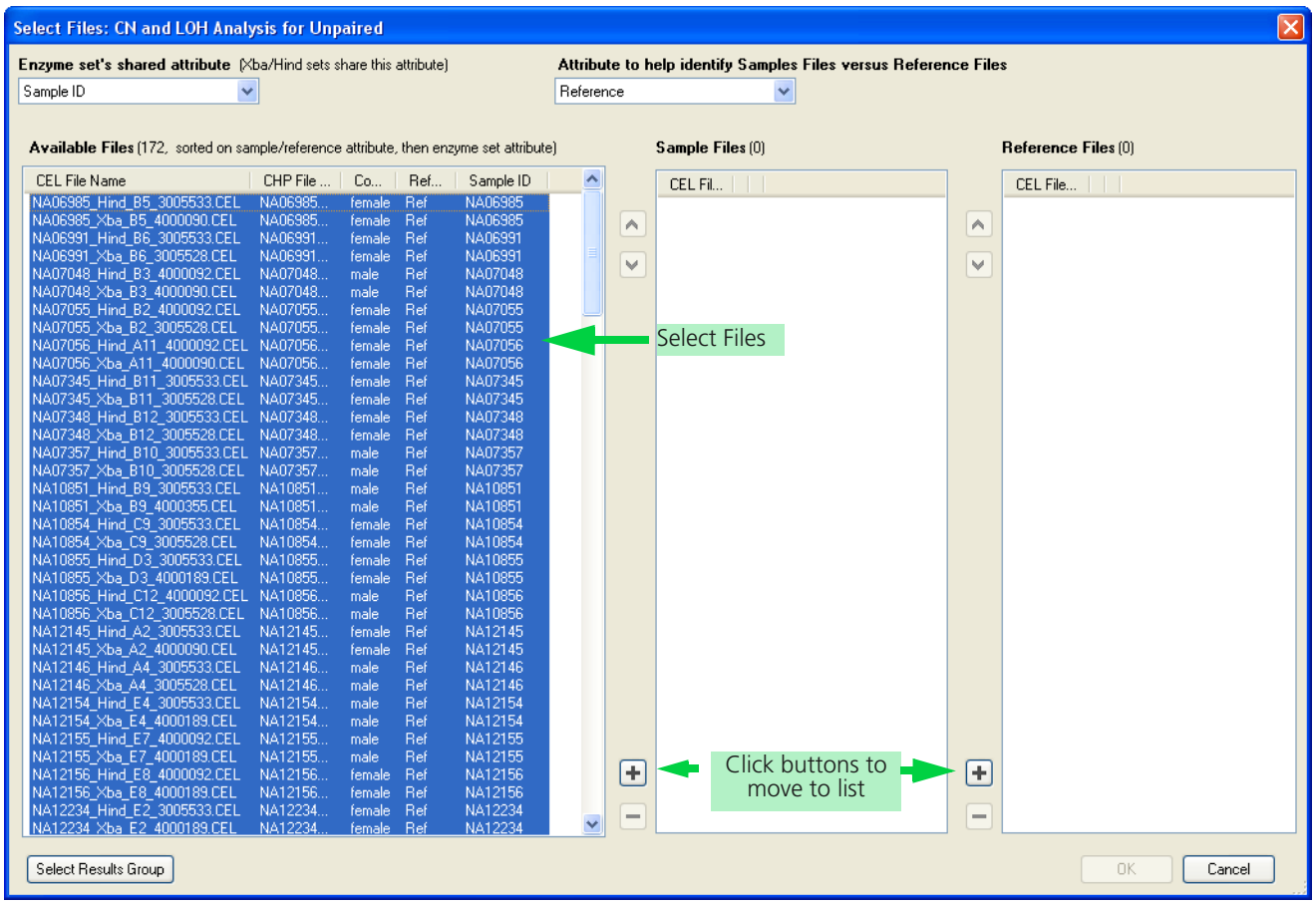
Figure 12.26 Sorted by Enzyme Set and Sample/Reference attributes

Available Files (172, sorted on sample/reference attribute, then enzyme set attribute)

CEL File Name	CHP File ...	Co...	Ref...	Sample ID
NA12874_Xba_H1_4000355.CEL	NA12874...	male	Ref	NA12874
NA12875_Hind_H2_4000092.CEL	NA12875...	female	Ref	NA12875
NA12875_Xba_H2_4000355.CEL	NA12875...	female	Ref	NA12875
NA12878_Hind_H6_4000092.CEL	NA12878...	female	Ref	NA12878
NA12878_Xba_H6_4000355.CEL	NA12878...	female	Ref	NA12878
NA12891_Hind_H4_4000092.CEL	NA12891...	male	Ref	NA12891
NA12891_Xba_H4_4000355.CEL	NA12891...	male	Ref	NA12891
NA12892_Hind_H5_4000092.CEL	NA12892...	female	Ref	NA12892
NA12892_Xba_H5_4000355.CEL	NA12892...	female	Ref	NA12892
NA06993_Hind_B4_4000092.CEL	NA06993...	male	Sample	NA06993
NA06993_Xba_B4_4000090.CEL	NA06993...	male	Sample	NA06993
NA06994_Hind_A7_3005533.CEL	NA06994...	male	Sample	NA06994
NA06994_Xba_A7_3005528.CEL	NA06994...	male	Sample	NA06994
NA07000_Hind_A8_3005533.CEL	NA07000...	female	Sample	NA07000
NA07000_Xba_A8_3005528.CEL	NA07000...	female	Sample	NA07000
NA07019_Hind_A12_4000092.CEL	NA07019...	female	Sample	NA07019
NA07019_Xba_A12_3005528.CEL	NA07019...	female	Sample	NA07019

12. Select files in the Available Files list (Figure 12.27).
Click the **Add** button to add data to the Sample Files list or Reference Files list.
Click the **Remove** button to remove data from a list.

Figure 12.27 Moving files in the Select Files dialog box



If the files in the Available Files list are highlighted, you will not be able to move them to the Sample or Reference lists until you have selected a results group for the file (Figure 12.28).

Figure 12.28 Highlighted file (need to select Results Group)

NA12234_Sty.CEL	NA12234_Sty.brlmm.chp	female
NA12264_Nsp.CEL	<Select Results Group>	male
NA12264_Sty.CEL	NA12264_Sty.brlmm.chp	male

See *Selecting Results Groups* on page 223 for more information.

13. Click the Up and Down buttons to change the file's position and align arrays by enzyme set.

NOTE: The reference set for unpaired analysis for Human Mapping100K/500K (CN4) analysis should consist of at least 25 samples, preferably all female. Reference samples should all be female for best results on the X chromosome. If X chromosome information is not important, male samples may be used in the reference set. For more information, see the Affymetrix website for the white paper "Copy Number and Loss of Heterozygosity Estimation Algorithms for the GeneChip Human Mapping Array Sets"

NOTE: You can also change the sort order of the Sample and Reference files list by clicking on the column headers in the list.

For more information about using shared attributes to pair files by enzyme set or sample/reference group, see *Using Shared Attributes to Group Samples* on page 226.

14. Click **OK**.

IMPORTANT: The Copy Number and LOH output files will be named using the Enzyme Set attribute for the arrays.

Different progress windows open as the analysis proceeds.

The Copy Number and LOH files use different naming conventions depending upon whether array enzyme sets are matched or not:

- If array enzyme sets are being matched in the analysis, the output files are named using the Enzyme Set attribute for the arrays.
- If array enzyme sets are not being matched in the analysis (if only CEL files processed using a single enzyme are being analyzed), the output files are named using the CEL file name for the Sample file.

After generating the Copy Number and/or LOH files, you can:

- View the QC data in the *Copy Number QC Summary Table for 100K/500K* on page 231
- *Generate a Segment Report* (page 285)
- *View the CN/LOH/CN Segment data in the GTC Browser* (page 305)
- *Export data to other software* (page 307)

The data file format is described in *Copy Number/LOH File Format for Human Mapping 100K/500K Array Data* on page 221.

The data file format is described in *Copy Number/LOH File Format for Human Mapping 100K/500K Array Data* on page 221.

Copy Number/LOH File Format for Human Mapping 100K/500K Array Data

The copy number and LOH data are in separate files for Human Mapping 100K/500K array data.

The Copy Number and LOH use different naming conventions depending upon whether array enzyme sets are matched or not:

- If array enzyme sets are being matched in the analysis, the output files are named using the Enzyme Set attribute for the arrays.
- If array enzyme sets are being matched in the analysis (if only CEL files processed using a single enzyme are being analyzed), the output files are named using the CEL file name for the Sample file.

Header Section

The resulting CN4.cnchp and CN4.lohchp data files contain the following information in the header:

- Information about the array (number of SNPs, probe array type, and library file)
- Algorithm parameters and command line that was executed (e.g. all advanced parameters that were used)
- Workflow (e.g. paired copy number)
- Sample Name
- Reference file(s) used

Data Section – For *.CN4.cnchp (Copy Number) Files

The resulting *.CN4.cnchp data files contain the data shown in the table below (Table 12.1).

NOTE: Those values that are labeled "paired analysis only" require that the Generate Allele-Specific Copy Number check box is selected in the Advanced Analysis options.

Table 12.1 Data items for CNCHP files

Item	Description
ProbeSet	SNP ID
Chromosome	Chromosome number
Position	Physical position of the SNP
Log2Ratio	Smoothed Log2 ratio value
HmmMedianLog2Ratio	Median Log2 ratio value of all contiguous SNPs in the given HMM copy number state segment
CNState	HMM copy number state
NegLog10PValue	Negative Log10 p-value indicating how different the median Log2 ratio of the HMM state is from the normal state (CN State 2) for that particular sample
Log2RatioMin	Smoothed Log2 ratio value for the allele with the lower signal intensity (paired analysis only)
HmmMedianLog2RatioMin	Median Log2 ratio value of all the contiguous SNPs in the given HMM copy number state segment of the allele with the lower signal intensity (paired analysis only)
CNStateMin	HMM copy number state of the allele with the lower signal intensity (paired analysis only)
NegLog10PValueMin	Negative Log10 p-value indicating how different the median Log2 ratio of the HMM state of the allele with the lower signal intensity is from the CN 2 State for that particular sample (paired analysis only)
Log2RatioMax	Smoothed Log2 ratio value for the allele with the higher signal intensity (paired analysis only)
HmmMedianLog2RatioMax	Median Log2 ratio value of all the contiguous SNPs in the given HMM copy number state segment of the allele with the higher signal intensity (paired analysis only)
CNStateMax	HMM copy number state of the allele with the higher signal intensity (paired analysis only)
NegLog10PValueMax	Negative Log10 p-value indicating how different the median Log2 ratio of the HMM state of the allele with the higher signal intensity is from the CN 2 State for that particular sample (paired analysis only)
Chip#	The Array ID (1 or 2) where the SNP resides: 1 = The first array in the virtual set as displayed in the Sample List box. 2 = The second array in the virtual set as displayed in the Sample List box.

Data Section – For *.CN4.lohchp (LOH) files

The resulting *.CN4.lohchp data files contain the data shown in the table below (Table 12.2).

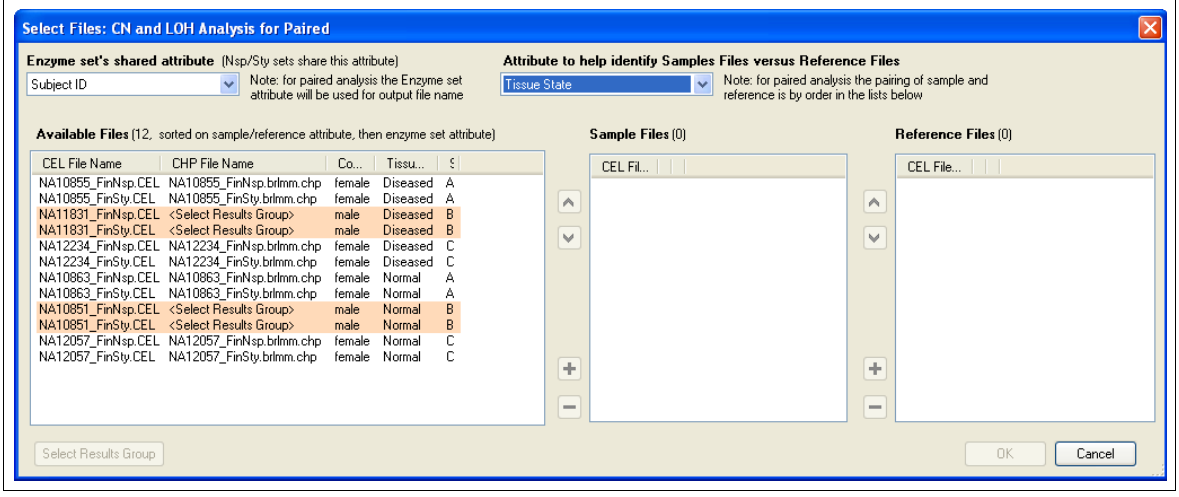
Table 12.2 Data for LOH files

Item	Description
ProbeSet	SNP ID
Chromosome	Chromosome number
Position	Physical position of the SNP
Call	Genotype call for the tumor/test sample
RefCall	Genotype call for the paired reference sample (paired analysis only)
RefHetRate	Heterozygosity rate of the given SNP in the reference samples (un-paired analysis only)
LOHState	1=LOH and 0=Retention
LOHProb	Likelihood that a SNP is in LOH state (closer to 1 indicates a strong likelihood of LOH)
RetProb	Likelihood that a SNP is in Retention state (closer to 1 indicates a strong likelihood of Retention)

Selecting Results Groups

If a CEL file selected for CN/LOH analysis has more than one set of genotype results, you will see the file highlighted in the Available files list (Figure 12.29).

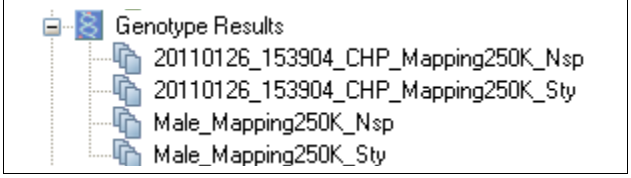
Figure 12.29 Highlighted files in the Available Files list



This will occur if a particular CEL file has been genotyped in more than a single batch, or if the same CHP file is present in more than one results group (Figure 12.30).

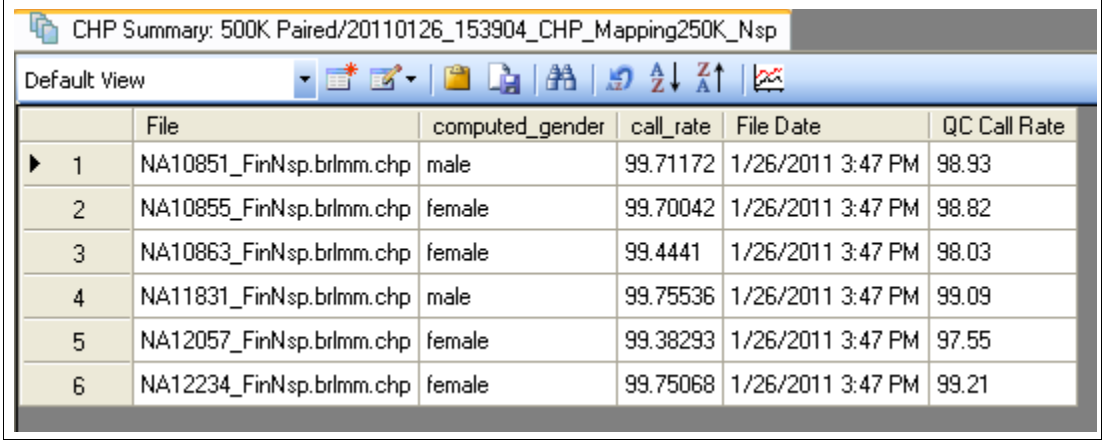
In the example below, the male sample has been separated out into an additional results set.

Figure 12.30 Multiple Genotype Results groups



The figure below (Figure 12.31) shows the full results set with all results files.

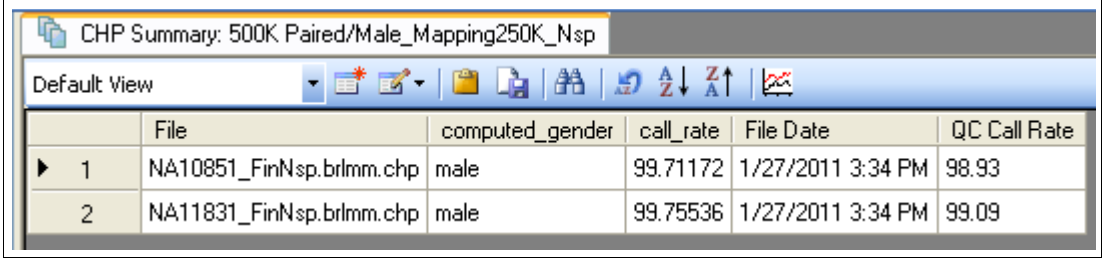
Figure 12.31 First REsults Set



	File	computed_gender	call_rate	File Date	QC Call Rate
1	NA10851_FinNsp.brlmm.chp	male	99.71172	1/26/2011 3:47 PM	98.93
2	NA10855_FinNsp.brlmm.chp	female	99.70042	1/26/2011 3:47 PM	98.82
3	NA10863_FinNsp.brlmm.chp	female	99.4441	1/26/2011 3:47 PM	98.03
4	NA11831_FinNsp.brlmm.chp	male	99.75536	1/26/2011 3:47 PM	99.09
5	NA12057_FinNsp.brlmm.chp	female	99.38293	1/26/2011 3:47 PM	97.55
6	NA12234_FinNsp.brlmm.chp	female	99.75068	1/26/2011 3:47 PM	99.21

The figure below (Figure 12.32) shows the male results set with data from male samples.

Figure 12.32 Second Results set with the same CHP file names



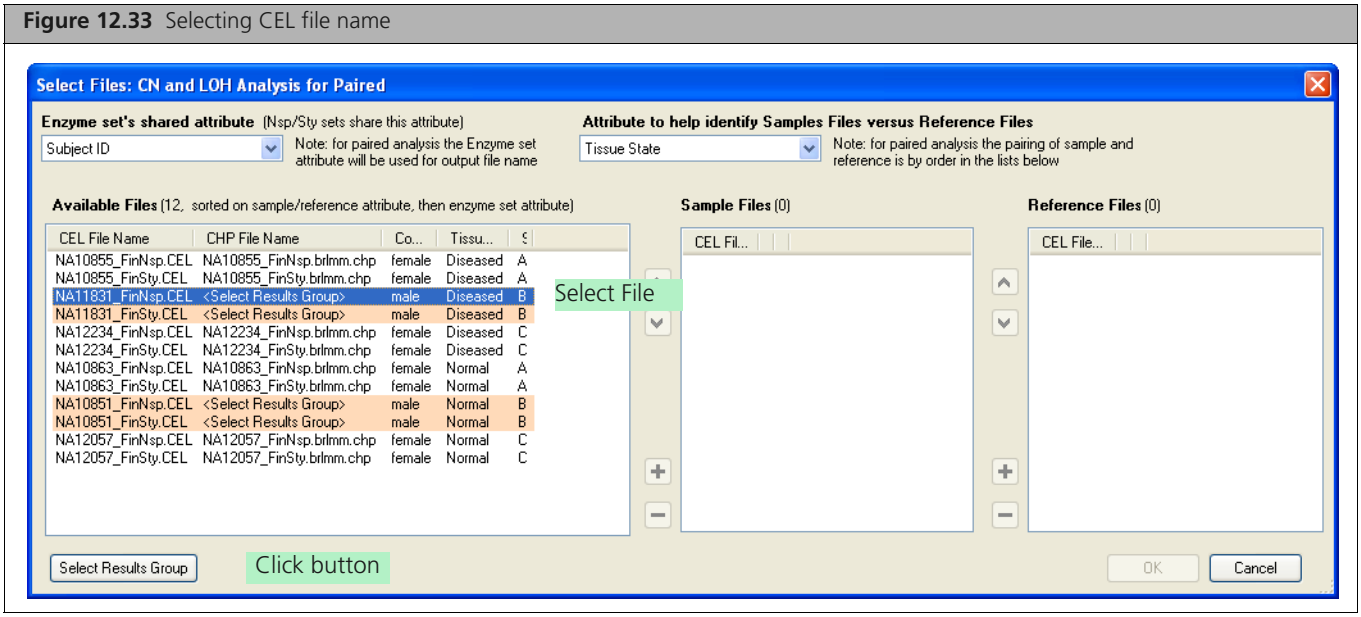
	File	computed_gender	call_rate	File Date	QC Call Rate
1	NA10851_FinNsp.brlmm.chp	male	99.71172	1/27/2011 3:34 PM	98.93
2	NA11831_FinNsp.brlmm.chp	male	99.75536	1/27/2011 3:34 PM	99.09

You will not be able to select the highlighted files and move them to the Sample or Reference Files lists until you choose a CHP file from a results group.

To select the Results set for a file:

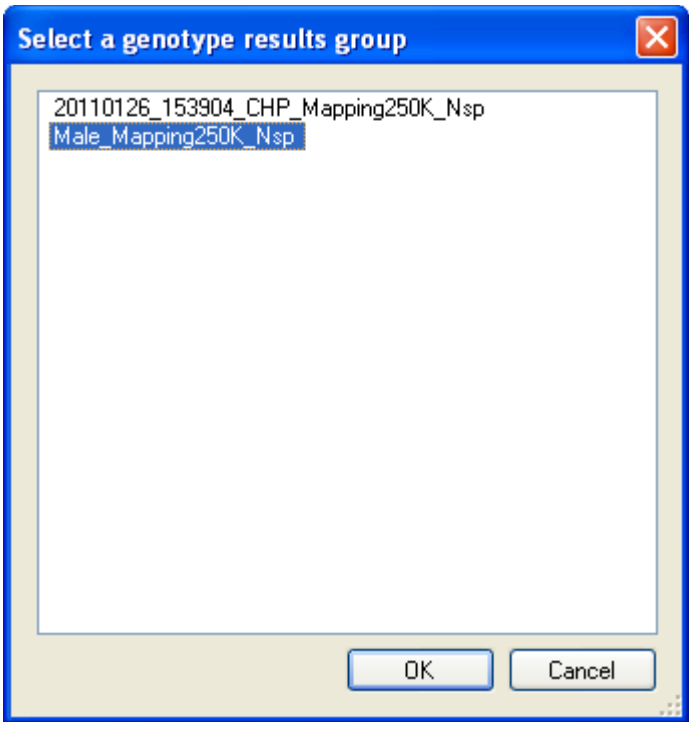
1. Select the CEL file name and click the **Select Results Group** button (Figure 12.33).

Figure 12.33 Selecting CEL file name



The Select Results Group dialog box opens (Figure 12.34).

Figure 12.34 Select Results Group dialog box.



2. Select the Results group with the file you wish to use and click **OK**.
   The file in the Available Files list displays the CHP file name (Figure 12.35).

Figure 12.35 File with selected Results Group

NA12234_Sty.CEL	NA12234_Sty.brlmm.chp	female
NA12264_Nsp.CEL	NA12264_Nsp.brlmm.chp	male
NA12264_Sty.CEL	NA12264_Sty.brlmm.chp	male

You can now select the file and move it to the Sample Files List or Reference Files list.

Using Shared Attributes to Group Samples

Attributes in the array files (.ARR, .XML) can be used to group samples for different analysis types.

You assign a common attribute to:

- Pair the two different enzyme set arrays arising from the same biological source/state.

  The Enzyme Set shared attribute FUNCTIONALLY couples the two array enzyme set types from the single biological sample, inextricably linking and interleaving the data together in the resulting single cnchp file (and/or single lohchp file). The Enzyme Set Attribute is a functional attribute required by GTC software to enable the CN4 algorithm to run correctly when paired CN/LOH or unpaired enzyme set arrays are analyzed for CN/LOH.
- Match up arrays for paired analysis from the same sample.

  The Sample/Reference pairing sorts out the list and is therefore helpful, but optional, and is not required by the algorithm in any way.

Using shared attributes allows you to sort files for easier selection and informs you if you have made certain mistakes in pairing files.

Enzyme Set Shared Attribute (Functionally required in all paired and enzyme set unpaired Copy Number/LOH analysis)

The Human Mapping 100K and 500K arrays use two different physical arrays to cover the entire set of SNPs.

- Human Mapping 100K includes the following arrays:
  - Mapping50K_Xba240
  - Mapping50K_Hind240
- Human Mapping 500K includes the following arrays:
  - Mapping250K_Nsp
  - Mapping250K_Sty

Running the same biological sample on both arrays in a set is necessary to completely cover the genome.

You can group analysis results from the two arrays for one sample into one copy number data (CNCHP) file using the Enzyme Set Shared Attribute to group arrays.

It is necessary to match enzyme sets with the Enzyme Set Attribute, whether you are performing a paired or unpaired CN/LOH analysis.

To group samples using enzyme set attributes:

1. Put the Sample (ARR or XML), Intensity (CEL) and Genotyping (CHP) files for both array types in the same data set.
2. Specify the necessary attributes for Enzyme Set in the Sample files. This should be done during initial sample registration, but you can add and edit the attributes using GCOS or AGCC later on.

   Each pair of enzyme set arrays needs to be assigned at least one shared attribute unique to the CN/LOH analyses of which it will be a part. For example, in the figure below (Figure 12.36), the Patient_State attribute is the attribute used to pair the enzyme set arrays from a single sample.

Figure 12.36 Table of sample files (run as array sets) displaying different sample attributes (for example, Patient_State) that can be used to pair the sample sets

	File	Patient ID	Patient_State	Sample Type	Gender	File Date	# CELs Per Sample
1	NA10851_Nsp.ARR	A	A_Disease	Disease	M	10/24/2007 3:03 PM	1
2	NA10851_Sty.ARR	A	A_Disease	Disease	M	10/24/2007 3:03 PM	1
3	NA10855_Nsp.ARR	B	B_Normal	Normal	M	10/24/2007 3:03 PM	1
4	NA10855_Sty.ARR	B	B_Normal	Normal	M	10/24/2007 3:03 PM	1
5	NA10863_Nsp.ARR	B	B_Disease	Disease	M	10/24/2007 3:03 PM	1
6	NA10863_Sty.ARR	B	B_Disease	Disease	M	10/24/2007 3:03 PM	1
7	NA11831_Nsp.ARR	A	A_Normal	Normal	M	10/24/2007 3:03 PM	1
8	NA11831_Sty.ARR	A	A_Normal	Normal	M	10/24/2007 3:03 PM	1
9	NA12056_Nsp.ARR	C	C_Disease	Disease	M	10/24/2007 3:03 PM	1
10	NA12056_Sty.ARR	C	C_Disease	Disease	M	10/24/2007 3:03 PM	1
11	NA12057_Nsp.ARR	D	D_Normal	Normal	F	10/24/2007 3:03 PM	1
12	NA12057_Sty.ARR	D	D_Normal	Normal	F	10/24/2007 3:03 PM	1
13	NA12234_Nsp.ARR	D	D_Disease	Disease	F	10/24/2007 3:03 PM	1
14	NA12234_Sty.ARR	D	D_Disease	Disease	F	10/24/2007 3:03 PM	1
15	NA12264_Nsp.ARR	E	E_Disease	Disease	M	10/24/2007 3:03 PM	1
16	NA12264_Sty.ARR	E	E_Disease	Disease	M	10/24/2007 3:03 PM	1
17	NA12707_Nsp.ARR	E	E_Normal	Normal	M	10/24/2007 3:03 PM	1
18	NA12707_Sty.ARR	E	E_Normal	Normal	M	10/24/2007 3:03 PM	1
19	NA12716_Nsp.ARR	C	C_Normal	Normal	M	10/24/2007 3:03 PM	1
20	NA12716_Sty.ARR	C	C_Normal	Normal	M	10/24/2007 3:03 PM	1

Sample vs. Reference Shared Attribute (Helpful but never required for analysis)

This attribute pairing is useful when performing paired CN analysis; it enables sorting the Sample/Reference data for easier selection and provides a basic check to make sure they were not mixed up.

To group files using Sample/Reference attributes:

1. Put the Sample (ARR or XML), Intensity (CEL) and Genotyping (CHP) files for both array types in the same data set.
2. Specify the necessary attributes for Sample vs. Reference in the Sample files. This should be done during initial sample registration, but you can add and edit the attributes using GCOS or AGCC later on.

   A Sample vs. Reference attribute should be designated for the files. All Sample files should be assigned one attribute value, and all Reference files should be assigned a different attribute value.

Figure 12.37 Table of Sample Files paired by Sample type attribute

	File	Patient ID	Patient_State	Sample Type	Gender	File Date	# CELs Per Sample
1	NA10851_Nsp.ARR	A	A_Disease	Disease	M	10/24/2007 3:03 PM	1
2	NA10851_Sty.ARR	A	A_Disease	Disease	M	10/24/2007 3:03 PM	1
3	NA10855_Nsp.ARR	B	B_Normal	Normal	M	10/24/2007 3:03 PM	1
4	NA10855_Sty.ARR	B	B_Normal	Normal	M	10/24/2007 3:03 PM	1
5	NA10863_Nsp.ARR	B	B_Disease	Disease	M	10/24/2007 3:03 PM	1
6	NA10863_Sty.ARR	B	B_Disease	Disease	M	10/24/2007 3:03 PM	1
7	NA11831_Nsp.ARR	A	A_Normal	Normal	M	10/24/2007 3:03 PM	1
8	NA11831_Sty.ARR	A	A_Normal	Normal	M	10/24/2007 3:03 PM	1
9	NA12056_Nsp.ARR	C	C_Disease	Disease	M	10/24/2007 3:03 PM	1
10	NA12056_Sty.ARR	C	C_Disease	Disease	M	10/24/2007 3:03 PM	1
11	NA12057_Nsp.ARR	D	D_Normal	Normal	F	10/24/2007 3:03 PM	1
12	NA12057_Sty.ARR	D	D_Normal	Normal	F	10/24/2007 3:03 PM	1
13	NA12234_Nsp.ARR	D	D_Disease	Disease	F	10/24/2007 3:03 PM	1
14	NA12234_Sty.ARR	D	D_Disease	Disease	F	10/24/2007 3:03 PM	1
15	NA12264_Nsp.ARR	E	E_Disease	Disease	M	10/24/2007 3:03 PM	1
16	NA12264_Sty.ARR	E	E_Disease	Disease	M	10/24/2007 3:03 PM	1
17	NA12707_Nsp.ARR	E	E_Normal	Normal	M	10/24/2007 3:03 PM	1
18	NA12707_Sty.ARR	E	E_Normal	Normal	M	10/24/2007 3:03 PM	1
19	NA12716_Nsp.ARR	C	C_Normal	Normal	M	10/24/2007 3:03 PM	1
20	NA12716_Sty.ARR	C	C_Normal	Normal	M	10/24/2007 3:03 PM	1

Example

As an example, the figure below (Figure 12.37) displays files for paired analysis on two sample types (Diseased/Normal) from five sources, A, B, C, D, E

Human Mapping 500K arrays are being used, so the samples for each sample type (Diseased or Normal) from each source needs to be run on two arrays, one for each enzyme set, for a total of twenty arrays.

Figure 12.38 Table of files and attributes with Source and Sample type

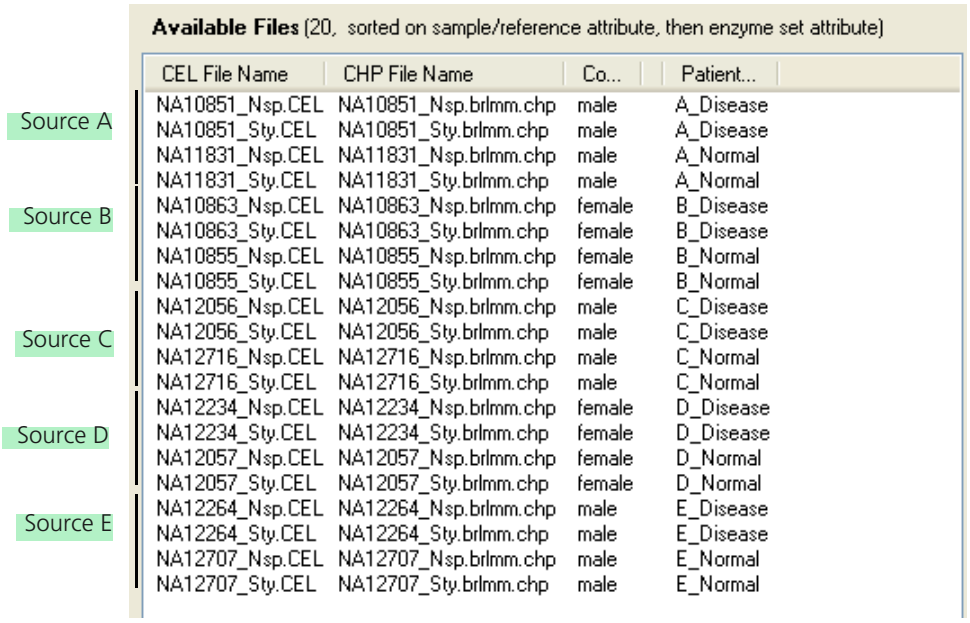
Source

Sample Type

	File	Patient ID	Patient_State	Sample Type	Gender	File Date	# CELs Per Sample
1	NA10851_Nsp.ARR	A	A_Disease	Disease	M	10/24/2007 3:03 PM	1
2	NA10851_Sty.ARR	A	A_Disease	Disease	M	10/24/2007 3:03 PM	1
3	NA10855_Nsp.ARR	B	B_Normal	Normal	M	10/24/2007 3:03 PM	1
4	NA10855_Sty.ARR	B	B_Normal	Normal	M	10/24/2007 3:03 PM	1
5	NA10863_Nsp.ARR	B	B_Disease	Disease	M	10/24/2007 3:03 PM	1
6	NA10863_Sty.ARR	B	B_Disease	Disease	M	10/24/2007 3:03 PM	1
7	NA11831_Nsp.ARR	A	A_Normal	Normal	M	10/24/2007 3:03 PM	1
8	NA11831_Sty.ARR	A	A_Normal	Normal	M	10/24/2007 3:03 PM	1
9	NA12056_Nsp.ARR	C	C_Disease	Disease	M	10/24/2007 3:03 PM	1
10	NA12056_Sty.ARR	C	C_Disease	Disease	M	10/24/2007 3:03 PM	1
11	NA12057_Nsp.ARR	D	D_Normal	Normal	F	10/24/2007 3:03 PM	1
12	NA12057_Sty.ARR	D	D_Normal	Normal	F	10/24/2007 3:03 PM	1
13	NA12234_Nsp.ARR	D	D_Disease	Disease	F	10/24/2007 3:03 PM	1
14	NA12234_Sty.ARR	D	D_Disease	Disease	F	10/24/2007 3:03 PM	1
15	NA12264_Nsp.ARR	E	E_Disease	Disease	M	10/24/2007 3:03 PM	1
16	NA12264_Sty.ARR	E	E_Disease	Disease	M	10/24/2007 3:03 PM	1
17	NA12707_Nsp.ARR	E	E_Normal	Normal	M	10/24/2007 3:03 PM	1
18	NA12707_Sty.ARR	E	E_Normal	Normal	M	10/24/2007 3:03 PM	1
19	NA12716_Nsp.ARR	C	C_Normal	Normal	M	10/24/2007 3:03 PM	1
20	NA12716_Sty.ARR	C	C_Normal	Normal	M	10/24/2007 3:03 PM	1

Sorting the available files by enzyme set results in the following (Figure 12.39):

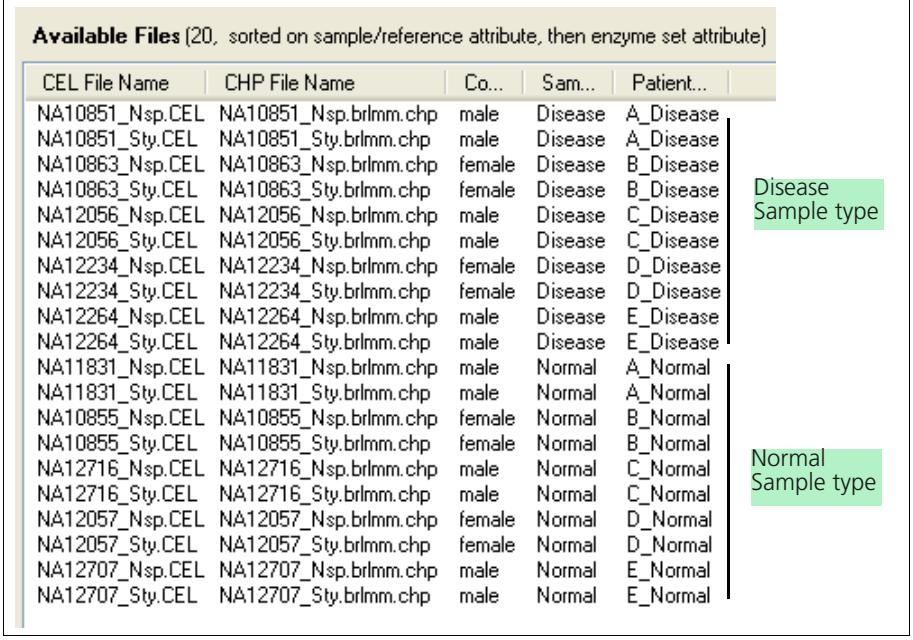
Figure 12.39 Available files Sorted by Enzyme Set Attribute



CEL File Name	CHP File Name	Co...	Patient...
NA10851_Nsp.CEL	NA10851_Nsp.brlmm.chp	male	A_Disease
NA10851_Sty.CEL	NA10851_Sty.brlmm.chp	male	A_Disease
NA11831_Nsp.CEL	NA11831_Nsp.brlmm.chp	male	A_Normal
NA11831_Sty.CEL	NA11831_Sty.brlmm.chp	male	A_Normal
NA10863_Nsp.CEL	NA10863_Nsp.brlmm.chp	female	B_Disease
NA10863_Sty.CEL	NA10863_Sty.brlmm.chp	female	B_Disease
NA10855_Nsp.CEL	NA10855_Nsp.brlmm.chp	female	B_Normal
NA10855_Sty.CEL	NA10855_Sty.brlmm.chp	female	B_Normal
NA12056_Nsp.CEL	NA12056_Nsp.brlmm.chp	male	C_Disease
NA12056_Sty.CEL	NA12056_Sty.brlmm.chp	male	C_Disease
NA12716_Nsp.CEL	NA12716_Nsp.brlmm.chp	male	C_Normal
NA12716_Sty.CEL	NA12716_Sty.brlmm.chp	male	C_Normal
NA12234_Nsp.CEL	NA12234_Nsp.brlmm.chp	female	D_Disease
NA12234_Sty.CEL	NA12234_Sty.brlmm.chp	female	D_Disease
NA12057_Nsp.CEL	NA12057_Nsp.brlmm.chp	female	D_Normal
NA12057_Sty.CEL	NA12057_Sty.brlmm.chp	female	D_Normal
NA12264_Nsp.CEL	NA12264_Nsp.brlmm.chp	male	E_Disease
NA12264_Sty.CEL	NA12264_Sty.brlmm.chp	male	E_Disease
NA12707_Nsp.CEL	NA12707_Nsp.brlmm.chp	male	E_Normal
NA12707_Sty.CEL	NA12707_Sty.brlmm.chp	male	E_Normal

Each sample from diseased tissue is given the value “Disease” in the Sample type attribute, while each sample from normal tissue is given the value “Normal.” The samples can be sorted again on the sample type attribute, as shown in the figure below (Figure 12.40).

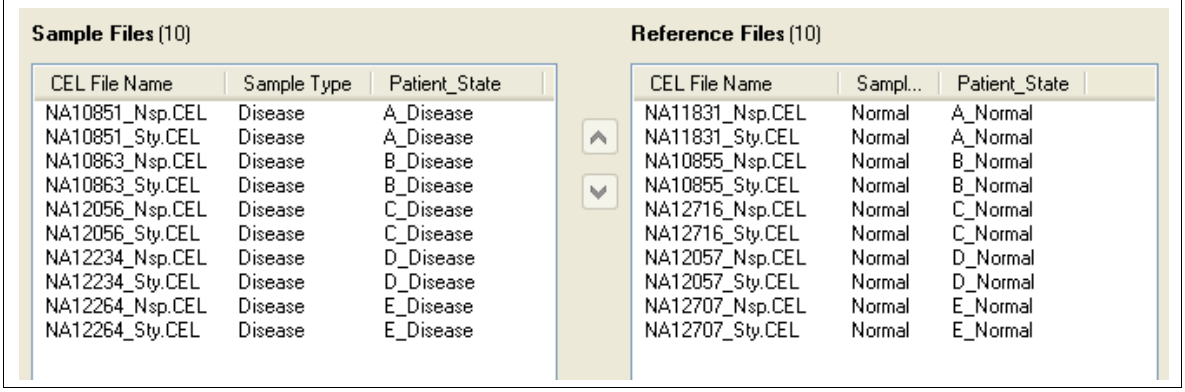
Figure 12.40 Sorted by Enzyme Set and Sample/Reference attributes



CEL File Name	CHP File Name	Co...	Sam...	Patient...
NA10851_Nsp.CEL	NA10851_Nsp.brlmm.chp	male	Disease	A_Disease
NA10851_Sty.CEL	NA10851_Sty.brlmm.chp	male	Disease	A_Disease
NA10863_Nsp.CEL	NA10863_Nsp.brlmm.chp	female	Disease	B_Disease
NA10863_Sty.CEL	NA10863_Sty.brlmm.chp	female	Disease	B_Disease
NA12056_Nsp.CEL	NA12056_Nsp.brlmm.chp	male	Disease	C_Disease
NA12056_Sty.CEL	NA12056_Sty.brlmm.chp	male	Disease	C_Disease
NA12234_Nsp.CEL	NA12234_Nsp.brlmm.chp	female	Disease	D_Disease
NA12234_Sty.CEL	NA12234_Sty.brlmm.chp	female	Disease	D_Disease
NA12264_Nsp.CEL	NA12264_Nsp.brlmm.chp	male	Disease	E_Disease
NA12264_Sty.CEL	NA12264_Sty.brlmm.chp	male	Disease	E_Disease
NA11831_Nsp.CEL	NA11831_Nsp.brlmm.chp	male	Normal	A_Normal
NA11831_Sty.CEL	NA11831_Sty.brlmm.chp	male	Normal	A_Normal
NA10855_Nsp.CEL	NA10855_Nsp.brlmm.chp	female	Normal	B_Normal
NA10855_Sty.CEL	NA10855_Sty.brlmm.chp	female	Normal	B_Normal
NA12716_Nsp.CEL	NA12716_Nsp.brlmm.chp	male	Normal	C_Normal
NA12716_Sty.CEL	NA12716_Sty.brlmm.chp	male	Normal	C_Normal
NA12057_Nsp.CEL	NA12057_Nsp.brlmm.chp	female	Normal	D_Normal
NA12057_Sty.CEL	NA12057_Sty.brlmm.chp	female	Normal	D_Normal
NA12707_Nsp.CEL	NA12707_Nsp.brlmm.chp	male	Normal	E_Normal
NA12707_Sty.CEL	NA12707_Sty.brlmm.chp	male	Normal	E_Normal

This allows the files to be paired up by Enzyme set and by sample/reference pair, as shown in the figure below (Figure 12.41).

Figure 12.41 Selected array files, paired by enzyme set and sample type



Sample Files (10)

CEL File Name	Sample Type	Patient_State
NA10851_Nsp.CEL	Disease	A_Disease
NA10851_Sty.CEL	Disease	A_Disease
NA10863_Nsp.CEL	Disease	B_Disease
NA10863_Sty.CEL	Disease	B_Disease
NA12056_Nsp.CEL	Disease	C_Disease
NA12056_Sty.CEL	Disease	C_Disease
NA12234_Nsp.CEL	Disease	D_Disease
NA12234_Sty.CEL	Disease	D_Disease
NA12264_Nsp.CEL	Disease	E_Disease
NA12264_Sty.CEL	Disease	E_Disease

Reference Files (10)

CEL File Name	Sampl...	Patient_State
NA11831_Nsp.CEL	Normal	A_Normal
NA11831_Sty.CEL	Normal	A_Normal
NA10855_Nsp.CEL	Normal	B_Normal
NA10855_Sty.CEL	Normal	B_Normal
NA12716_Nsp.CEL	Normal	C_Normal
NA12716_Sty.CEL	Normal	C_Normal
NA12057_Nsp.CEL	Normal	D_Normal
NA12057_Sty.CEL	Normal	D_Normal
NA12707_Nsp.CEL	Normal	E_Normal
NA12707_Sty.CEL	Normal	E_Normal

Copy Number QC Summary Table for 100K/500K

The Copy Number QC Summary Table displays QC information about the copy number and LOH analyses.

Use the *GTC Browser* (page 305) to view Copy Number, LOH, and CN Segments data in a genomic context.

The Copy Number QC Summary Table uses all the table options as described in *Table Features* on page 198.

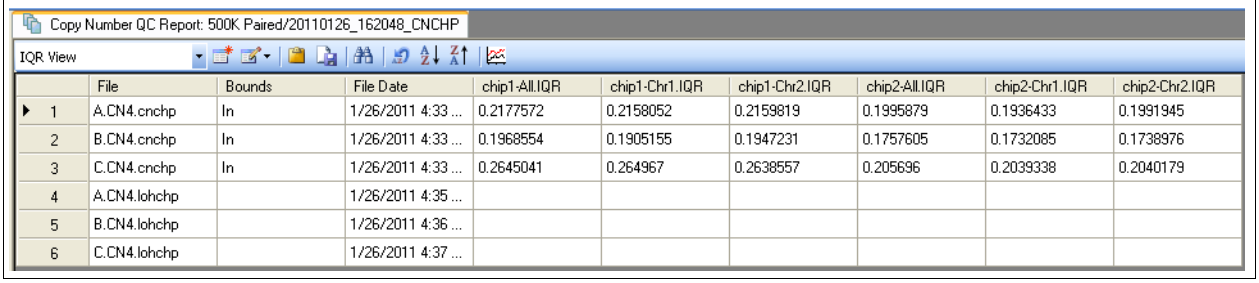
To open the QC Summary table:

- Right-click a Copy Number/LOH Results set and select **Show Copy Number QC Summary Table**; or

  From the Workspace menu, select **Copy Number/LOH Results > Show Copy Number QC Summary Table**.

  The QC Summary table opens (Figure 12.42).

Figure 12.42 Copy Number QC Report for 100K arrays, all columns view



Copy Number QC Report: 500K Paired/20110126_162048_CNCHP

IQR View

	File	Bounds	File Date	chip1-All.IQR	chip1-Chr1.IQR	chip1-Chr2.IQR	chip2-All.IQR	chip2-Chr1.IQR	chip2-Chr2.IQR
1	A.CN4.cnchp	In	1/26/2011 4:33 ...	0.2177572	0.2158052	0.2159819	0.1995879	0.1936433	0.1991945
2	B.CN4.cnchp	In	1/26/2011 4:33 ...	0.1968554	0.1905155	0.1947231	0.1757605	0.1732085	0.1738976
3	C.CN4.cnchp	In	1/26/2011 4:33 ...	0.2645041	0.264967	0.2638557	0.205696	0.2039338	0.2040179
4	A.CN4.lohchp		1/26/2011 4:35 ...						
5	B.CN4.lohchp		1/26/2011 4:36 ...						
6	C.CN4.lohchp		1/26/2011 4:37 ...						

The following information is displayed for Human Mapping 100K/500K arrays in All Columns View:

File	File name.
Bounds	In or out of QC bounds. See *Setting QC Thresholds* on page 312 for more information.
IQR for all chromosomes	Interquartile range average for all chromosomes. The interquartile range (IQR) of the un-smoothed log2ratio smoothed total CN is displayed for each sample. The IQR values are displayed for each chromosome as well as for the whole sample. In a paired analysis, the IQR values are reported for each allele independently. The interquartile range is a measure of dispersion or spread. It is the difference between the 75th percentile (often called Q3 or 3rd quantile) and the 25th percentile (Q1 or first quantile). The formula for interquartile range is therefore: Q3-Q1. Since the IQR represents the central 50% of the data, it is not affected by outliers or extreme values and is hence a robust measure of dispersion. In general the sample-level IQR should be comparable to the chromosomal IQR for the given sample. A discordance in a chromosomal observation is potentially indicative of a biological change.
IQR for individual chromosomes	Interquartile range for each individual chromosome.
File Date	Date the file was created.

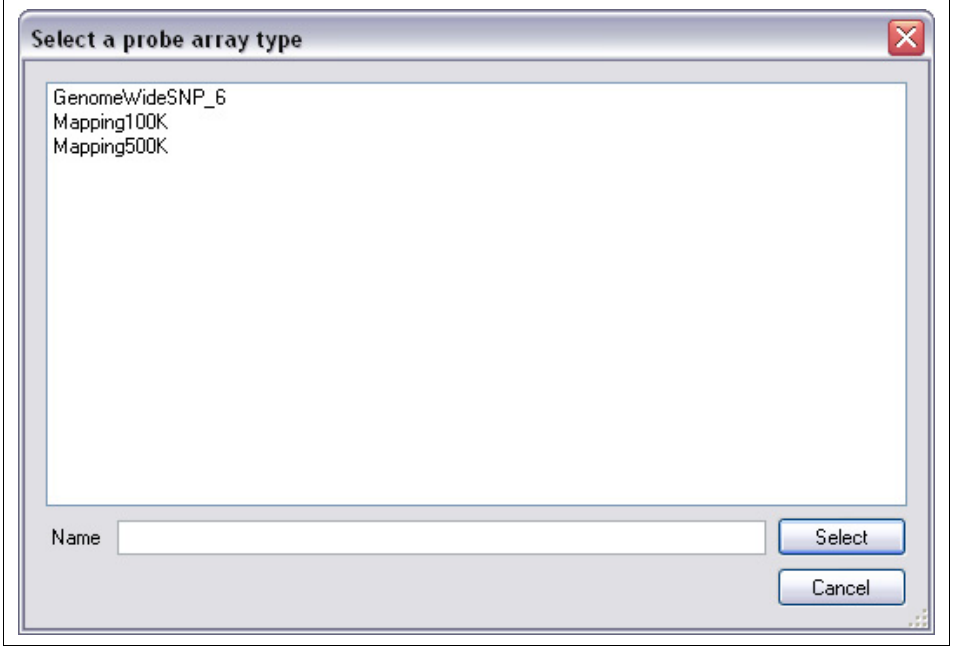
Changing Algorithm Parameters for Human Mapping 100K/500K Analysis

You can change algorithm parameters for the copy number and LOH analysis for Human Mapping 100K/500K arrays.

To open the Configurations dialog box:

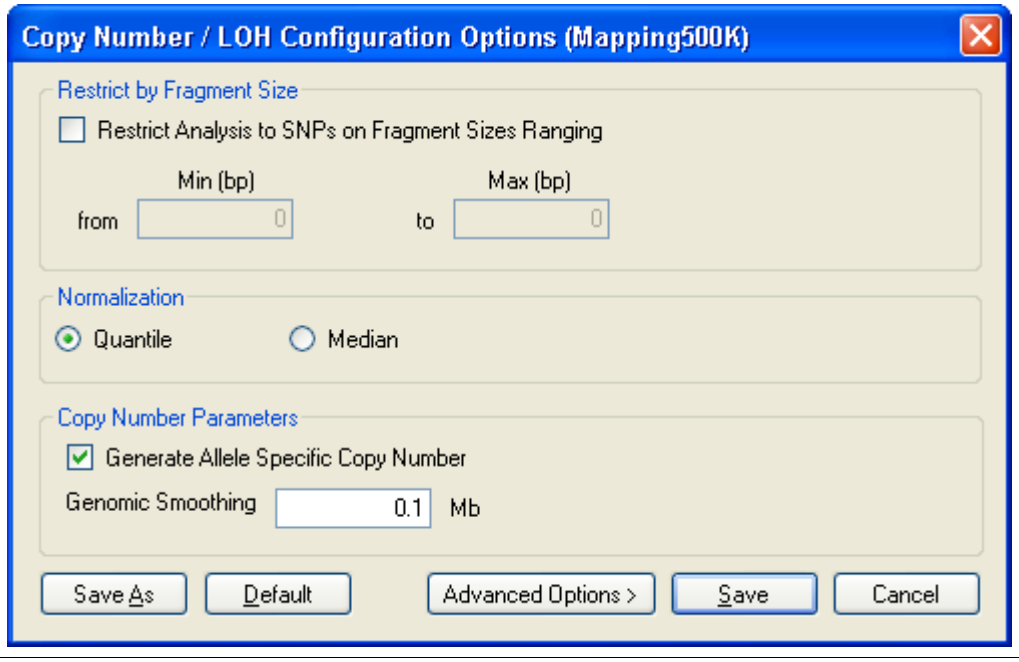
1. From the Edit menu, select **Copy Number Configurations > New Configuration**.
   The Select Probe Array Type dialog box opens (Figure 12.43).

Figure 12.43 Select Probe Array Type dialog box



2. Select Mapping100K or Mapping500K from the list and click **Select**.
   The Copy Number/LOH Configuration Options dialog box opens (Figure 12.44).

Figure 12.44 Basic Configuration Options for Human Mapping 100K/500K arrays

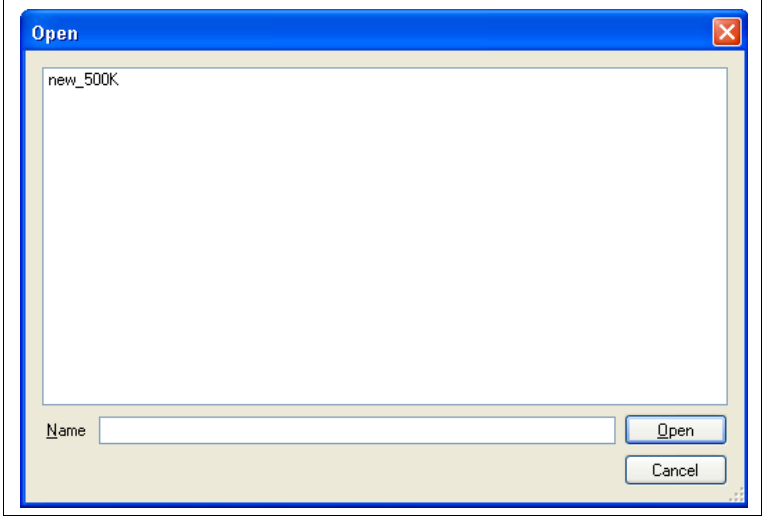


3. Enter values for configuration Options.
   The parameters are described in:
   - *Basic Options* on page 234
   - *Advanced Options* on page 237
4. Save the new configuration file:
   - To save as new configuration: Click **Save As**.
   - Save as default configuration: Click **Default**.

To edit a previously created Configuration.

1. From the Edit menu, select **Copy Number Configurations > Open Configuration**.
   The Open dialog box opens (Figure 12.45).

Figure 12.45 Open dialog box

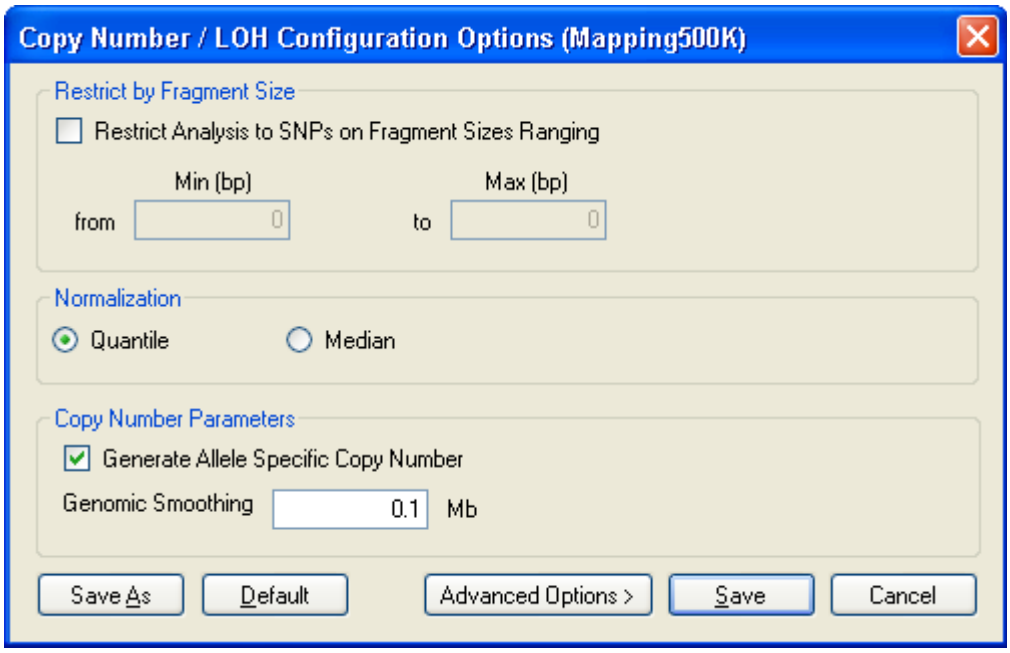


2. Select the configuration file to be edited and click **Open.**
   The Basic Options dialog box opens.
3. Enter values for configuration Options.
   The parameters are described in:
   - *Basic Options* on page 234
   - *Advanced Options* on page 237

Basic Options

The basic options are displayed when the dialog box first opens (Figure 12.46).

Figure 12.46 Basic Configuration Options



The Basic Options allow you to change parameters for:

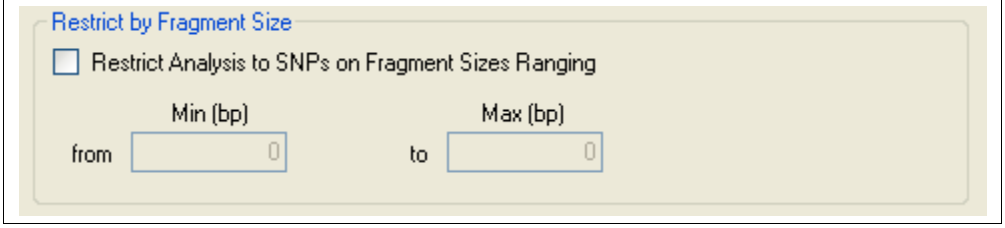
- *Restrict by Fragment Size* on page 234
- *Normalization* on page 235
- *Copy Number Parameters* on page 235

See below for an explanation of these parameters.

Restrict by Fragment Size

This option enables the analysis to be performed on only a subset of SNPs based on the fragment size where the SNPs reside (Figure 12.47). By default, this option is unchecked and all SNPs are included in the analysis.

Figure 12.47 Restrict by Fragment Size



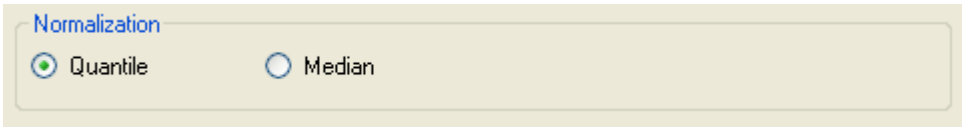
To enable this option:

1. Check the box next to Restrict Analysis to SNPs on Fragment Sizes Ranging.
2. Enter the size of fragments that you want to be included in the analysis.
3. Proceed to further customize the analysis configuration as outlined below or save the configuration changes to exit the dialog box.

Normalization

This option enables specification of the probe-level normalization. Select one of the following two options in the Normalization group box (Figure 12.48).

Figure 12.48 Normalization group box



Quantile

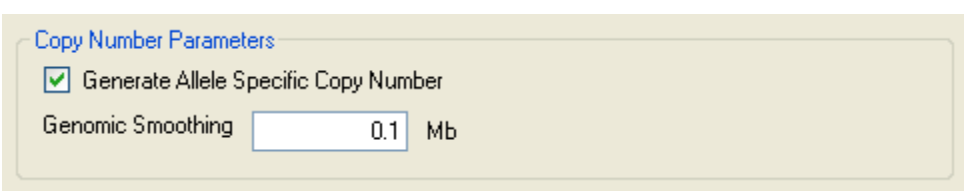
Quantile normalization performs a sketch normalization, based on perfect match (PM) probes across the CEL files. Quantile is the default setting.

Median

Median scaling performs a linear scaling based on the median of all CEL files included in the analysis. All PM and mismatch (MM) probes are included to compute the median intensity of a CEL file.

Copy Number Parameters

Figure 12.49 Copy Number Parameters



Generate Allele Specific Copy Number

For paired analysis, an allele-specific analysis can be performed on the SNPs, which are heterozygous in the paired normal. This option can be disabled by unchecking the Generate Allele Specific Copy Number box.

Genomic Smoothing

The genomic smoothing option allows the user to specify the genomic smoothing length (in megabases) to be used. The genomic smoothing that is applied is a Gaussian smoothing. The default bandwidth value is 100 Kb (0.1 Mb) that results in a window size of 400 Kb. This default is optimized for Human Mapping 500K analyses. For Human Mapping 100K analyses, use 0.5 Mb. Genomic smoothing can be disabled by applying a smoothing bandwidth of 0 bp. See *Copy Number Parameter Settings* on page 240 for recommended CN parameter settings.

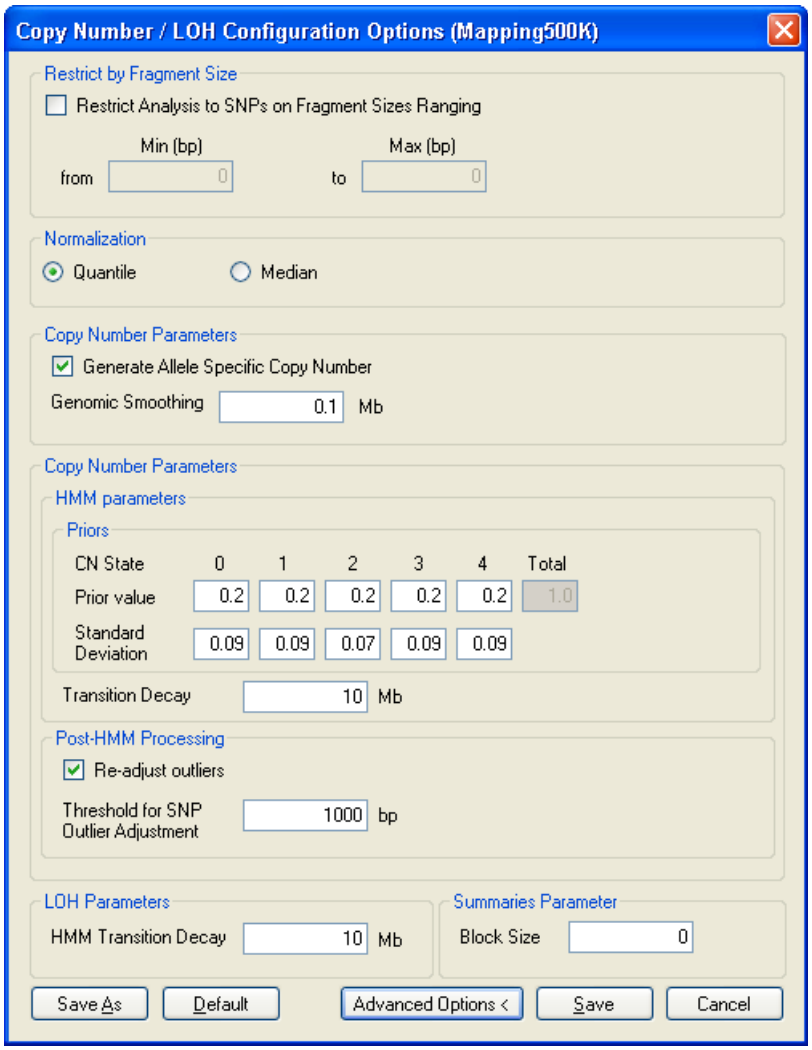
NOTE: The smoothing bandwidth should be determined based on the type of aberration in the sample. For example, if you are interested in small aberrations such as micro-deletions, you will want to use a smaller genomic smoothing length or no smoothing, comparable to or less than the size of the micro effect that is being studied. If you are looking for large chromosomal deletions, you may choose to use a large Mb smoothing bandwidth.

Advanced Options

Click the Advanced Options button to display the following options (Figure 12.50):

- HMM Parameters
- Post-HMM Processing
- LOH Parameters

Figure 12.50 Options DB with Advanced Options shown

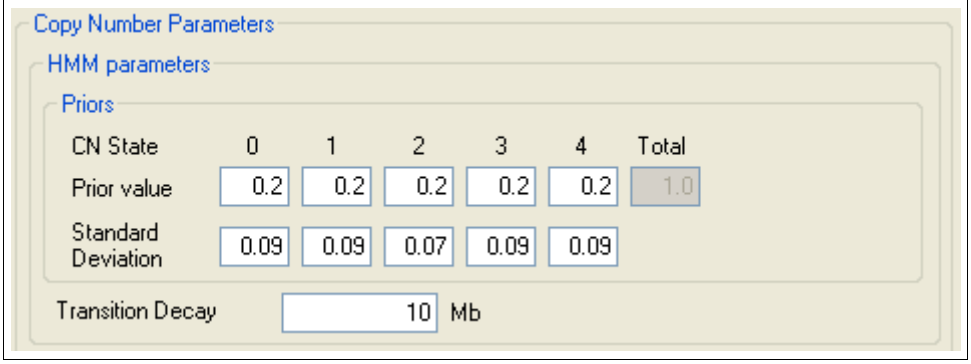


Copy Number Parameters/HMM Parameters

The Copy Number HMM Parameters adjust (Figure 12.51):

- CN State: Prior Value and Standard Deviation
- Transition Decay

Figure 12.51 HMM Parameters



CN State – Prior Value

A 5-state Hidden Markov Model (HMM) is applied for smoothing and segmenting the CN data.

The HMM has 5 possible states:

State 0 = CN of 0; homozygous deletion

State 1 = CN of 1; heterozygous deletion

State 2 = CN of 2; normal diploid

State 3 = CN of 3; single copy gain

State 4 = CN of 4; amplification

The default for each state is 0.2 indicating that each SNP has equal prior probability of being in any one of the 5 states. Generally speaking, the prior should not be adjusted unless it is known that the bulk of the data is comprised of hemizygous deletions. In this case, the prior corresponding to State 1 can be changed from 0.2 to 0.96 with all other prior states adjusted accordingly to equal a total of 1.

NOTE: The prior values entered are only initial estimates. The HMM optimizes this parameter based on the data.

Standard Deviation

Standard deviation is one of the parameters that affect the probability with which the underlying CN state is emitted to produce the observed state. Specifically, it reflects the underlying variance or dispersion in each CN state. The standard deviation of each underlying state can be adjusted. As a rule of thumb, the lower the Genomic Smoothing value, a higher standard deviation should be used for each CN state. This basically implies that with increased noise (due to less smoothing) the variance of the CN states should be increased.

The default is 0.07 for state 2 and 0.09 for all other states (0, 1, 3, 4). (See *Copy Number Parameter Settings* on page 240 for suggested changes to this parameter).

Transition Decay

This parameter controls the expected correlation between adjacent SNPs. The copy number state of any given SNP is partially dependent on that of its neighboring SNPs and is weighted based on the distance between them. By adjusting this parameter, neighboring SNPs can either have more or less of a dependence on each other.

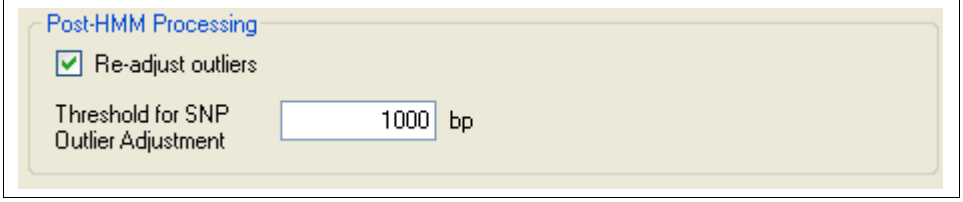
The default value is 10 Mb.

To reduce the influence of neighboring SNPs, decrease this value (transition faster).
For example, if you set the decay to 1 Mb, and if a given SNP is in CN State 1, the probability that the flanking SNPs to the right will continue to be in State 1 is much lower compared to the case where the transition decay is 100 Mb.

To increase the influence of neighboring SNPs, increase this value (transition slower).

Post-HMM Processing

Figure 12.52 Re-adjusting outliers



Re-adjust outliers

This parameter enables adjusting the CN state of singleton SNPs in a different state in comparison to the states of the flanking SNPs.

For example, if there is a single SNP in a 1 Mb region that is called CN State 3 by the HMM, but all surrounding SNPs are called CN State 2, then by checking the Re-adjust outliers checkbox, this singleton SNP will be changed from CN State 3 to CN State 2, provided it is within the threshold for SNP outlier adjustment. See Threshold for SNP Outlier Adjustment

If the surrounding states of the singleton SNP are two different states, the algorithm computes a weight median to determine which state to assign to the singleton SNP.

NOTE: Weighting of the median is determined by the distance to the flanking SNPs.

Threshold for SNP Outlier Adjustment

This parameter is linked to the re-adjust outliers parameter. It is the distance that is applied to determine if the flanking SNPs should impact the readjustment of the singleton SNP.

The default value is 1000 bp (the singleton SNP is in the center of this region).

NOTE: These parameters are highly correlated with the Gaussian smoothing used. If heavily smoothed (for example, >1Mb), the readjustment should be turned off. If the readjustment is enabled at the default threshold distance, it may not have any effect.

The readjustment parameter should be disabled for detection of micro-aberrations.

Suggested Cytogenetics Settings for Human Mapping 100K/500K Arrays

You may wish to save the HMM Parameters settings when performing cytogenetic analysis. Suggested values are:

Table 12.3 Recommended copy number parameter settings

CN State	Prior Value	Standard Deviation
0	0.2	0.23
1	0.2	0.23
2	0.2	0.2
3	0.2	0.23

Copy Number Parameter Settings

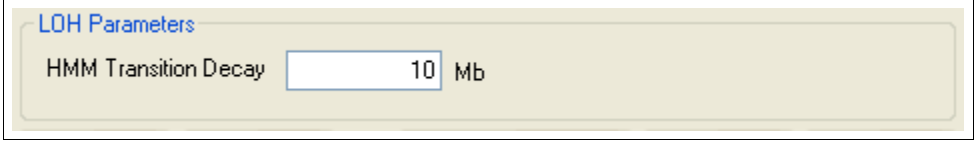
Analysis can be optimized to the specific copy number experiment by changing the algorithm parameters. The table below describes a set of recommended parameter settings for some common experimental conditions.

Table 12.4 Recommended copy number parameter settings

Copy Number	Footprint of Change	Restrict by Fragment Size	Ref. Set	Probe-level normalization	Gaussian Smoothing (kb)	HMM Priors	HMM Transition Decay (Mb)	HMM Std. Deviation	Adjust Outliers
Micro-deletions	<4Mb		Unpaired ≥ 25	Median Scaling	Low	Equal	≤1000	Refer to BW versus SD table (algorithm in manual)	off
Chr X changes	Size of chr X		Unpaired ≥ 25	Quantile	100	Equal	1000	0.09 for states 0, 1, 3, 4 & 0.07 for state 2	on
Trisomy/ Disomy	Variable		Unpaired ≥ 25	Quantile	100	Equal	1000	0.09 for states 0, 1, 3, 4 & 0.07 for state 2	on
Tumor-Normal pairs	Variable		1	Median/Quantile	100	Equal	1	0.09 for states 0, 1, 3, 4 & 0.07 for state 2	on
Homozy-gous deletions	Variable		Unpaired ≥ 25	Quantile	100	State 0=0.96 All other states = 0.01	10	0.09 for states 0, 1, 3, 4 & 0.07 for state 2	on
Pseudo-autosomal regions on X (Male)	"95 SNPs (Nsp) "140 SNPs (Sty)		Unpaired ≥ 25	Quantile	500	Equal	10	0.06 for states 0, 1, 3, 4 & 0.03 for state 2	on
Karyotype	1–5 Mb		Unpaired ≥ 25	Quantile	50	Equal	1	0.11 for states 0,1,3,4 & 0.08 for state 2	on
FISH (BAC clones)	200 Kb		Unpaired ≥ 25	Quantile	50	Equal	1	0.11 for states 0,1,3,4 & 0.08 for state 2	on
Analysis of FFPE samples	Variable	(exclude SNPs on larger PCR fragments)	Unpaired ≥ 30	Quantile	100	Equal	1-100	0.09 for states 0,1,3,4 & 0.07 for state 2	on

LOH Parameters

Figure 12.53 LOH Parameters - Advanced Options



Analysis can be optimized to the specific LOH experiment by changing the algorithm parameters. describes a set of recommended parameter settings for some common experimental conditions.

Table 12.5 Recommended LOH parameters

LOH	Reference Set	HMM Transition Decay (Mb)
Tumor – Normal Pairs	1	10
Unpaired	>30 from mixed population	10
Unpaired	~30 from same population	10

Copy Number & LOH Analysis for Genome-Wide Human SNP 6.0 Arrays

GTC 4.2 can be used to perform the following analyses for the Genome-Wide Human SNP Array 6.0:

- Copy Number (CN)
- Loss of Heterozygosity (LOH)

The following analyses are performed on the CN data generated during CN/LOH analysis:

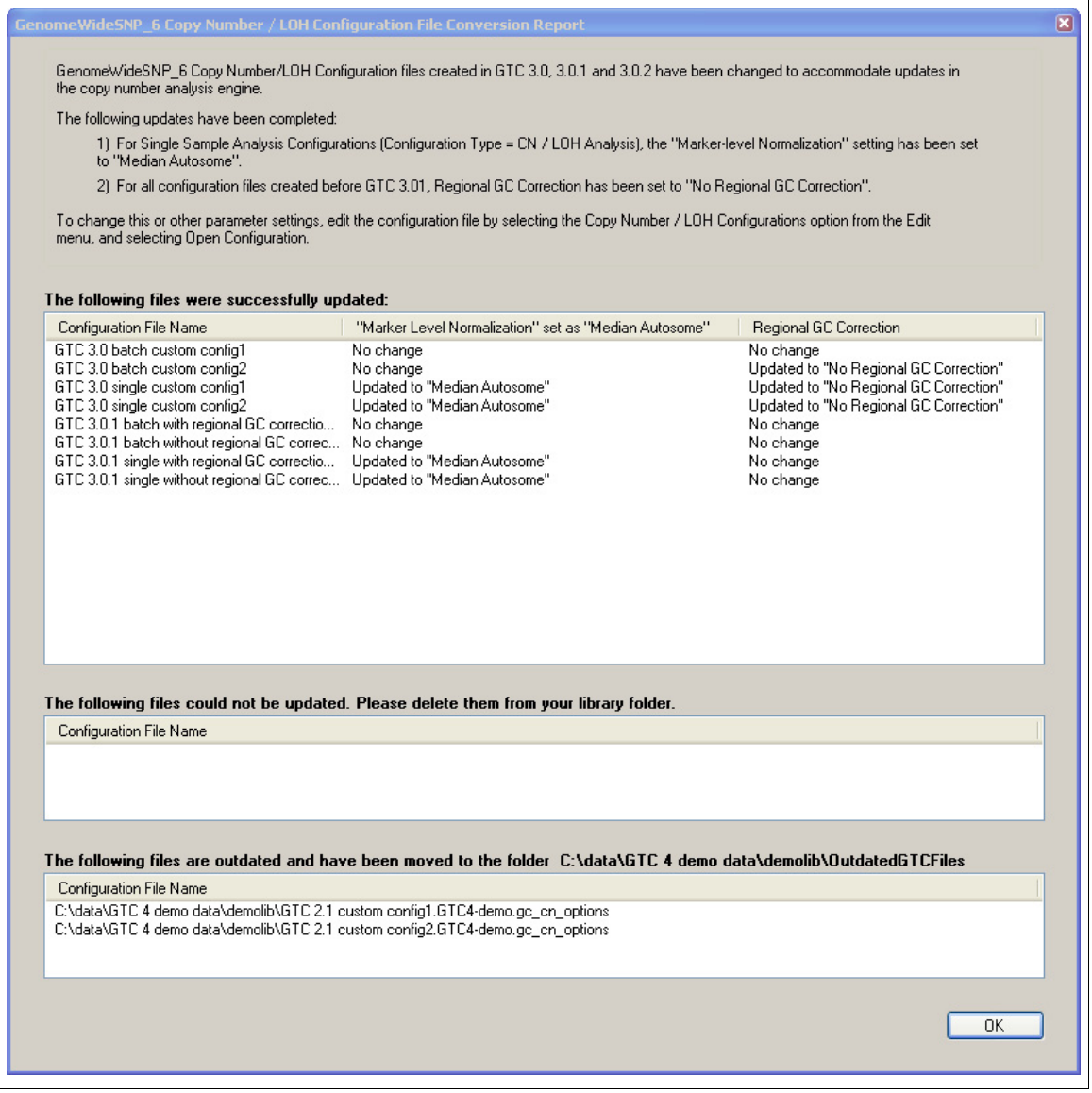
- Copy Number Segment Reporting
- Custom Region Copy Number Segment Reporting

NOTE: **Copy Number Variation (CNV) analysis is performed in a separate step from CN/LOH analysis. The CNV data can be viewed in the Heat Map with the CN data. See Chapter 15, *Copy Number Variation Analysis* on page 315 for more information.**

GTC 4.2 provides an updated default CN configuration file to accommodate updates in CN analysis. For the configuration type CN/LOH Analysis, the Marker-level Normalization option is set to Median Autosome in the default configuration file. You can manually change the Marker-level Normalization option by editing the configuration file (for more details, see *Changing CN/LOH Algorithm Configurations for SNP 6.0 Analysis* on page 266).

CN configuration files from GTC 3.0, 3.0.1, and 3.0.2 are automatically updated when GTC 4.2 launches, or when a new user profile is selected, or when the library path is changed. Configuration files from GTC 2.0 or 2.1 are not updated. The updates are listed in a Conversion Report (Figure 13.1).

Figure 13.1 CN configuration update



Only SNP 6.0 CEL files are needed for analysis by BRLMM-P+; genotyping (CHP) files are not required.

IMPORTANT: Copy Number and LOH analysis algorithms performed on SNP 6.0 array data are collectively referred to in Genotyping Console as "CN5" in output file names.

NOTE: GTC 4.2 does not perform copy number, LOH, or Copy number region analysis on data from SNP 5.0 and Axiom Array types, however the included Axion CNV Tool application does. See *Axiom CNV Summary Tool and Viewer* on page 364 for more information.

IMPORTANT: Affymetrix recommends that you perform Copy Number/LOH analysis with all files stored locally. For more details on hard disk space requirements, see Appendix J, *Hard Disk Requirements* on page 363.

IMPORTANT: Affymetrix recommends that you perform Copy Number/LOH analysis with regional GC correction configuration.

The basic workflow for Copy Number/LOH analysis for SNP 6.0 arrays involves:

1. *Performing Copy Number/LOH analysis on a selection of CEL files* (page 244).
   There are two options for this analysis:
   - *CN/LOH Reference Model File Creation and Analysis (Batch Sample Mode)* on page 245.
   - *CN/LOH Analysis with a Previously Created Reference Model File (Single Sample Mode)* on page 253.
2. *Performing the Copy Number Segment analysis on the SNP 6.0 CN data files* (page 285).

   NOTE: Segment Reporting Analysis can also be performed on 100K/500K data.

   For SNP 6.0 data, the Segment Report also provides gender calls, including reports for samples with unknown (or ambiguous) genders.
3. *Viewing QC data in table format* (page 261).
4. *Viewing the CN/LOH data in the GTC Browser* (page 305).
5. *Viewing the Copy Number and Copy Number Variation (CNV) data in the Heat Map Viewer* (page 323).

   NOTE: CNV analysis is performed in a separate step from CN/LOH analysis. The CNV data can be viewed in the Heat Map with the CN data. See Chapter 15, *Copy Number Variation Analysis* on page 315 for more information.
6. *Exporting data into formats that can be used by secondary analysis software* (page 307).
7. You can also:
   - *Change the QC threshold settings* (page 312).
   - *Change the algorithm parameters for SNP 6.0 analysis* (page 266).

NOTE: Small numerical differences may occur between different runs even with the same inputs due to an interaction between rounding from double to single precision and the way the application handles memory management.

Copy Number/LOH Analysis for SNP 6.0 Arrays

NOTE: Affymetrix recommends that you perform Copy Number/LOH analysis with all files stored locally.

The CN/LOH analysis for SNP 6.0 arrays outputs files with the extension CN5.cnchp; these files contain both copy number and LOH data.

The following types of analysis can be performed:

- *CN/LOH Reference Model File Creation and Analysis (Batch Sample Mode)* on page 245

  This analysis first creates a Reference Model file using the CEL files for the selected samples. Then each CEL file used to create this Reference Model file is re-analyzed against the new Reference Model file. From this comparison, the sample's Copy Number and LOH data are generated. The genotype calls made on the fly by the BRLMM-P+ algorithm are used for the LOH analysis.

  The analysis provides Gender Calls — Female or Male.

- *CN/LOH Analysis with a Previously Created Reference Model File (Single Sample Mode)* on page 253

  In this analysis you compare the selected sample CEL files to a previously created Reference Model file, either the HapMap270 file supplied by Affymetrix or a Reference Model file you have created using the CN/LOH Reference Model File Creation and Analysis process described above. In this "Single sample" workflow, the LOH analysis is done with the genotype calls made on the fly by the BRLMM-P+ algorithm using the Reference Model data.

  The analysis provides Gender Calls — Female or Male.

NOTE: Small numerical differences may occur between different runs even with the same inputs due to an interaction between rounding from double to single precision and the way the application handles memory management.

NOTE: CN/LOH analysis can be run either with regional GC correction or without regional GC correction. Either configuration works with both batch sample mode and single sample mode.

NOTE: Previous GTC 3.0 configuration files will automatically be updated by GTC 4.2 and run without GC correction and with updated score threshold (1.0) and configurable Marker-level Normalization.

NOTE: Analysis performed with regional GC correction will need NetAffx NA26.1 or higher version of annotation files. Analysis performed without regional GC correction will need NetAffx NA25 or higher version of annotation files.

Copy Number and LOH analyses are done during the same analysis run and the data are kept in the same CN5.cnchp file.

CN/LOH Reference Model File Creation and Analysis (Batch Sample Mode)

IMPORTANT: Affymetrix recommends that you perform Copy Number/LOH analysis with regional GC correction configuration.

IMPORTANT: Affymetrix recommends that you run Copy Number/LOH analysis with batch sample mode and regional GC correction using arrays run at the same lab using the same reagent lots to reduce general variability and to correct GC waviness.

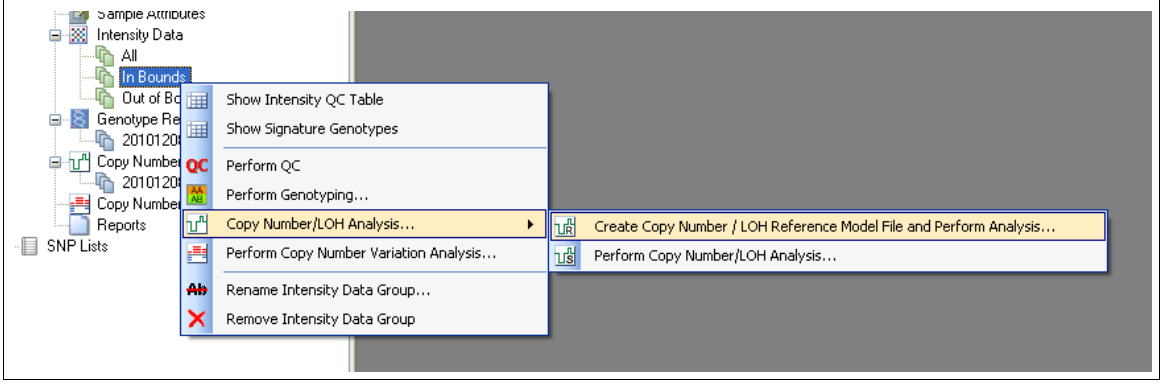
IMPORTANT: See Appendix J, *Hard Disk Requirements* on page 363 for more details on hard disk space requirements.

This analysis first creates a Reference Model file using the CEL files for the selected samples. Then each CEL file used to create this Reference Model file is re-analyzed against the new Reference Model file. From this comparison, the sample's Copy Number and LOH data are generated. The genotype calls made on the fly by the BRLMM-P+ algorithm are used for the LOH analysis. The Reference Model Files end in the filename extension .ref.

To create a Reference Model File and perform SNP 6.0 CN/LOH analysis:

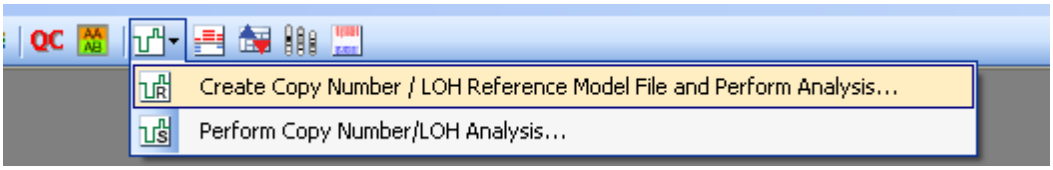
1. Select the Intensity Data file set from the Data tree.
2. Do one of the following:
   - From the Workspace menu, select **Intensity Data > Create Copy Number/LOH Reference Model File and Perform Analysis**; or
   - Right-click on the Intensity Data file set in the data tree and select **Copy Number/LOH Analysis > Create Copy Number/LOH Reference Model File and Perform Analysis** from the pop-up menu (Figure 13.2); or

Figure 13.2 Selecting CN/LOH Analysis from the data tree



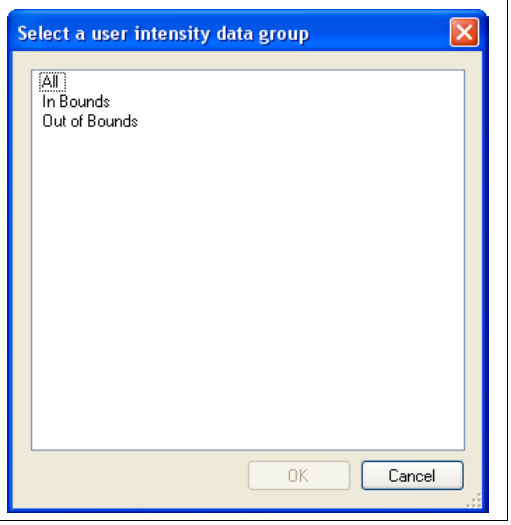
   - Click the **Create Copy Number/LOH Analysis** button in the tool bar and select **Create Copy Number/LOH Reference Model File and Perform Analysis…** from the menu (Figure 13.3).

Figure 13.3 Selecting CN/LOH analysis from the tool bar



If you have not selected a particular Intensity Data set, the Select a user intensity data group dialog box opens (Figure 13.4).

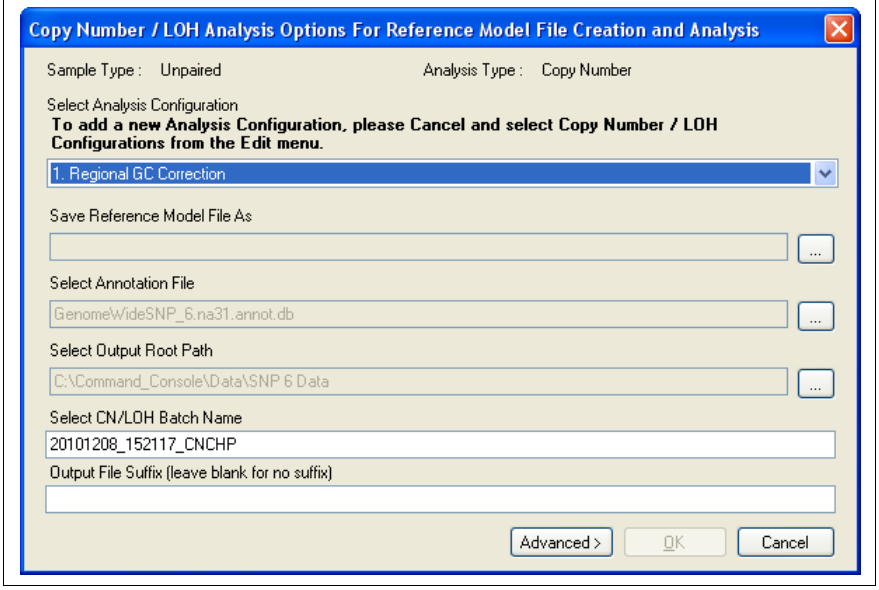
Figure 13.4 Copy Number Analysis Options dialog box



3. Select a data group and click **OK** in the Select a user intensity data group dialog box.

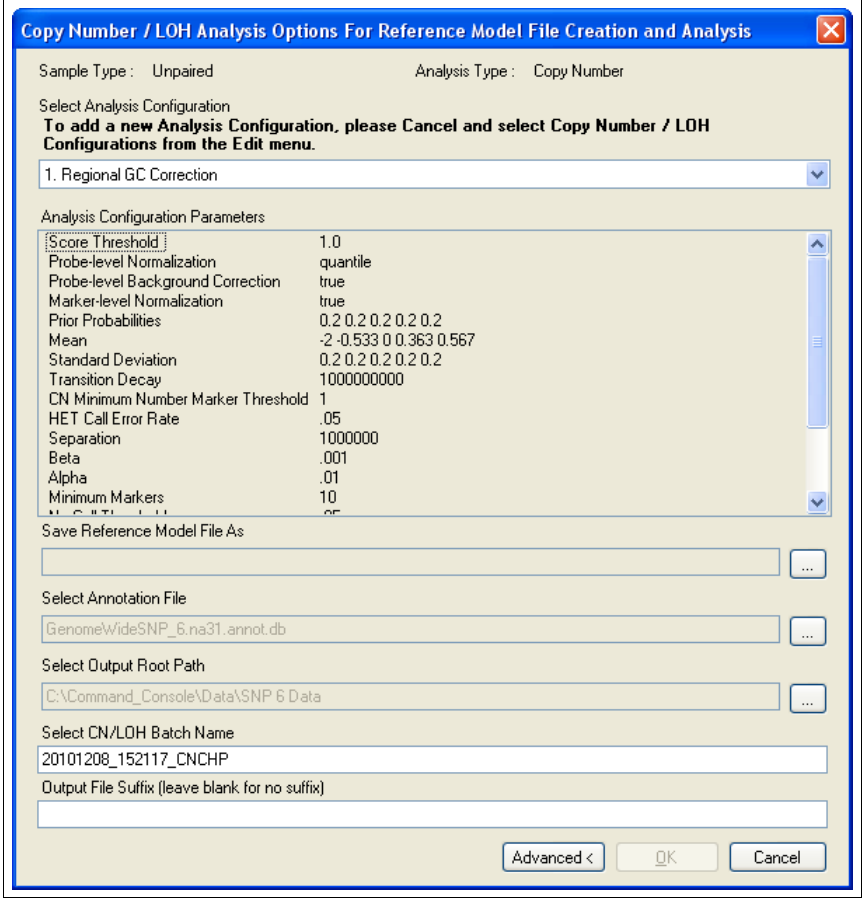
   The Copy Number/LOH Analysis Options for Reference Model File Creation and Analysis dialog box opens (Figure 13.5).

Figure 13.5 Copy Number Analysis/LOH Analysis Options for Reference Model File Creation and Analysis dialog box



Click the **Advanced** button to review analysis configuration parameters (Figure 13.6).

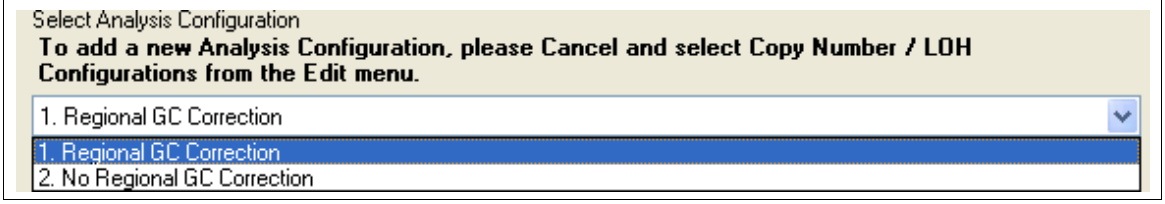
Figure 13.6 Displaying the Analysis Configuration Parameters



4. Select a different Analysis Configuration (Figure 13.7).

   Analysis configurations are sets of parameters used in the analysis. See *Changing CN/LOH Algorithm Configurations for SNP 6.0 Analysis* on page 266 for more information on creating a new analysis configuration.

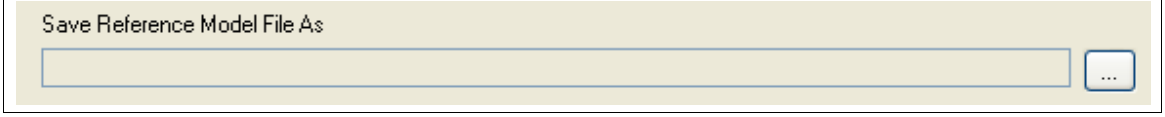
Figure 13.7 Copy Number Analysis Options dialog box, Configuration drop-down list



   - Select a different configuration from the drop-down list.

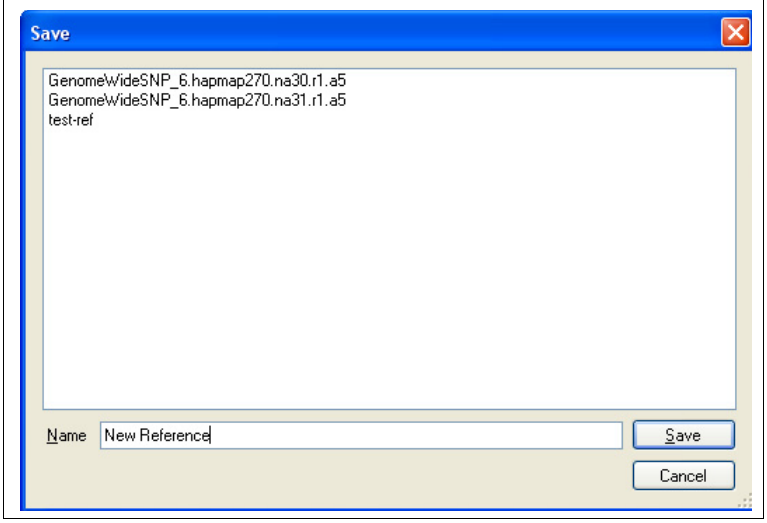
5. Enter a name for the new Reference File (Figure 13.8):

Figure 13.8 Save Reference Model File As



A. Click the Save Reference Model File As browse button ... .
The Save dialog box opens (Figure 13.9).

Figure 13.9 Save dialog box



B. Enter a name for the file in the Name box.

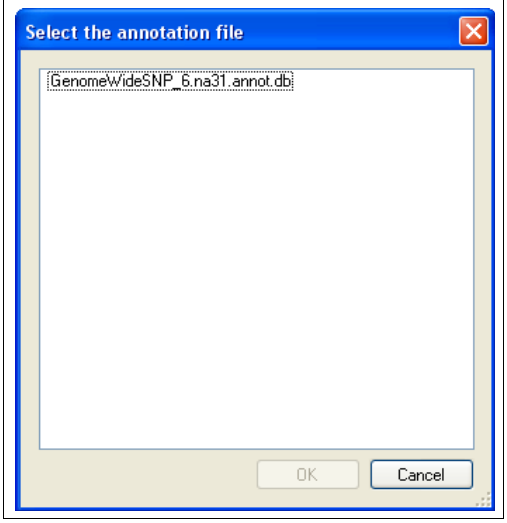
C. Click **Save** in the Save dialog box.

6. Select a different annotation file (Figure 13.10) (optional).
This option enables you to select an annotation file for the analysis.

Figure 13.10 Select Annotation File

A. Click the Select Annotation File browse button ... .
The Select the annotation file dialog box (Figure 13.11) opens.

Figure 13.11 Select annotation file dialog box



NOTE: The NetAffx annotation file must be of NA26.1 or higher version if configuration files are with regional GC correction. If the configuration files are without regional GC correction, the NetAffx annotation file can be of NA25 or higher version.

NOTE: Only official released SNP6 NetAffx annotation files are filtered in this dialogue window. Annotations for other array platforms or custom annotation files will not be filtered.

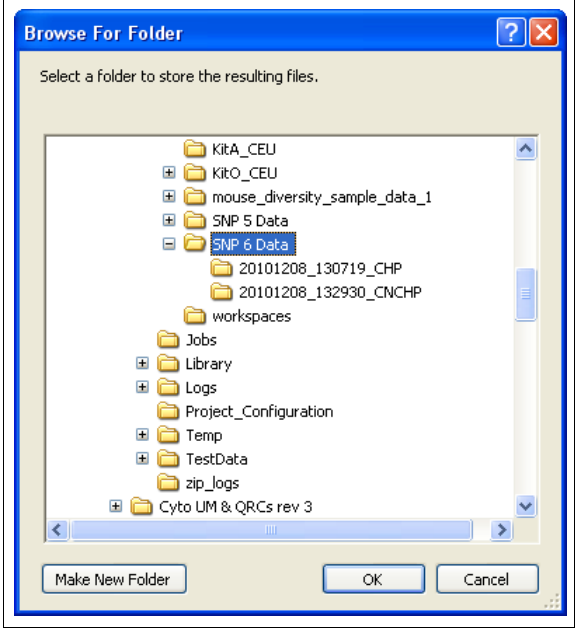
B. Click **OK** in the Select Annotation File dialog box.

7. Select Output Root path (Figure 13.12) (optional):
   This option changes the location where the CN/LOH files are placed.

Figure 13.12 Select Output Root Path

A. Click the Select Output Root Path browse button ... .
   The Browse for Folder dialog box opens (Figure 13.13).

Figure 13.13 Browse for Folder dialog box



B. Select a new location for the CN/LOH data files and click **OK** in the Browse for Folder dialog box.

8. Select CN/LOH Batch Name (Figure 13.14) (optional):
   This option changes the name of the folder in which the CNCHP files are placed. A name based on the analysis type and the time and date of the analysis is automatically assigned to the folder unless you change it.

Figure 13.14 Select CN/LOH Batch Name

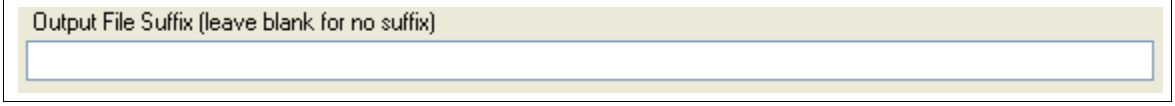
- Click in the box and enter the Batch Name.

NOTE: This is the name of the folder where the different Data Results files are kept. To view report files, access the folder through Windows Explorer.

9. Enter File Suffix for the CNCHP files (Figure 13.15) (optional):

   This option adds a suffix to the CNCHP files to help you track them. Click in the box and enter a suffix.

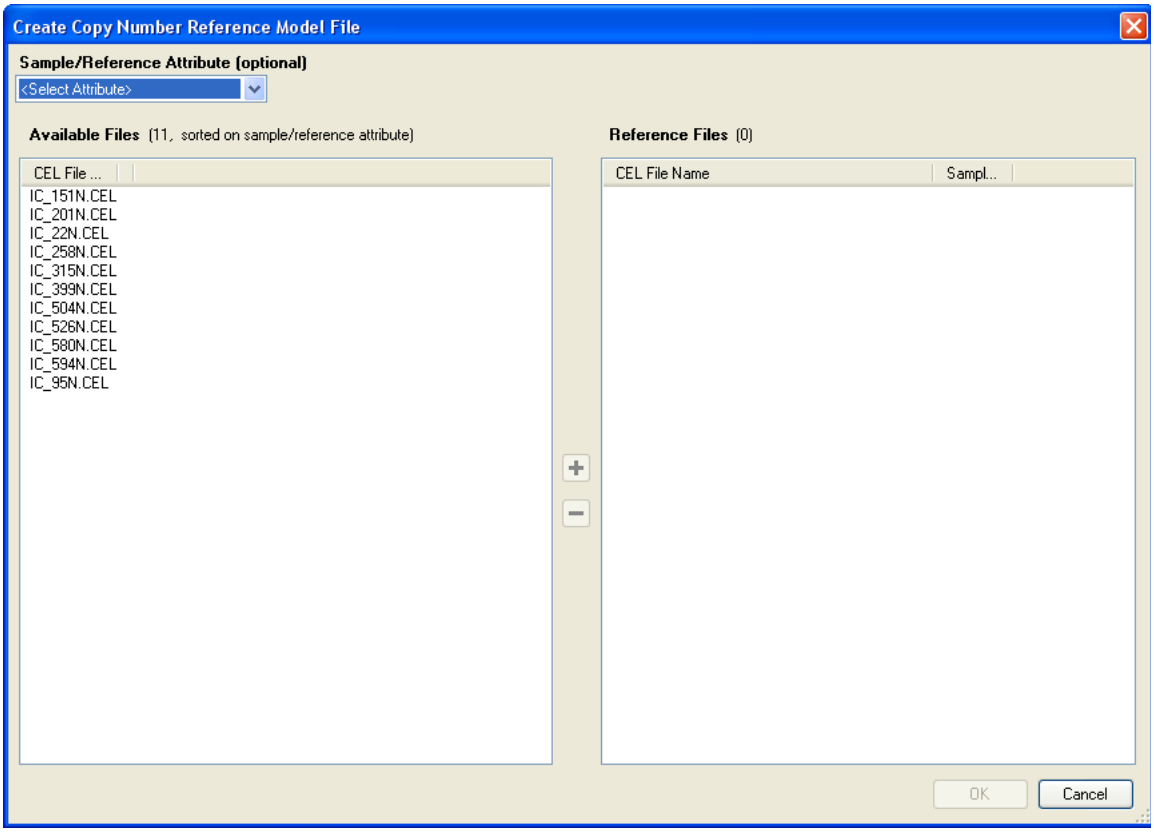
Figure 13.15 Output File Suffix



10. Click **OK** in the Copy Number/LOH Analysis Options for Reference Model File Creation and Analysis dialog box.

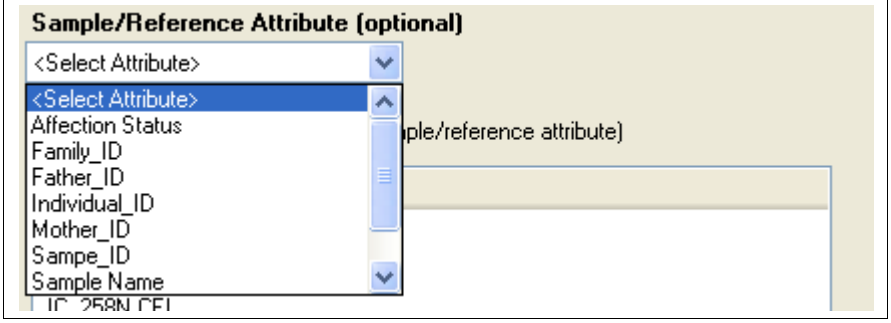
    The Create Copy Number Reference Model File dialog box opens (Figure 13.16).

Figure 13.16 Create Copy Number Reference Model File dialog box



The Sample/Reference Attribute (optional) dropdown list (Figure 13.17) enables you to sort the CEL files by an attribute in the corresponding Sample (ARR) files.

Figure 13.17 Sample/Reference Attribute (optional) dropdown list



11. Select files in the Available Files list. A minimum of five files is required to run the analysis.

NOTE: To create a useful Reference Model File, it is recommended that you select 44 or more samples if possible, although the software will accept as few as 5. For obtaining good data on the X and Y chromosomes, you should use a minimum of 15 files from female samples and 15 files from male samples to generate the Reference Model file. See *Notes on Selecting Files for Creating Reference Model Files* on page 253 for more information.

NOTE: If all other parameters and files are the same, reference model files generated with or without regional GC correction are exactly the same. You do not have to regenerate a reference model file twice with different settings in the regional GC correction option.

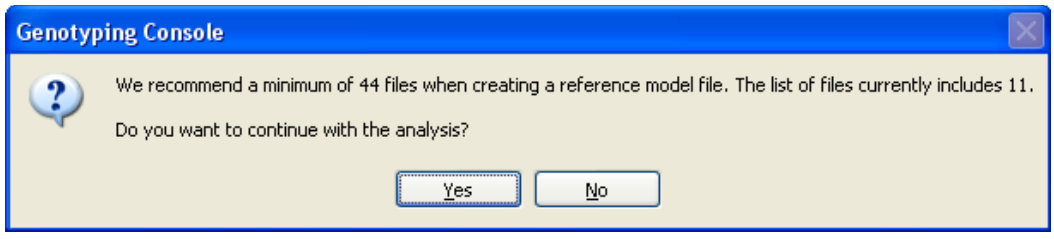
Click the **Add** button to add data to the Reference list.

Click the **Remove** button to remove data from the Reference list.

12. Click **OK**.

If you have selected fewer than the recommended number of samples, a warning appears (Figure 13.18).

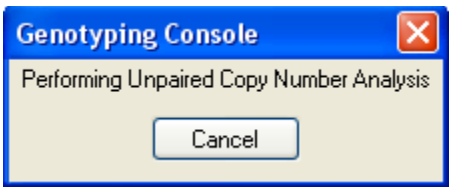
Figure 13.18 File number warning



Click **Yes** to proceed with the analysis.

A notice shows that the analysis is in progress (Figure 13.19).

Figure 13.19 Performing Unpaired Copy Number Analysis notice



After generating the Copy Number/LOH (CN5.cnchp) files, you can:

- *View CN QC data in tables* (page 261).
- *Use the new Reference Model file to perform additional single-sample CN/LOH analysis* (below).
- *Generate Copy Number Segment Reports* (page 285).
- *View the data in the GTC Browser* (page 305)
- *View the CN data in the Heat Map viewer* (page 242)
- *Export data to other software* (page 307)

Notes on Selecting Files for Creating Reference Model Files

Affymetrix recommends using a minimum of 44 samples when creating a Reference Model File. A minimum of five files is required to generate the reference model file.

Affymetrix recommends using a set of mixed gender samples when creating a Reference Model File for analysis.

Affymetrix recommends using at least 15 female samples when creating a Reference Model File for analysis of the X chromosome.

Affymetrix recommends using at least 15 male samples when creating a Reference Model File for analysis of the Y chromosome.

CN/LOH Analysis with a Previously Created Reference Model File (Single Sample Mode)

IMPORTANT: Affymetrix recommends that you perform Copy Number/LOH analysis with regional GC correction.

IMPORTANT: Affymetrix recommends that you run Copy Number/LOH analysis with batch sample mode and regional GC correction using arrays run at the same lab using the same reagent lots to reduce general variability and to correct GC waviness.

IMPORTANT: Affymetrix recommends that you perform Copy Number/LOH analysis with all files stored locally. For more details on hard disk space requirements, see **Appendix J, *Hard Disk Requirements* on page 363**.

In this analysis you compare the selected sample CEL files to a previously created Reference Model file, either the HapMap270 one supplied by Affymetrix or a reference you have created using the CN/LOH Reference Model File Creation and Analysis process described above. In this workflow no CHP files are required; instead the LOH analysis is done with the genotype calls made on the fly by the BRLMM-P+ algorithm.

NOTE: You can perform a single sample analysis on more than one CEL file at a time; single sample means that each CEL file is compared to a reference model file.

Notes on Selecting Files for Analysis against a Previously Created Reference Model File

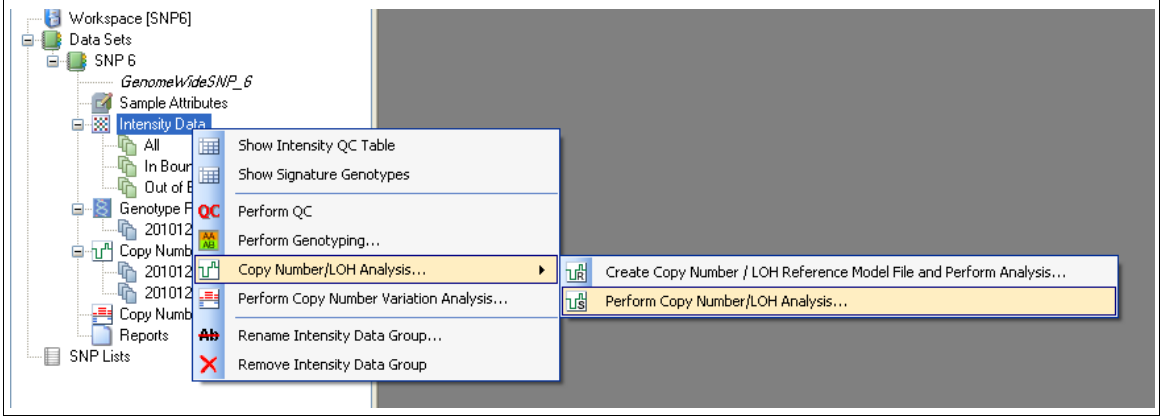
Affymetrix recommends not analyzing only female samples against a Reference Model File previously generated with only male samples when running CN/LOH analysis.

See Appendix A, *Algorithms* on page 343 for references to the BRLMM-P+ algorithm.

To perform CN/LOH analysis with a previously created Reference Model File:

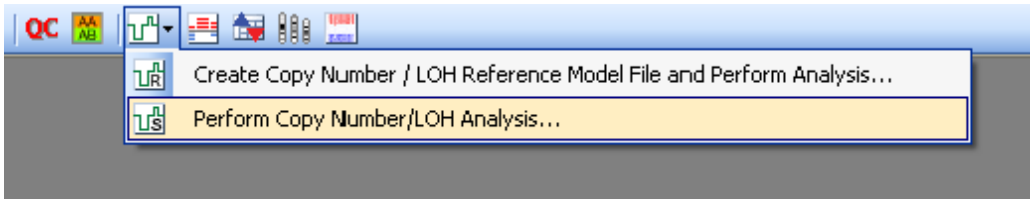
1. Select the Intensity Data file set.
2. Do one of the following:
   - From the Workspace menu, select **Intensity Data > Perform Copy Number/LOH Analysis**; or
   - Right-click on the Intensity Data file set in the data tree and select **Copy Number/LOH Analysis > Perform Copy Number/LOH Analysis** from the pop-up menu (Figure 13.20); or

Figure 13.20 Selecting Copy Number/LOH Analysis from the data tree



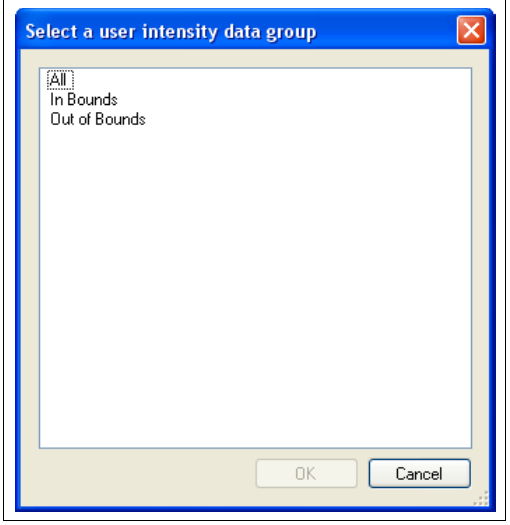
   - Click the **Create Copy Number/LOH Analysis** button in the tool bar and select **Perform Copy Number/LOH Analysis …** from the menu (Figure 13.21).

Figure 13.21 Selecting Copy Number/LOH Analysis from the tool bar



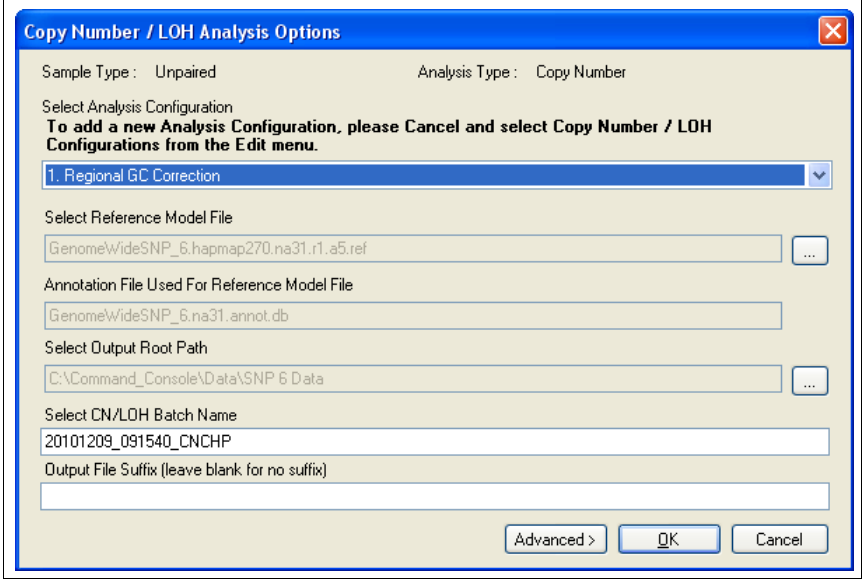
If you have not selected a particular Intensity Data set, the Select a user intensity data group dialog box opens (Figure 13.22).

Figure 13.22 Select a user intensity data group dialog box



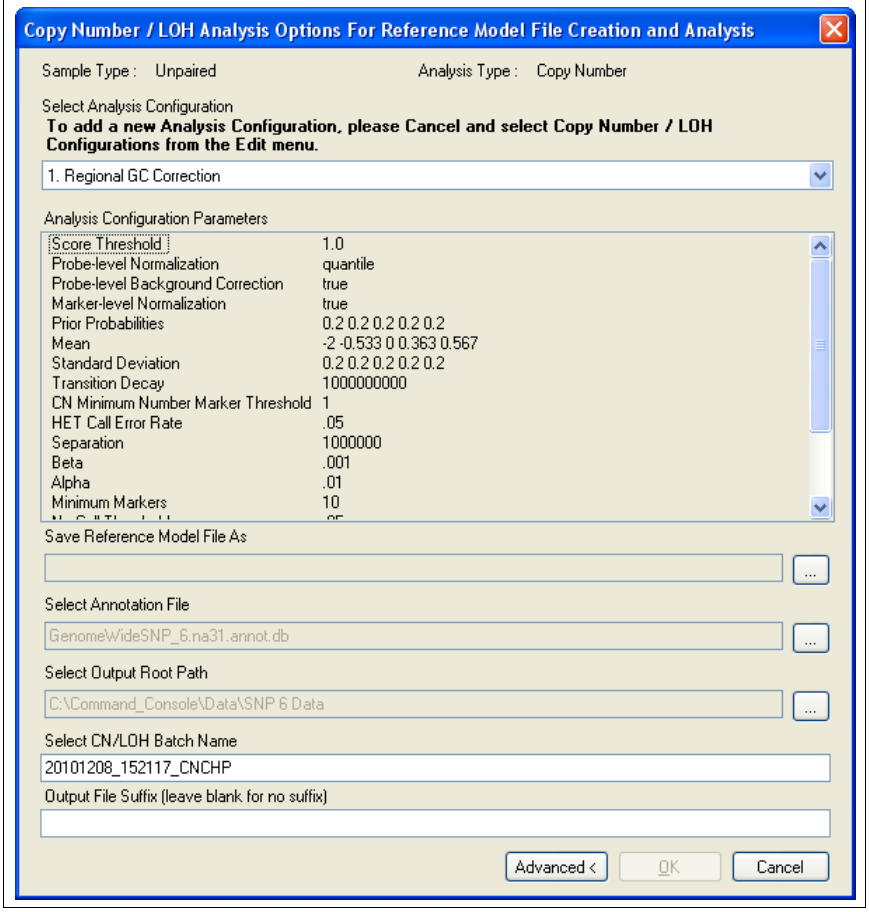
3. Select a data group and click **OK** in the Select a user intensity data group dialog box. The Copy Number Analysis Options dialog box opens (Figure 13.23).

Figure 13.23 Copy Number/LOH Analysis Options dialog box



Click the **Advanced** button to review analysis configuration parameters (Figure 13.24).

Figure 13.24 Displaying the Analysis Configuration Parameters



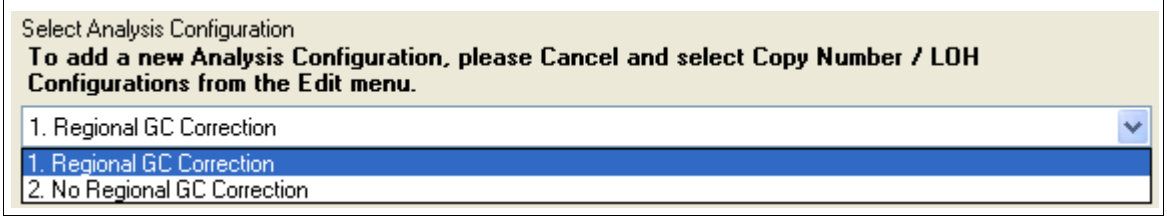
NOTE: You cannot change the annotation files in this analysis once a specific reference model file is chosen. The annotation files used to create the Reference file are automatically selected.

4. Select a different Analysis Configuration without regional GC correction or any other custom configuration files (Figure 13.25).

   Analysis configurations are sets of parameters used in the analysis. See *Changing CN/LOH Algorithm Configurations for SNP 6.0 Analysis* on page 266 for more information on creating a new analysis configuration.

NOTE: For some parameters, you cannot select different values than those used in the generation of the reference file used for the analysis.

Figure 13.25 Copy Number/LOH Analysis Options, configurations drop-down list



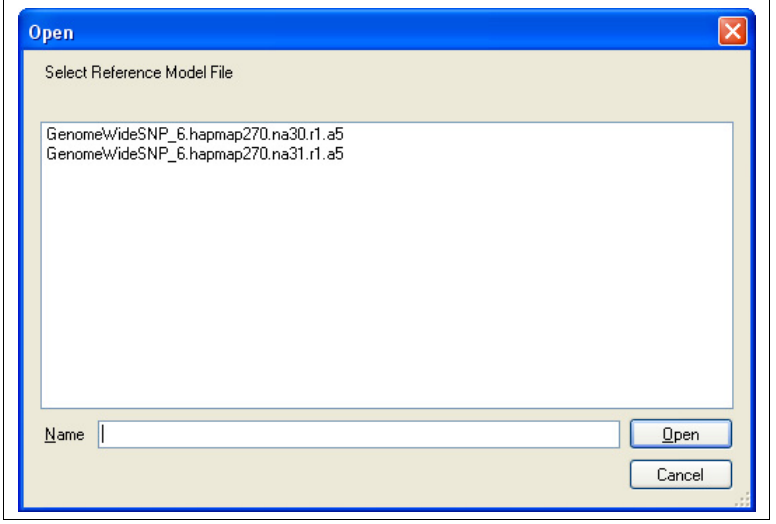
- Select a different configuration from the drop-down list.

5. Select a Reference File for the analysis:

Figure 13.26 Select Reference Model File box

A. Click the Select Reference Model File As browse button ... (Figure 13.26). The Select Reference Model file dialog box opens (Figure 13.27).

Figure 13.27 Select Reference Model File dialog box



B. Select a reference file from the list and click **Open** in the Select Reference Model File dialog box. The correct annotation file is automatically selected (Figure 13.28).

Figure 13.28 Annotation File Used for Reference Model File

NOTE: If you choose the GenomeWideSNP_6.hapmap270.na26.1.r1.a5.ref as reference model file, you are required to have NetAffx NA26.1 version of annotation files.

NOTE: You cannot change the annotation files in this analysis once a specific reference model file is chosen. The annotation files used to create the Reference file are automatically selected.

NOTE: The NetAffx annotation files must be of NA26.1 or higher version if configuration files are with regional GC correction. If the configuration files are without regional GC correction, the NetAffx annotation files can be of NA25 or higher version.

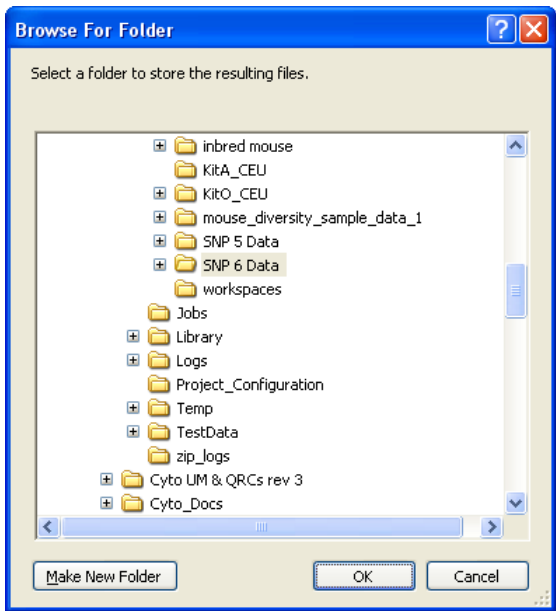
6. Select Output Root path (Figure 13.29) (optional).

Figure 13.29 Select Output Root Path

This option changes the location where the CN/LOH files are placed.

A. Click the Select Output Root Path browse button ... .
The Browse for Folder dialog box appears (Figure 13.30).

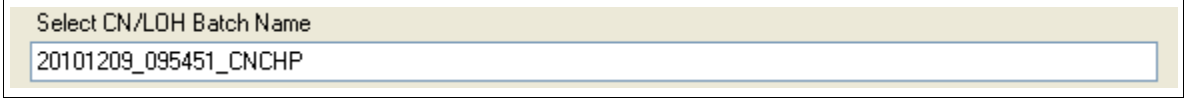
Figure 13.30 Browse for Folder dialog box



B. Select a new location for the CN/LOH data files and click **OK** in the Browse for Folder dialog box.

7. Select CN/LOH Batch Name (Figure 13.31):

Figure 13.31 Select CN/LOH Batch Name



This option changes the name of the folder in which the CNCHP files are placed.

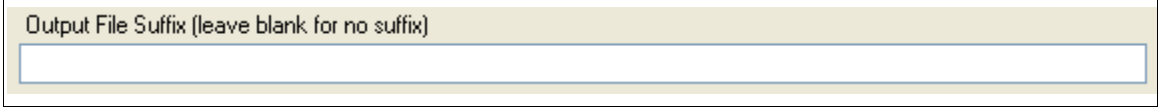
A name based on the analysis type and the time and date of the analysis is automatically assigned to the folder unless you change it.

NOTE: This folder is the location where the different Data Results files are kept. You can access the folder through Windows Explorer to view report files.

- Click in the box and enter the Batch Name.

8. Enter a File Suffix for the CNCHP files (Figure 13.32):

Figure 13.32 Output File Suffix

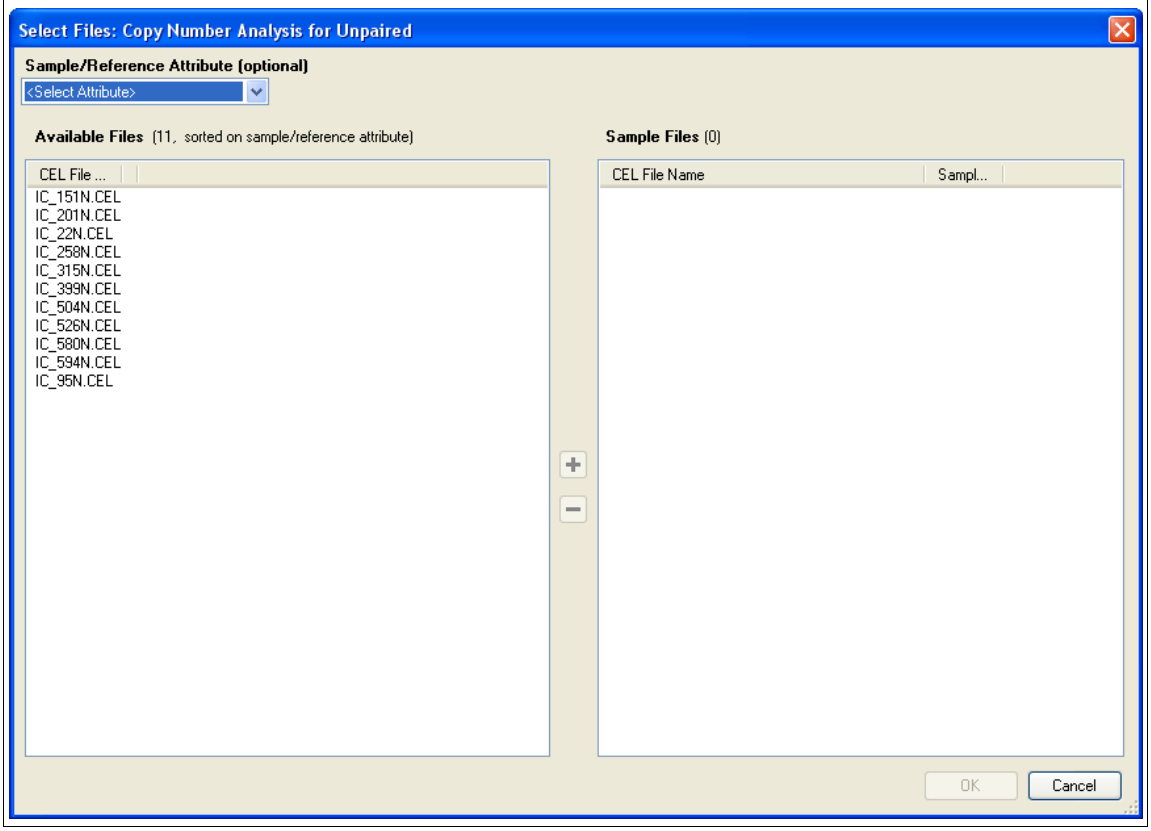


This option adds a suffix to the CNCHP files to help you track them.

- Click in the Box and enter a suffix.

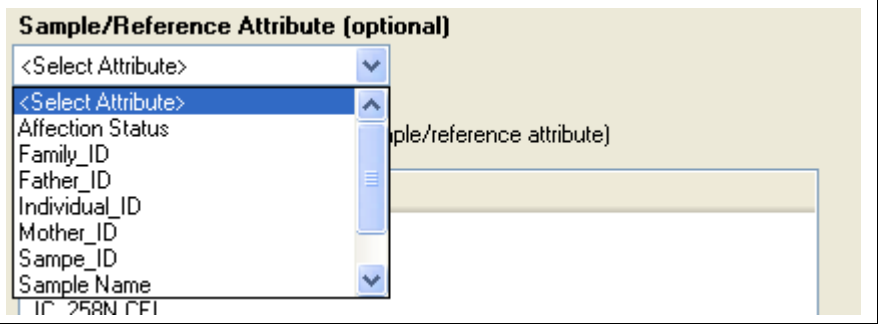
9. Click **OK** in the Copy Number/LOH Analysis Options dialog box.
The Select Files dialog box opens (Figure 13.33).

Figure 13.33 Select Files: Copy Number Analysis for Unpaired dialog box



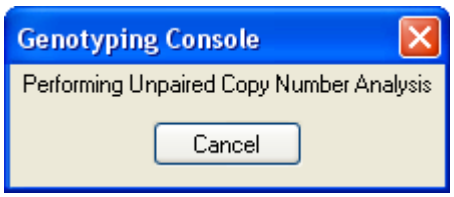
The Sample/Reference Attribute (optional) dropdown list (Figure 13.34) enables you to sort the CEL files by an attribute in the corresponding Sample (ARR) files.

Figure 13.34 Sample/Reference Attribute (optional) dropdown list



10. Select files in the Available Files list.
    Click the **Add** button to add data to the sample or reference list.
    Click the **Remove** button to remove data from a list.
11. Click **OK** in the Select Files: Copy Number Analysis for Unpaired dialog box.
    A notice shows that the analysis is in progress (Figure 13.35).

Figure 13.35 Performing Unpaired Copy Number Analysis notice



After generating the Copy Number/LOH (CN5.cnchp) files, you can:

- *View CN QC data in tables* (page 261).
- *Use the new Reference Model file to perform additional single-sample CN/LOH analysis* (below).
- *Generate Copy Number Segment Reports* (page 285).
- *View the data in the GTC Browser* (page 305)
- *View the CN data in the Heat Map viewer* (page 242)
- *Export data to other software* (page 307)

The file format is described below.

Copy Number/LOH Data File Format for Genome-Wide Human SNP Array 6.0 Data

For Genome-Wide Human SNP Array 6.0 analysis, the Copy Number and LOH data are kept in the same file.

Header Section

The header section contains the following information:

- Information about the Software and Algorithm version used to generate the data
- File name, creation and modification times, and unique identifier
- Array type
- Genome version and library information
- CN/LOH Algorithm parameters

- Reference Model File used
- Number of Markers for each chromosome
- X and Y chromosome information

Data Section – for *.CN5.cnchp files

The data section contains information on the following output fields found in *.CN5.cnchp files:

allele difference	Difference between the A channel signal and the B channel signal, with each signal standardized with respect to their median values in the reference
cnstate	Hidden Markov Model (HMM) copy number state
smoothsignal	Smoothed log2 ratios or smoothed log2 ratios calibrated to Copy Number and anti-logged (depending on the options setting)
loh	Loss of Heterozygosity, 1=LOH, and 0=retention
log2ratio	Log2 ratio value

Adjusting Normalization and Background Parameters for Reference Model File and Sample Files

The Copy Number algorithms depend on comparing signal for each marker in each sample against a reference formed from a group of samples. The underlying assumption is that for each marker the reference signal state in the group will be CN=2 (except for the Y chromosome, where the reference state is CN=1), and hence deviations from the reference signal can be seen by forming the log ratio of each marker's signal compared to its reference value. For the autosomes, the reference value for each probe set is formed by taking the median of summarized probe set signals across all samples in the reference. For each SNP probe set, summary signal is calculated after normalizing intensities by using probe logarithmic intensity error (PLIER) with non-standard options for each of the SNP allele probe sets and summing the result of both alleles. For each CN probe set, summary signal is the normalized intensity only. For chromosome X, the reference value is formed using only the samples determined not to have a single X and assumes the majority of such samples are diploid. For chromosome Y, the reference value is formed using only the samples determined to have a Y present.

NOTE: Forming a reference where a large fraction of the samples have one or more chromosomal aneuploidies in common will give you unreliable results for the chromosomes affected by aneuploidy.

In the process of calculating signal various normalization steps are made so that signal from each sample can be meaningfully compared with each other. If these normalization steps are not the same, then the comparison is no longer meaningful. In particular, in single sample workflow, new samples are normalized in the same way as the reference.

For information about changing the algorithm parameters, see *Changing CN/LOH Algorithm Configurations for SNP 6.0 Analysis* on page 266.

For information about the algorithm description, see Appendix A, *Algorithms* on page 343.

CN/LOH QC Report Table for the Genome-Wide Human SNP Array 6.0

Use the *GTC Browser* (page 305) to view Copy Number, LOH, and CN Segments data in a genomic context.

Use the *Heat Map Viewer* (page 323) to view Copy Number data along with Copy Number Variation data, if available.

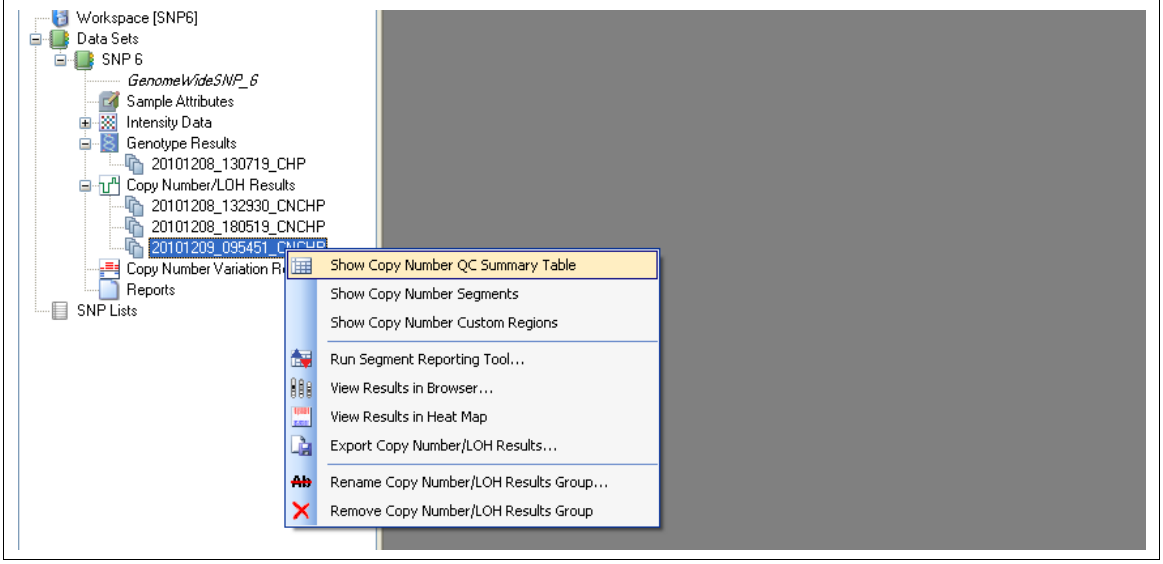
The Copy Number QC Summary Table displays QC information about the copy number and LOH analyses.

The Copy Number QC Summary Table uses all the table options as described in *Table Features* on page 198.

To open the QC Summary table:

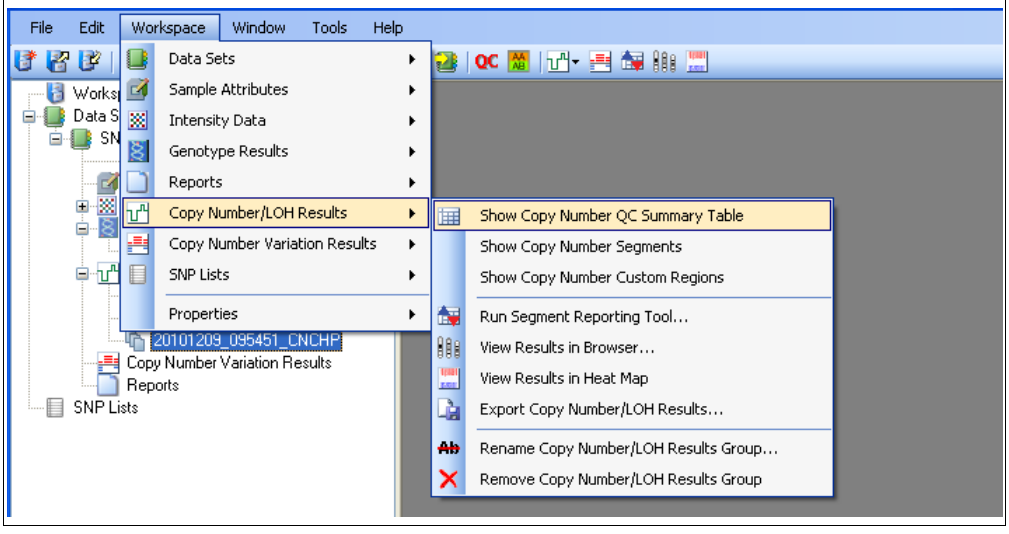
- Right-click on the Copy Number/LOH Results set of interest and select **Show Copy Number QC Summary Table** (Figure 13.36); or

Figure 13.36 Right-click data tree menu, Copy Number/LOH Results



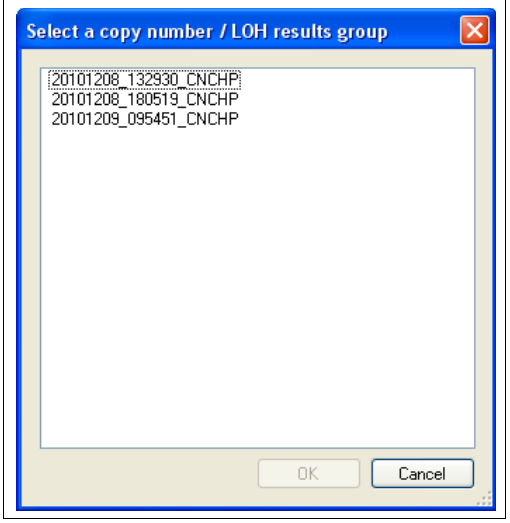
From the Workspace menu, select **Copy Number/LOH Results > Show Copy Number QC Summary Table** (Figure 13.37).

Figure 13.37 Workplace menu, Copy Number?LOH Results



If you have not selected a specific Copy Number/LOH Results group, the Select a copy number/LOH results group dialog box opens (Figure 13.38).

Figure 13.38 Select a copy number/LOH results group dialog box



Select a results group and click OK in the dialog box.

The Copy Number QC Report table opens (Figure 13.39).

Figure 13.39 Copy Number/LOH QC Report for Genome-Wide Human SNP 6.0 array (All Columns View)

Copy Number QC Report: SNP 6/20101208_132930_CNCHP

All Columns View

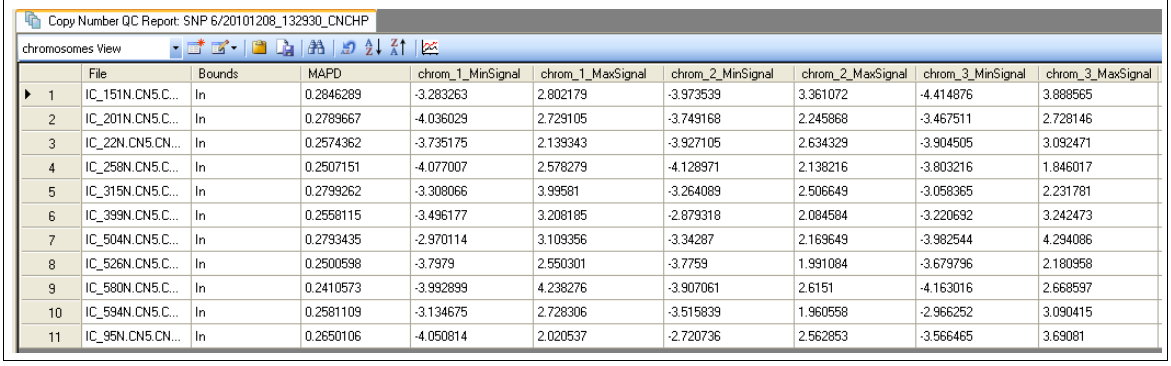
	File	Bounds	Gender	MedianAutosomeM	MAPD	iqr	all_probeset_rle_m	gc-correction-size	sample-median-cn-	sample-hom-
1	IC_151N.CN5.C...	In	Unknown	-0.0001244964	0.2846289	0.4225537	0.2704934	0.06150581	1.999983	0.6921684
2	IC_201N.CN5.C...	In	Unknown	0.001156741	0.2789667	0.4158621	0.2660266	0.03917071	1.997756	0.7058049
3	IC_22N.CN5.CN...	In	Unknown	0.0006417721	0.2574362	0.380118	0.2456367	0.03681666	1.999367	0.7192711
4	IC_258N.CN5.C...	In	Unknown	-0.0007790106	0.2507151	0.3610894	0.2350006	0.02168429	2.002646	0.7066078
5	IC_315N.CN5.C...	In	Unknown	0.0006783329	0.2799262	0.4142789	0.2669449	0.03670239	1.996865	0.6961487
6	IC_399N.CN5.C...	In	Unknown	-0.001160133	0.2558115	0.3671939	0.23989	0.02877164	2.003711	0.7064695
7	IC_504N.CN5.C...	In	Unknown	0.0008474602	0.2793435	0.4153564	0.2665798	0.04949306	1.997286	0.6981958
8	IC_526N.CN5.C...	In	Unknown	-0.001202096	0.2500598	0.3634907	0.2362318	0.03272379	2.0046	0.7169994
9	IC_580N.CN5.C...	In	Unknown	0.0006175083	0.2410573	0.3450513	0.2283465	0.02811114	1.999946	0.7148483
10	IC_594N.CN5.C...	In	Unknown	-0.0002132202	0.2581109	0.3686569	0.2407759	0.02709992	2.001262	0.7059542
11	IC_95N.CN5.CN...	In	Unknown	-0.0006088587	0.2650106	0.383046	0.2477547	0.03829601	2.001902	0.7092178

The Copy Number/LOH QC Report provides the following information for SNP 6.0 data (Table 13.1):

Table 13.1 SNP 6.0 Copy Number/LOH QC Report data

Item	Description
File	File name.
Bounds	In or out of QC bounds. See *Setting QC Thresholds* on page 312 for more information.
Gender	Gender call for the sample. It can be: • Male • Female • Unknown
MedianAutosomeMedian	Defined by taking the median of the medians of the log2 ratios of all autosomes, then subtracting this from each log2 ratio (including X and Y). This correction assumes the majority of the autosomes represent normal diploid DNA and this correction removes subtle array to array biases in normalization.
MAPD	Median absolute pairwise difference. See *MAPD and Copy Number QC on the Genome-Wide Human SNP Array 6.0* on page 265 for more information.
iqr	Interquartile range average for all chromosomes.
all_probeset_rle_mean	The mean absolute relative log expression (RLE) – This metric is generated by taking the probe set summary for a given array and calculating the difference in log base 2 from the median value of that probe set over all the arrays. The mean is then computed from the absolute RLE for all the probe sets for a given CEL file.
gc-correction-size	The median of the absolute value of the differences between uncorrected log2 ratios and GC waviness corrected log2 ratios.
sample-median-cn state	The median of all the calibrated (mapped in CN state space) log2 ratios for the sample.
sample-hom-frequency	The frequency (homozygous calls / all SNP calls) of SNP homozygous calls for the sample.
sample-het-frequency	The frequency (heterozygous calls / all SNP calls) of SNP heterozygous calls for the sample.
waviness-sd	The residual standard deviation (SD) after correcting for adjacent probe set to probe set SD based on autosomal log2 ratios. The waviness-sd is a measure of the signal variability in longer range waviness.
chrom_MinSignal (one for each chromosome)	Minimum Log2Ratio for a given chromosome and sample (Figure 13.40).
Chrom_MaxSignal (one for each chromosome)	Maximum Log2Ratio for a given chromosome and sample (Figure 13.40).
File Date	Date file was created.

Figure 13.40 Copy Number/LOH QC Report showing chrom_MinSignal and chrom_Max_Signal for chromosomes 1, 2, and 3



Copy Number QC Report: SNP 6/20101208_132930_CNCHP

chromosomes View

	File	Bounds	MAPD	chrom_1_MinSignal	chrom_1_MaxSignal	chrom_2_MinSignal	chrom_2_MaxSignal	chrom_3_MinSignal	chrom_3_MaxSignal
1	IC_151N.CN5.C...	In	0.2846289	-3.283263	2.802179	-3.973539	3.361072	-4.414876	3.888565
2	IC_201N.CN5.C...	In	0.2789667	-4.036029	2.729105	-3.749168	2.245868	-3.467511	2.728146
3	IC_22N.CN5.CN...	In	0.2574362	-3.735175	2.139343	-3.927105	2.634329	-3.904505	3.092471
4	IC_258N.CN5.C...	In	0.2507151	-4.077007	2.578279	-4.128971	2.138216	-3.803216	1.846017
5	IC_315N.CN5.C...	In	0.2799262	-3.308066	3.99581	-3.264089	2.506649	-3.058365	2.231781
6	IC_399N.CN5.C...	In	0.2558115	-3.496177	3.208185	-2.879318	2.084584	-3.220692	3.242473
7	IC_504N.CN5.C...	In	0.2793435	-2.970114	3.109356	-3.34287	2.169649	-3.982544	4.294086
8	IC_526N.CN5.C...	In	0.2500598	-3.7979	2.550301	-3.7759	1.991084	-3.679796	2.180958
9	IC_580N.CN5.C...	In	0.2410573	-3.992899	4.238276	-3.907061	2.6151	-4.163016	2.668597
10	IC_594N.CN5.C...	In	0.2581109	-3.134675	2.728306	-3.515839	1.960558	-2.966252	3.090415
11	IC_95N.CN5.CN...	In	0.2650106	-4.050814	2.020537	-2.720736	2.562853	-3.566465	3.69081

MAPD and Copy Number QC on the Genome-Wide Human SNP Array 6.0

MAPD is defined as the Median of the Absolute values of all Pairwise Differences between log2 ratios for a given array. Each pair is defined as adjacent in terms of genomic distance, with SNP markers and CN markers being treated equally. Hence any two markers that are adjacent in the genomic coordinates are a pair. Except at the beginning and the end of a chromosome every marker belongs to 2 pairs as it is adjacent to the marker preceding it and the marker following it on the genome.

MAPD is a per array estimate of variability, like Standard Deviation (SD). If the log2 ratios are distributed normally with a constant SD then MAPD/0.96 is equal to SD. MAPD is a robust QC check against high biological variability in log2 ratios induced by conditions such as cancer.

Variability in log2 ratios in an array arises from two distinct sources:

- Intrinsic variability in the starting material, hyb cocktail preparation, the array, the scanner
- Apparent variability induced by the fact that the arrays used to produce the reference file may have systematic differences from the array currently being analyzed.

Regardless of the source of the variability, increased variability in the log2 ratios decreases the quality of CN calls. Very high MAPD indicates that the log2 ratio differences for the given array are too large to recommend the array for further analysis. Variability in general will be reduced by using a reference set generated from arrays run at the same lab using the same reagent lots.

As in genotyping, there can be substantial batch effects or lab-to-lab systematic effects. If a reference is generated from arrays run in a lab other than the lab where the arrays used for analysis are run, such systematic differences inflate the apparent variability between the reference and the analysis set. Affymetrix has observed that using the supplied Affymetrix reference with arrays run in different labs will inflate MAPD by around 50%, although a factor of 2 is possible.

If an array with MAPD generated from the Affymetrix reference is greater than 0.35, then we recommend against using that array in an analysis.

When using a Reference Model File made up of arrays NOT generated in the same lab using the same reagent lots: CNCHP files with a MAPD value greater than 0.35 should not be used for further analysis.

When using a Reference Model File made up of arrays that WERE generated in the same lab using the same reagent lots: CNCHP files with a MAPD value greater than 0.3 should not be used for further analysis.

Changing CN/LOH Algorithm Configurations for SNP 6.0 Analysis

IMPORTANT: Affymetrix recommends that you perform Copy Number/LOH analysis with regional GC correction configuration.

NOTE: You cannot edit a configuration file that was created in GTC 2.1 or earlier. You can only edit configuration files that were created in GTC 3.0 or higher.

- *Creating a New Algorithm Configuration* on page 266
- *Restoring Configuration Settings to Default Values* on page 269
- *Editing a Configuration* on page 270

The configuration options are described in:

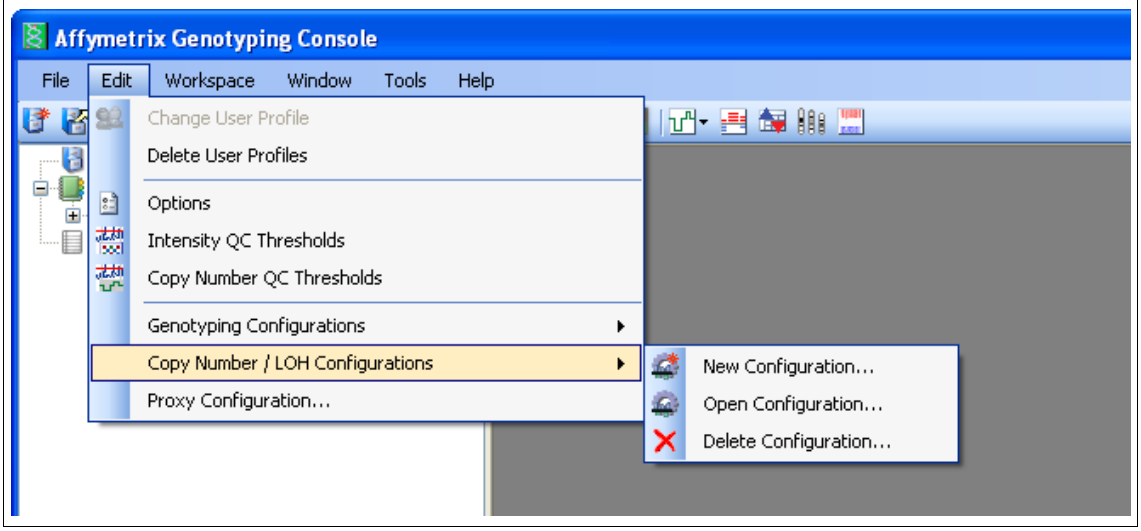
- *Basic Configuration Options for SNP 6.0 CN/LOH Analysis* on page 274
- *Advanced Configuration Options for SNP 6.0 CN/LOH Analysis* on page 276

Creating a New Algorithm Configuration

To create a new algorithm configuration:

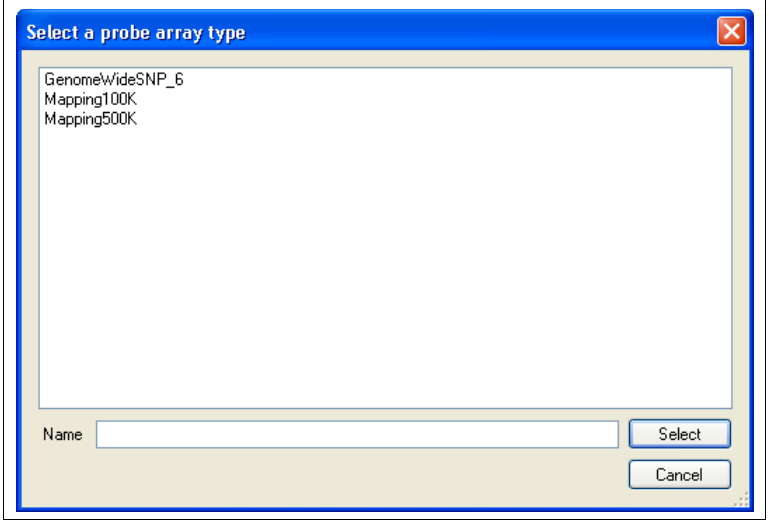
1. From the Edit menu, select **Copy Number Configurations > New Configuration** (Figure 13.41).

Figure 13.41 Select Probe Array Type dialog box



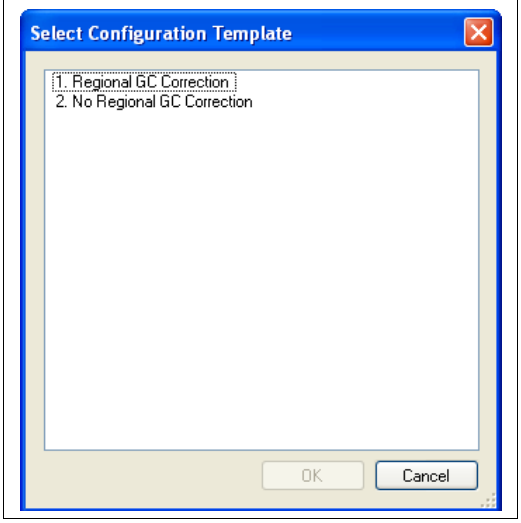
The Select Probe Array Type dialog box opens (Figure 13.42).

Figure 13.42 Select Probe Array Type dialog box



2. Select GenomeWideSNP_6 from the list and click **Select**.
   The Select Configuration Template dialog box opens.

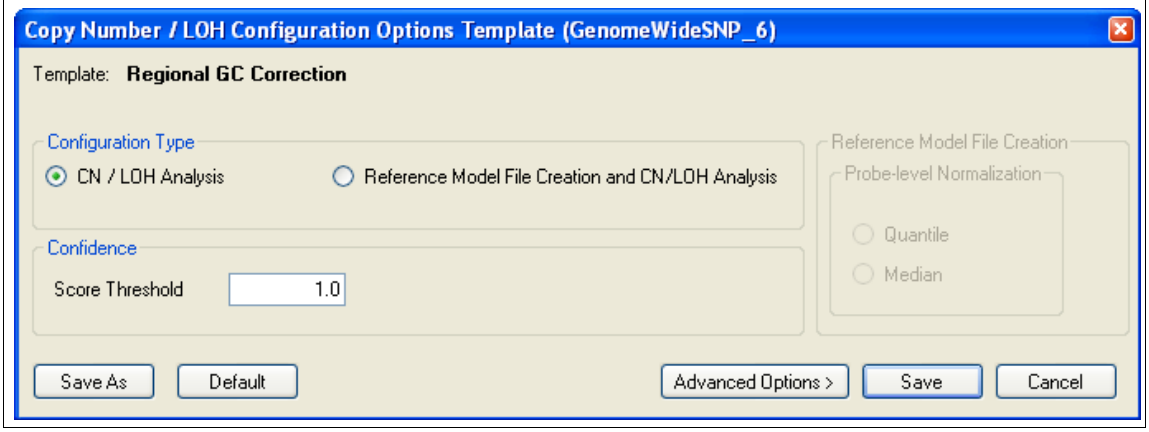
Figure 13.43 Select Configuration Template with or without regional GC correction



NOTE: The same configuration parameters are available for both the Regional GC Correction and No Regional GC Correction templates, but the default values for some parameters are different in the two template types. See *HMM Parameters* on page 278 for more information.

3. Select the configuration template and click **OK**.
   The CN/LOH Configuration Options Template dialog box opens (Figure 13.44).

Figure 13.44 CN/LOH configuration Options Template dialog box, Basic Options displayed



The Array type selected is displayed in the title bar.

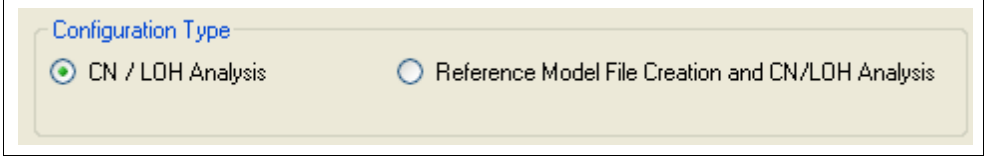
The template selected is displayed at the top of the dialog box.

The same parameters can be adjusted for Regional GC Correction and for No Regional GC Correction, although some of the default parameter values differ. These differences are explained in *HMM Parameters* on page 278.

4. Select the Configuration Type (Figure 13.45).
   - CN/LOH Analysis
   - Reference Model File Creation and CN/LOH Analysis

   Certain options are available only when the *Reference Model File Creation and CN/LOH Analysis* option is selected.

Figure 13.45 Selecting Configuration Type



See *Configuration Type* on page 274.

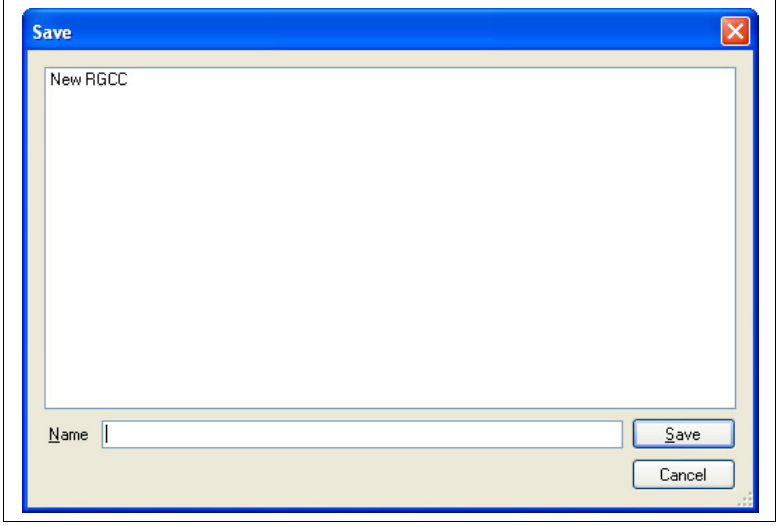
5. Enter values for configuration Options.

   The options are described in:
   - *Basic Configuration Options for SNP 6.0 CN/LOH Analysis* on page 274
   - *Advanced Configuration Options for SNP 6.0 CN/LOH Analysis* on page 276
6. Click **Save as** or **Save** in the CN/LOH Configuration Options Template dialog box.

   The Save dialog box opens (Figure 13.46).

NOTE: See *Restoring Configuration Settings to Default Values* on page 269 to restore the default values.

Figure 13.46 Save dialog box



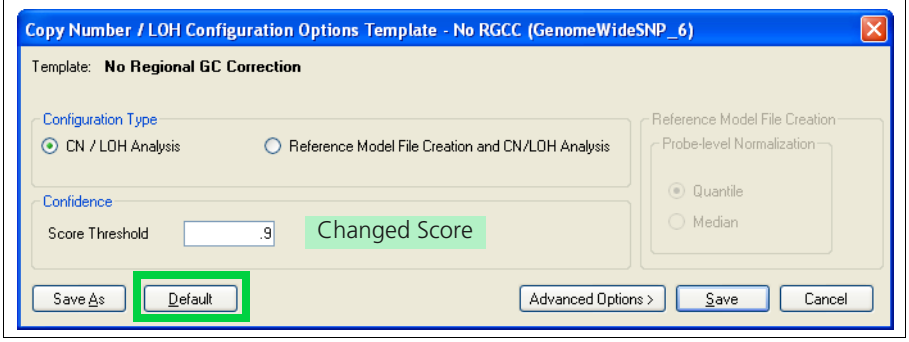
7. Enter a name for the configuration and click **Save** in the Save dialog box.

Restoring Configuration Settings to Default Values

To restore the default values for configuration settings:

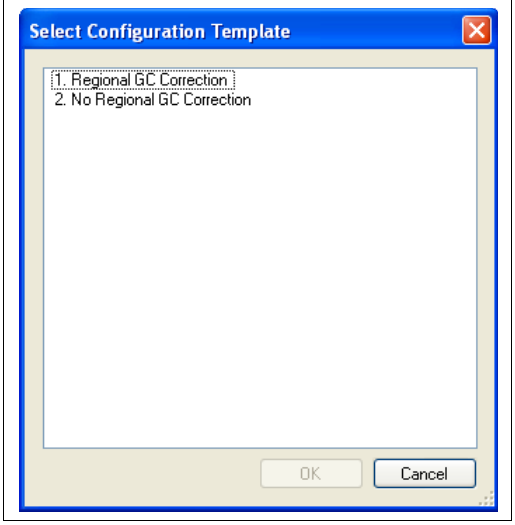
1. Click the **Default** button in the CN/LOH Configuration Options Template dialog box (Figure 13.47).

Figure 13.47 CN/LOH Configuration Options Template dialog box with score threshold changed



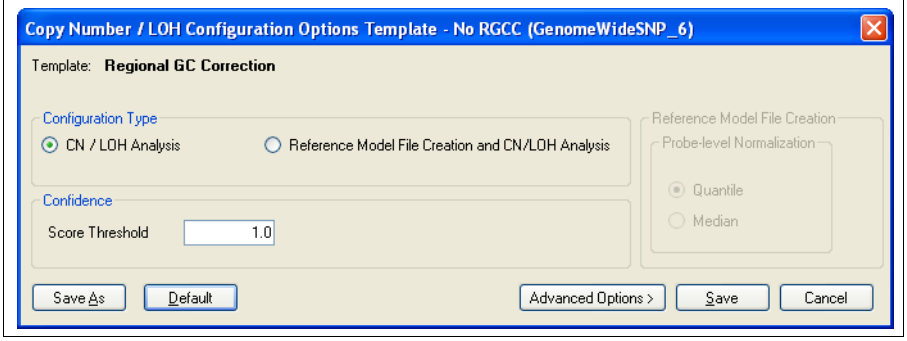
The Select Configuration Template dialog box opens (Figure 13.48).

Figure 13.48 Select Configuration Template dialog box



2. Select a template and click **OK** in the Select Configuration Template dialog box.
   The default configuration values for the selected template are restored in the CN/LOH Configuration Options Template dialog box (Figure 13.49).

Figure 13.49 CN/LOH Configuration Options Template dialog box with default values restored



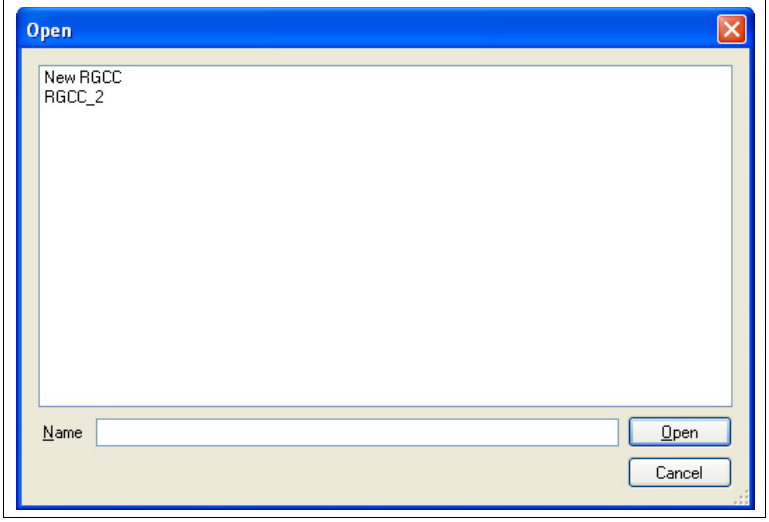
Editing a Configuration

NOTE: GTC cannot edit a configuration file that was created in GTC 2.1 or earlier. You can only edit configuration files that were created in GTC 3.0 and higher.

To edit a configuration that was created in GTC 3.0 and higher:

1. From the Edit menu, select **Copy Number Configurations > Open Configuration**.
   The Open dialog box opens (Figure 13.50).

Figure 13.50 Open dialog box

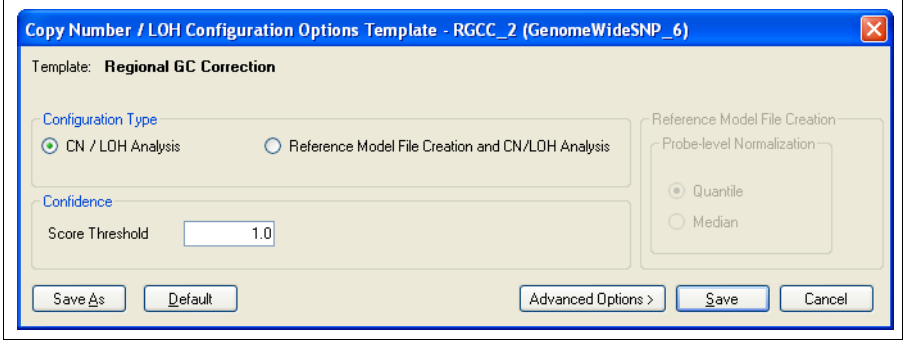


NOTE: The Open dialog box shows configuration files that were created in GTC 3.0 and higher. You can edit these files directly (see below for more information).

NOTE: Configuration files that were created in GTC 2.1 or earlier will be displayed in the Open dialog box to help you avoid overwriting them. You cannot edit these files directly (see below for more information).

2. Select the configuration file and click **Open**.
   The Copy Number/LOH Configurations Options Template dialog box opens (Figure 13.51).

Figure 13.51 Copy Number/LOH Configurations Options Template Dialog box, Edit mode



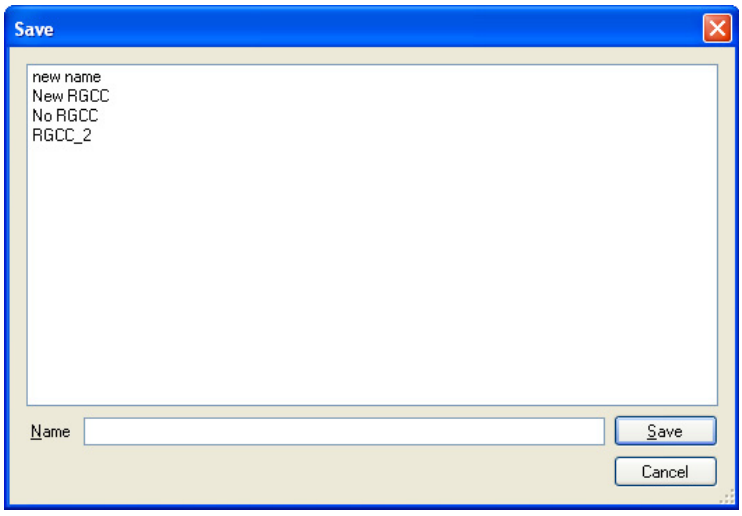
   The file name is displayed in the title bar of the dialog box when editing a configuration.

3. Select the configuration options and enter new parameter values
   The parameters are described in:
   - *Basic Configuration Options for SNP 6.0 CN/LOH Analysis* on page 274
   - *Advanced Configuration Options for SNP 6.0 CN/LOH Analysis* on page 276

4. Save the changes to the configuration:
   - To save the new values in the original configuration: Click **Save** in the Copy Number/LOH Configurations Options Template Dialog box.
     The configuration is updated with the new values.

- To save as new configuration:
  1) Click **Save as** in the Copy Number/LOH Configurations Options Template Dialog box. The Save dialog box opens (Figure 13.52).

Figure 13.52 Save dialog box for edited configuration



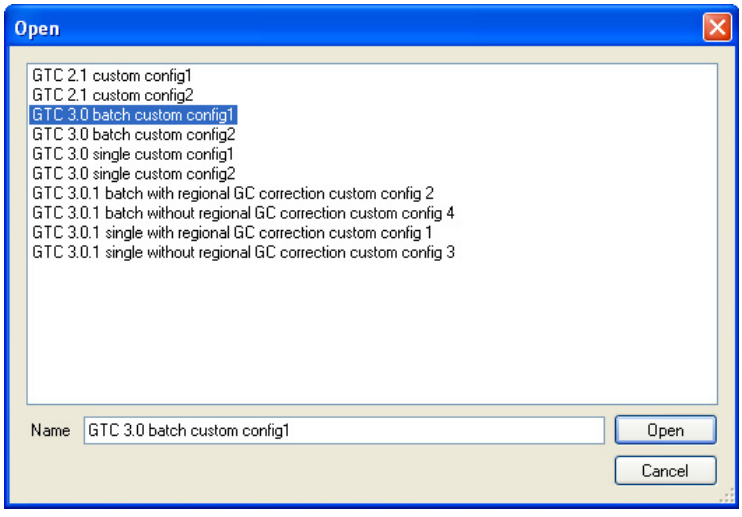
  2) Enter a configuration name and click the Save button in the Save dialog box. The new configuration is saved.

NOTE: See ***Restoring Configuration Settings to Default Values*** **on page 269** to restore the default values.

To edit a configuration that was created in GTC 3.0:

1. From the Edit menu, select **Copy Number Configurations > Open Configuration**. The Open dialog box appears (Figure 13.53).

Figure 13.53 Open dialog box



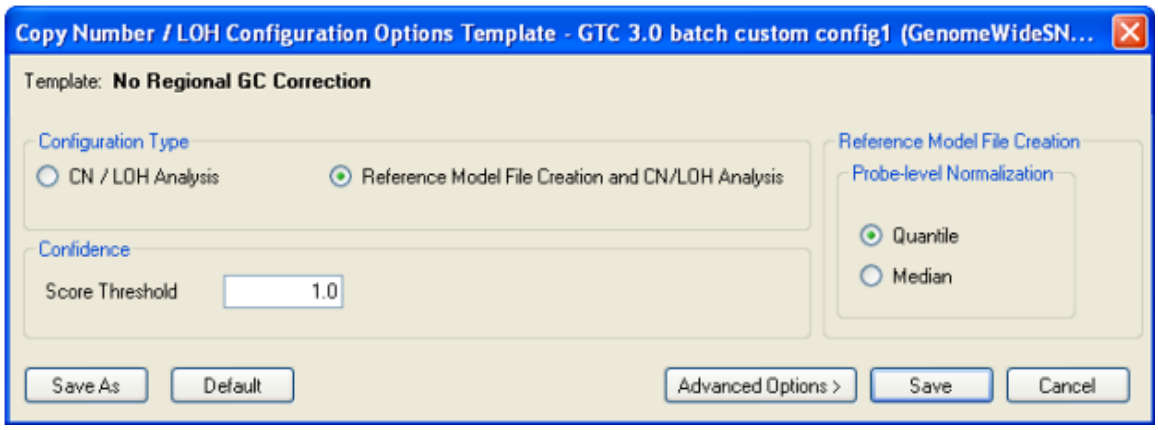
NOTE: Configuration files that were created in GTC 3.0 will be displayed in the Open dialog box for selection. You can edit these files directly (see below for more information). If GTC 3.0 configuration files are opened in GTC 3.0.1, these files will be treated as configurations without regional GC correction.

NOTE: If you are editing a configuration file created in GTC 3.0, you need to update the Score Threshold from 0.05 to 1.0 as the Affymetrix recommended new setting.

2. Select the configuration file and click **Open**.

   The CN/LOH Configuration Options Template dialog box opens with additional "**No Regional GC Correction**" added (Figure 13.54).

Figure 13.54 CN/LOH Configuration Options Template dialog box



3. Select the configuration options and enter a score threshold.
4. Save the changes to the configuration:
   - To save as new configuration: Click **Save as**.

NOTE: See ***Restoring Configuration Settings to Default Values*** **on page 269** to restore the default values.

To transfer parameters from a configuration created in GTC 2.1:

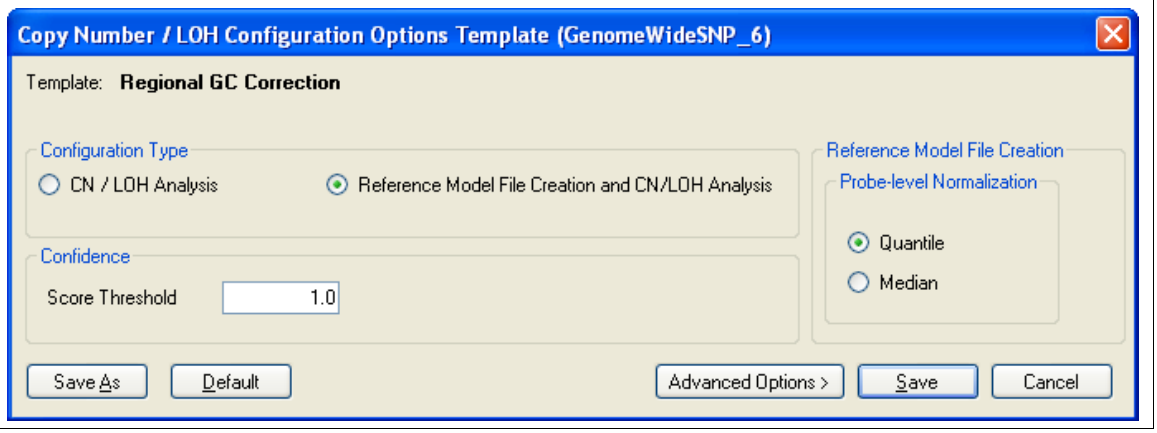
1. From the Windows Explorer, find the old configuration file and open it using text editor software.
2. Write down or print the old configuration file.
3. Make a new configuration file in GTC 3 using the old configuration parameters (a few parameters are new to GTC 3 configuration).

   The parameters are described in:
   - *Basic Configuration Options for SNP 6.0 CN/LOH Analysis* on page 274
   - *Advanced Configuration Options for SNP 6.0 CN/LOH Analysis* on page 276

Basic Configuration Options for SNP 6.0 CN/LOH Analysis

The basic options are displayed when the dialog box first opens (Figure 13.55).

Figure 13.55 Basic Options, regional GC correction template chosen



The Basic Options allow you to change:

- *Configuration Type* on page 274
- *Confidence Score Threshold Parameter* on page 274
- *Probe-level Normalization for Reference Model File Creation Parameter* on page 275

Configuration Type

You can create a configuration for the following types of analysis (Figure 13.56):

- CN/LOH Analysis
- Reference Model File Creation and CN/LOH Analysis

Figure 13.56 Configuration Type

NOTE: Some configuration options are available only when the Reference Model File Creation and CN/LOH Analysis option is selected.

Confidence Score Threshold Parameter

Confidence Score Threshold (Figure 13.57) is the maximum score at which the algorithm will make a genotype call.

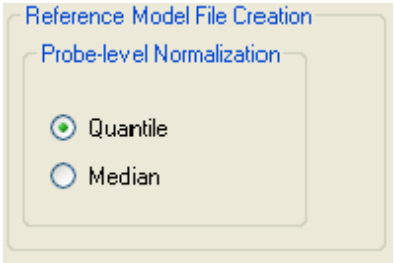
Figure 13.57 Confidence Score Threshold

Larger values of the score/confidence threshold indicate less certain calls. Calls with confidence scores above the threshold are assigned a no-call.

Probe-level Normalization for Reference Model File Creation Parameter

You can select different options for probe-level normalization for reference model file creation (Figure 13.58).

Figure 13.58 Reference Model File Creation: Probe-level Normalization



NOTE: This option is available only when the Reference Model File Creation and CN/LOH Analysis configuration option is selected.

- **Quantile normalization** is recommended for copy number analysis of association and cytogenetics samples. Quantile normalization is most appropriate for samples where most of the chromosomes are relatively normal.

  Quantile Normalization makes the entire distribution of data from the different arrays the same. The assumption for this method is that the signal distributions from all of the arrays should be similar. The data from each array is sorted and ranked from the lowest to the highest with each rank representing a quantile. The average intensity of each quantile is calculated across all the arrays. Then for each array in the set, the measured intensity in a given quantile is replaced with the calculated average intensity. All arrays in the data set now have identical distributions.
- In contrast, many cancer samples contain significant abnormalities that impact much of the genome; therefore, **median normalization** is recommended.

  Median Normalization scales all of the arrays in a set so that they have the same median intensity. This is a linear normalization method that will normalize all of the arrays to the median value of the medians for the individual arrays.

For a more detailed discussion of normalization please refer to *A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Variance and Bias*. Bioinformatics. 2003 Jan 22;19(2):185-93. B. M. Bolstad, R. A. Irizarry, M. Astrand and T. P. Speed.

IMPORTANT: For any single sample Copy Number/LOH Analysis run, the Probe-level Normalization and Probe-level Background Correction parameters must be and will be set to the same values for the Analysis as the parameters used to generate the Reference Model File used in the Analysis.

Advanced Configuration Options for SNP 6.0 CN/LOH Analysis

Click the **Advanced Options** button (Figure 13.59) to display the Advanced Options for CN/LOH Configuration (Figure 13.60).

Figure 13.59 Basic Options, regional GC correction template chosen

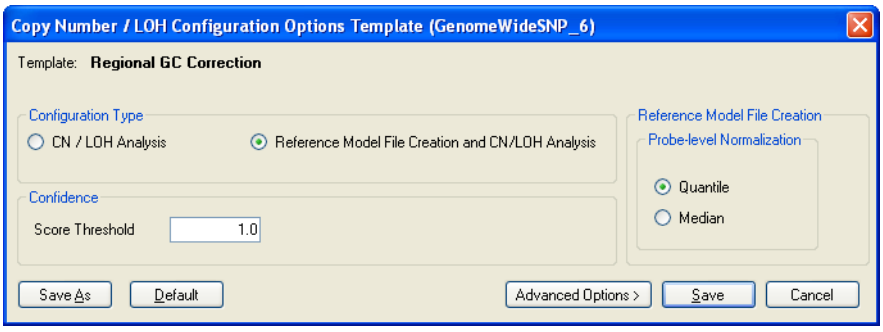
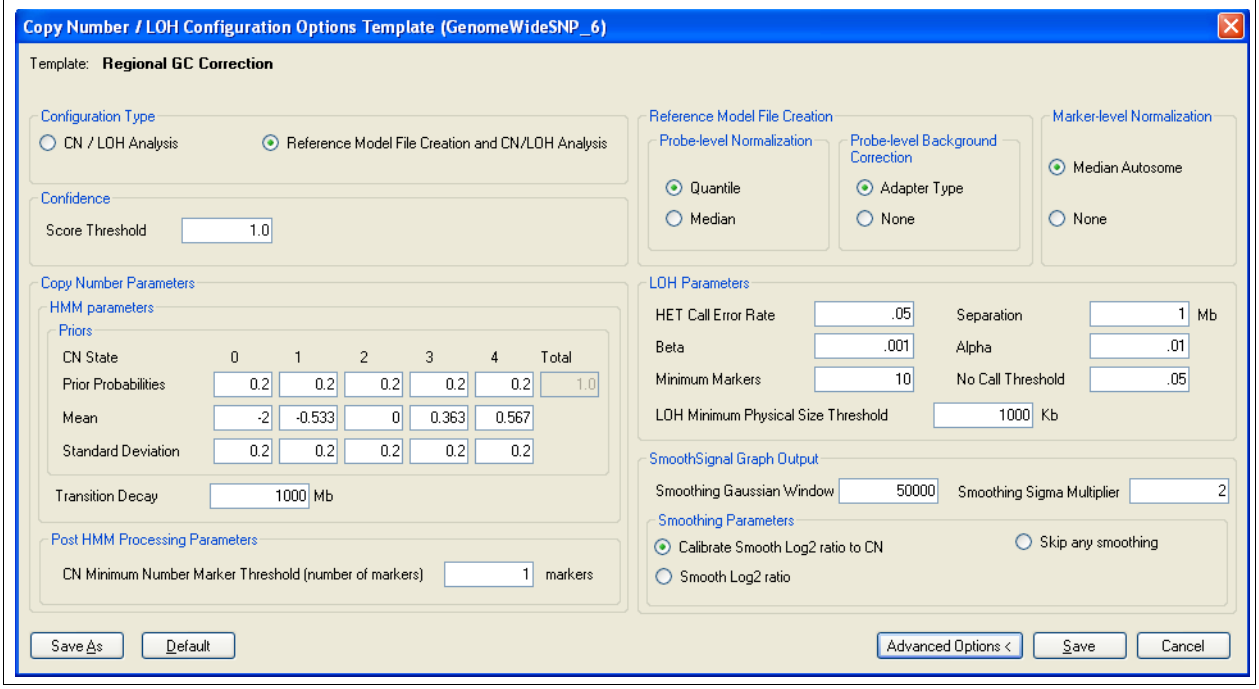


Figure 13.60 CN/LOH Configurations Options Template Dialog box: Advanced Options displayed



The Advanced Options include:

- *Reference Model File Creation: Probe Level Background Correction* on page 277
- *Marker-level Normalization* on page 277
- *Copy Number Parameters:* on page 278

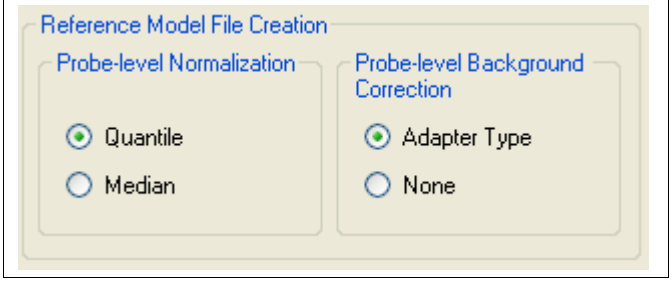
- *LOH Parameters* on page 281
- *SmoothSignal Graph Output* on page 283

Reference Model File Creation: Probe Level Background Correction

NOTE: This option is available only when the Reference Model File Creation and CN/LOH Analysis configuration option is selected.

The SNP 6.0 assay uses both Sty and Nsp enzymes to cut the original DNA into fragments. Each enzyme has four alternative recognition sites (adapters). Fragment-specific amplification has been observed depending on the particular pair of adapters used to cut out fragments. Such fragment-specific effects are typically very similar within a set of samples run together, but between sample sets such effects are occasionally quite different.

Figure 13.61 Reference Model File Creation Options: Probe-level Background Correction

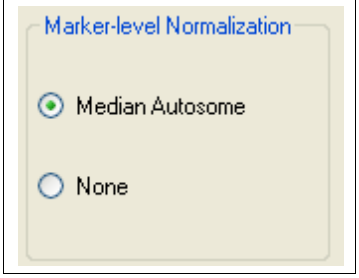


The probe-level "Adapter Type" normalization (Figure 13.61) is used to ensure the fragment effects are uniform across all samples. For any single sample Copy Number/LOH Analysis run, the Probe-level Normalization and Probe-level Background Correction configuration parameters should be set the same for the analysis as these parameters were set during the generation of the Reference Model file used in the analysis.

Marker-level Normalization

You can select the Median Autosome marker-level normalization (Figure 13.62) as an optional normalization done after log2 ratios are calculated: the log2 ratios are adjusted by subtracting the median of the median log2 ratio of all the autosomes.

Figure 13.62 Marker-level Normalization options



This adjustment can be useful for samples with primarily diploid autosomes when probe-level normalization may be affected by an aneuploidy such as high CN gain in Chromosome X. Note that this adjustment is not a probe-level background correction. This is only recommended for samples where most of the chromosomes are relatively normal.

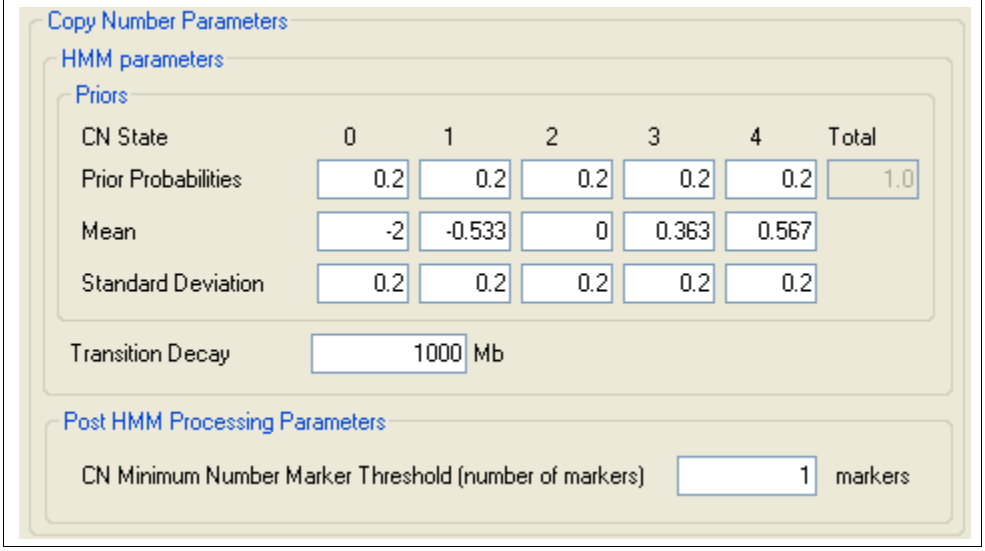
Copy Number Parameters:

These advanced parameters (Figure 13.63) enable you to adjust the Copy Number performance

You can adjust:

- *HMM Parameters* on page 278
- *Post-HMM Processing Parameters* on page 280

Figure 13.63 Default Copy Number Parameters for Regional GC Correction Template



HMM Parameters

CN State represents the possible values that the HMM can find. The HMM looks for CN states 0, 1, 2, 3 and 4-or-greater. CN state of 5 or more will also be represented as CN State 4.

A 5-state Hidden Markov Model (HMM) is applied for smoothing and segmenting the CN data. The user tunable parameters for the HMM are:

- *Priors Settings* on page 278
- *Transition Decay* on page 280

Priors Settings

The HMM has 5 possible states:

State 0 = CN of 0; homozygous deletion

State 1 = CN of 1; heterozygous deletion

State 2 = CN of 2; normal diploid

State 3 = CN of 3; single copy gain

State 4 = CN 4; amplification

For each of these states you can modify the following priors values:

- Copy Number State

  The default for each state is 0.2 indicating that each SNP has equal prior probability of being in any one of the 5 states. We have not extensively tested the impact of modifying this initial estimate on the performance of the HMM.

> **NOTE:** The prior values entered are only initial estimates. The HMM optimizes this parameter based on the data.

- Mean

  The mean is the expected log2 ratio of each CN state. For example if the reference is diploid for each marker then the expected log2 ratio for CN = 2 is 0. A Chromosome X titration experiment was performed using samples that have differing numbers of X chromosomes spanning the range of the HMM. The observed log2 ratios for different copy numbers of Chromosome X were used to set the default mean for each state (except CN = 0, which is unchanged from CNAT 4).

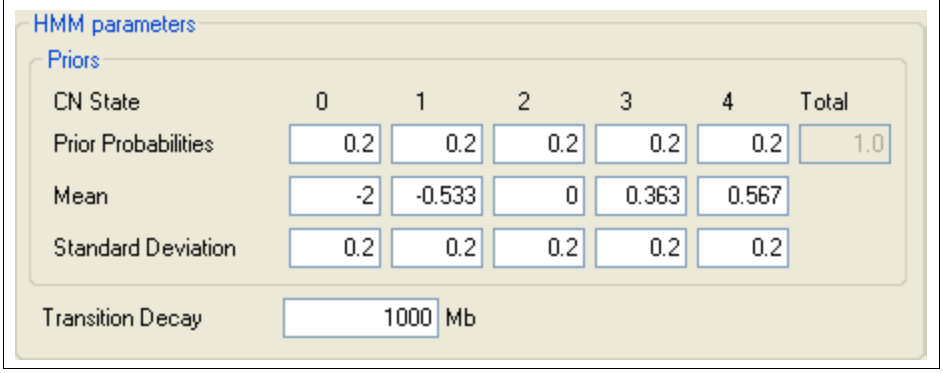
- Standard Deviation

  Standard deviation is one of the parameters that affect the probability with which the underlying CN state is emitted to produce the observed state. Specifically, it reflects the underlying variance or dispersion in each CN state. The standard deviation of each underlying state can be adjusted. The defaults in the SNP 6.0 parameters are a little lower than the observed SD's in each state, but when adjusted during testing to match the observed SD's did not improve the results of the HMM.

The HMM prior options have different parameters for regional GC and no Regional GC

The Hidden Markov Model (HMM) **with regional GC correction** is modified for SNP 6.0 with the following changes in the Mean values for different CN states (Figure 13.64).

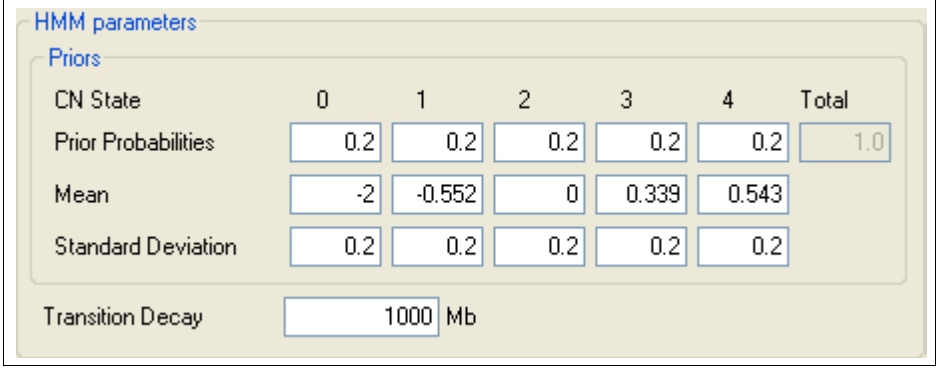
Figure 13.64 HMM parameters for SNP 6.0 with Regional GC Correction



In the SNP 6.0 CN/LOH configuration **without regional GC correction** (Figure 13.65), the same Hidden Markov Model (HMM) is used for SNP 6.0 as that used in CNAT 4, with the following notable exceptions for SNP 6.0:

- Smoothing log2 ratios prior to using the HMM is not possible
- Signal in log2 ratios for SNP markers is always "logSum"
- The "sumLog" signal summary is not possible

Figure 13.65 HMM parameters for SNP 6.0 with no Regional GC Correction

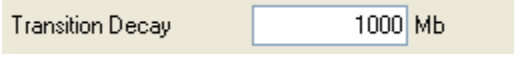


Accordingly, other than smoothing, the same parameters in CNAT 4 are exposed as advanced options. These parameters are used to define how the HMM calculates per marker Copy Number from log2 ratios.

Transition Decay

The Transition Decay parameter (Figure 13.66) controls the expected correlation between adjacent markers. The copy number state of any given marker is partially dependent on that of its neighboring markers and is weighted based on the distance between them. By adjusting this parameter, neighboring markers can either have more or less of a dependence on each other.

Figure 13.66 Transition Decay parameter



The default value is 1000 Mb.

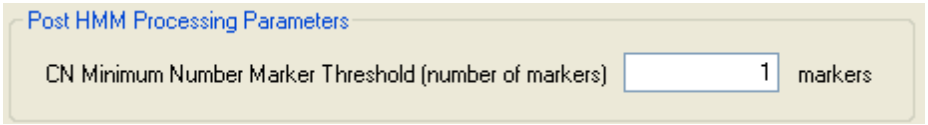
To reduce the influence of neighboring markers, decrease this value (transition faster). For example, if you set the decay to 1 Mb, and if a given marker is in CN State 1, the probability that the flanking markers to the right will continue to be in State 1 is much lower compared to the case where the transition decay is 100 Mb.

To increase the influence of neighboring markers, increase this value (transition slower).

Post-HMM Processing Parameters

Occasionally a marker (typically a CN probe) on the SNP 6.0 array performs erratically for unknown reasons. The outcome may be occasional singleton calls of CN different from unchanging CN in flanking markers (both CN and SNP) surrounding this marker.

Figure 13.67 Post HMM processing parameters



Setting the CN Minimum Number Marker Threshold parameter (Figure 13.67) to 1 changes the CN determination of such markers to agree with the other markers surrounding it.

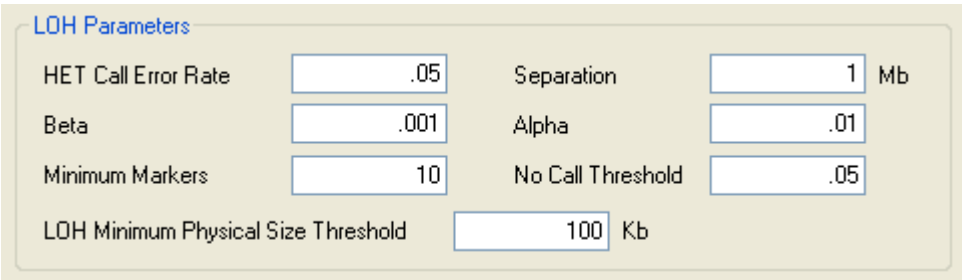
For example, if there is a single marker that is called CN State 1 by the HMM, but the surrounding markers are called CN State 2, then this singleton SNP will be changed from CN State 1 to CN State 2.

Setting this parameter to 0 leaves the original CN State value unchanged. For the SNP 6.0 array this field refers to the number of flanking markers and can only be 0 or 1.

LOH Parameters

The SNP 6.0 LOH algorithm looks for runs of homozygous SNP calls, taking into account the overall het rate and the likely error rate in calling.

Figure 13.68 LOH parameters



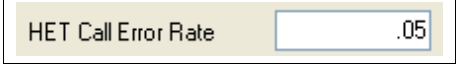
The following LOH Parameters can be adjusted (Figure 13.68).

HET Call Error Rate

The Genotyping algorithms perform well in the context of signal from diploid SNPs, with very low error rates. However, when signal arises from a non-diploid SNP, the genotyping error rate is higher. In the case of LOH associated with a CN =1 region, (e.g., as in a single X chromosome without special treatment by the genotyping algorithm) then, while we would expect no hets at all to be called, in practice with current default SNP 6.0 genotyping parameters, it is more usual to see around 5% het call rates depending on sample quality.

Lower quality data will result in a higher het call error rate. The algorithm auto-adjusts the het call error rate in the following case: if LOH is being called as part of a reference model generation and the default no-call threshold (.05) for genotyping is used, then the algorithm will adjust the het call error rate upwards if necessary, depending on the observed no-call genotyping rate. In all other cases the het call error is left as the value in the panel (Figure 13.69).

Figure 13.69 HET Call Error Rate



The het call error rate is tuned for LOH in hemizygous deletions (i.e. a loss of a portion of 1 chromosome out of a pair). In fact small regions of Copy Neutral LOH are very common; they may arise from portions of paired chromosomes that can be traced through different lines of descent back to a single ancestor and so these regions are identical and hence homozygous. To detect such Copy Neutral LOH, a het call error rate of .02 is more appropriate.

Alpha and Beta

The LOH algorithm depends on 2 concepts:

- Alpha (if LOH is present, this is the chance that the algorithm would fail to call it)

  Given that LOH is truly present, what are the odds that it is not found given the het call error rate? This is referred to as Type I error and is traditionally referred to as “Alpha” in statistics. Decreasing alpha decreases the odds the algorithm will falsely rule against LOH but increases the odds it will falsely find LOH.

- Beta (if normal Heterozygosity is present, this is the chance of mistakenly calling it LOH)

  Given the usual or expected rate of heterozygosity in a region what are the odds of falsely finding LOH? This is referred to as Type II error or statistical power and is traditionally referred to as "Beta" in statistics. Decreasing Beta decreases the odds the algorithm will falsely find LOH but increases the odds it will fail to find LOH when it is present.

  The Alpha and Beta parameters can be adjusted.

Figure 13.70 LOH parameters

Minimum Markers

The Minimum Markers parameter sets the minimum number of SNPs to be used in evaluating LOH. The algorithm calculates the number of SNPs needed to satisfy alpha and beta. For the supplied alpha and beta defaults this calculated number is well in excess of the default (10 marker) minimum, but if you decide to change the alpha and beta parameters then the Minimum Markers parameter can be used as a safety net.

Separation

LOH is calculated based on the assumption that LOH is found over a contiguous region of the genome. When gaps occur in the genome (such as across a centromere), LOH can be calculated separately for each stretch of the genome. The typical distance between SNPs on SNP 6.0 is on the order of 1,300 bases. The separation parameter controls how many base pairs must separate 2 markers before the LOH algorithm starts calculating the LOH value for a new stretch of genome.

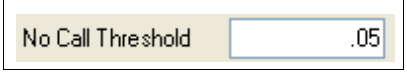
Figure 13.71 Separation

At the separation parameter's default setting the LOH algorithm will treat each chromosome as a region.

No Call Threshold

In any one sample not all SNP results provide equally reliable data. Some SNP results give high quality information about the genotype they call, and others give low quality information. Including low quality SNP calls increases the het call error rate over any improvement in the algorithm's accuracy by including these extra SNPs. The quality of the SNP call is captured by its Confidence value (as defined in the genotyping algorithm), and the No Call Threshold excludes SNPs with a greater Confidence value than this parameter value (Figure 13.72).

Figure 13.72 No Call Threshold



LOH Minimum Physical Size Threshold

This parameter sets a minimum size that LOH blocks must exceed to be reported as LOH (Figure 13.73).

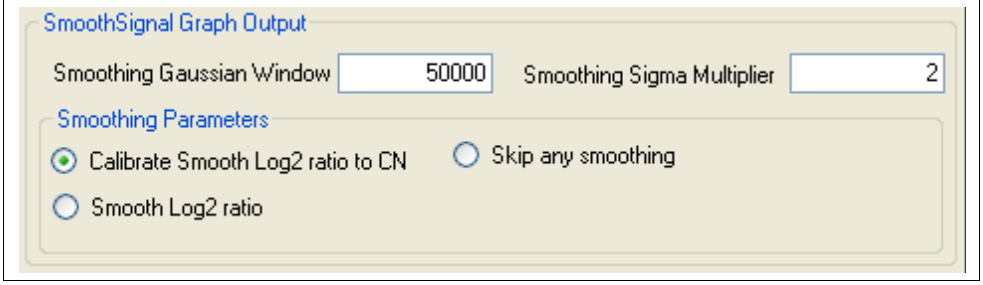
Figure 13.73 LOH Minimum Physical Size Threshold

As described above small regions of Copy Neutral LOH are very common; they may arise from portions of paired chromosomes that can be traced through different lines of descent back to a single ancestor and so these regions are identical and hence homozygous. Thus, some Copy Neutral LOH regions are associated with haplotype blocks. As region size increases the odds we see Copy Neutral LOH related to a haplotype block decreases. The HapMap phase 1 study shows roughly 70% of common haplotype blocks in humans are less than about 100 kilobases; further increasing The LOH Minimum Physical Size Threshold above the 100kb value will screen out Copy Neutral LOH regions arising from haplotype blocks at the expense of losing smaller LOH regions arising from other events (e.g., a deletion event).

SmoothSignal Graph Output

These options control the signal smoothing functions (Figure 13.74).

Figure 13.74 Smoothing Signal Graph Output



Smoothing Gaussian Window

The log2 ratio is the raw estimate of log of CN signal compared to an expected state of CN=2 for each marker. These raw estimates can be smoothed using a Gaussian kernel to lower noise to improve per marker Signal to Noise ratio at the expense of blurring the boundaries where the CN state changes. For each marker, the smooth is constructed using a weighted mean of the log2 ratios of surrounding markers with weights proportional to the Gaussian transform of their genomic distance from that marker. The Gaussian transform has Standard Deviation equal to the "Smoothing Gaussian Window." (Figure 13.75).

Figure 13.75 Smoothing Gaussian Window

In usual signal processing terminology this parameter is known as the bandwidth.

Setting this value to 0 will result in no smoothing.

Smoothing Sigma Multiplier

In principle, the Gaussian smooth uses all markers. In practice surrounding markers far from any particular marker have little numerical impact on the final smoothed value. The Smoothing Sigma Multiplier parameter (Figure 13.76) determines the number of Standard Deviations away from the given marker where markers will be included in the smooth.

Figure 13.76 Smoothing Sigma Multiplier

Note that larger values will result in increased compute times for the algorithm.

Setting this value to 0 will result in no smoothing.

Smoothing Parameters options

You can select from different smoothing parameter options (Figure 13.77).

Figure 13.77 Smoothing Parameters

- Calibrate Smooth Log2 ratio to CN

  Checking this option calibrates Smooth Log2 ratio to the HMM mean parameters for different CN states and inverts the resultant smoothed log2 ratio to normal Copy Number. So a 0 value in the smoothed log2 ratio will become 2 after inverting. If the HMM mean corresponding to CN state of 1 is -.55 then a smoothed log2 ratio with a value of -.55 will be inverted to CN = 1. If the Smoothing Gaussian Window is 0 or the Smoothing Sigma Multiplier is 0, then calibration and inversion to CN units occurs without any smoothing.
- Smooth Log2 Ratio

  Checking this option results in the smoothed log2 ratios only.
- Skip any smoothing

  Checking this option will prevent smoothed log2 ratios from being calculated and included in the CNCHP file output.

Common Functions for Copy Number/LOH Analyses

This chapter covers the copy number/LOH functions that are common to both Human Mapping 100K/500K arrays and the Genome-Wide Human SNP Array 6.0. These functions include:

- *Using the Segment Reporting Tool* on page 285
- *Loading Data into the GTC Browser* on page 305
- *Export Copy Number/LOH data* on page 307
- *Setting QC Thresholds* on page 312

Using the Segment Reporting Tool

You can use the Segment Reporting Tool (SRT) to locate segments with copy number changes in the CN data for 100K/500K and SNP 6.0 array data. The SRT detects both common and unique-to-a-sample copy number change segments.

For SNP 6.0 data the SRT also produces a gender call for the sample, based on the detected copy number state for the X and Y chromosomes. See *CN Segment Report (SNP 6.0 only)* on page 356 for more information about the CN Segment Report Tool's gender call.

More information is given in:

- *Introduction* on page 285
- *Running the Segment Reports Tool* on page 287
- *Segment Report Tool Results Files* on page 298

NOTE: The SRT requires annotation files (*.annot.db) to analyze CNCHP files generated in earlier versions of GTC. For Human Mapping 100K or 500K data, the SRT requires na24 version of the annotation file (*.na24.annot.db). For Genome-Wide SNP Array 6.0 data, the SRT requires na25 to na29 version of annotation files (*.annot.db), depending on the annotation version that was used to generated the CNCHP file.

Introduction

The Segment Reporting Tool runs three different processes on the data:

1. *Detect CN Segments that meet initial requirements* (below)
2. *Filter out segments that overlap with known CNV regions (optional)* on page 286
3. *Generate Custom Reports Using a Custom Regions File (optional)* on page 286

   At the end of these processes, a Segment Report (*.cn_segments) is generated for each copy number file. An optional Segment Summary Report that concatenates the segments from each Segment Report can also be generated.

   If the Custom Region option is used, a Custom Regions Report file is created for each .CNCHP file analyzed. An optional Custom Regions Summary Report that concatenates the segments from each Custom Regions report can also be generated.

Detect CN Segments that meet initial requirements

This process detects all the copy number change segments in the CNCHP files that meet the initial filtering parameters for:

- Minimum number of markers per Segment
- Minimum genomic size of a Segment

Filter out segments that overlap with known CNV regions (optional)

The SRT can filter out segments that overlap with known CNV regions based on a user-defined percentage of markers in the segment.

If the filter value is set to 25%, all segments that overlap known CNV Regions by more than 25% are **not** included in the report. Segments that overlap by 25% or less **are included** in the reports.

The SRT produces a **Segment Report** file (*.cn_segments) for each copy number file that is analyzed. The Segment Report files contain information on the copy number segments detected in a given CNCHP file.

The Segment Report files can be viewed:

- In the Copy Number Segment Report table of GTC 4.2
- In the GTC Browser

See *Segment Report File* on page 298 for more information.

The SRT can also generate an optional **Segment Summary Report** file (*.cn_segments_summary) concatenating the segment data for all of the CNCHP files analyzed in a particular run.

The *Segment Summary Report* on page 301 can be viewed in a spreadsheet program.

Generate Custom Reports Using a Custom Regions File (optional)

You can also use a Custom Regions file to generate Custom Regions reports for each copy number file. The Custom Regions file defines regions of the genome of interest. The Custom Regions Report allows analysis of "favorite" regions of the genome without needing to filter the cn_segments_report manually for these regions, or viewing data in the GTC Browser.

A sample template Custom Input Regions file (Custom Regions template cn_input_regions.bed) is located in the Library folder.

See *Custom Region File Format* on page 296 for more information.

The **Custom Regions** reports include information on segments with copy number changes in the defined custom regions. The report includes:

- custom region name
- overlap of the region with the segment and vice versa
- genomic location
- size
- # of markers in the segment
- overlap with known CNVs
- other annotation

The Custom Regions Report files can be viewed:

- In the Copy Number Segment Report table of GTC 4.2
- In the GTC Browser

You can also generate an optional **Custom Regions Summary Report** file (.custom_regions_summary) concatenating the custom regions data for all of the files analyzed in a particular run.

The *Segment Summary Report* on page 301 can be viewed in a spreadsheet program.

IMPORTANT: **The Custom Input Regions file can be loaded into the GTC Browser as a track. This allows you to view your Custom Regions in a genomic context.**

Running the Segment Reports Tool

The basic operation of the Segment Reporting Tool is described below.

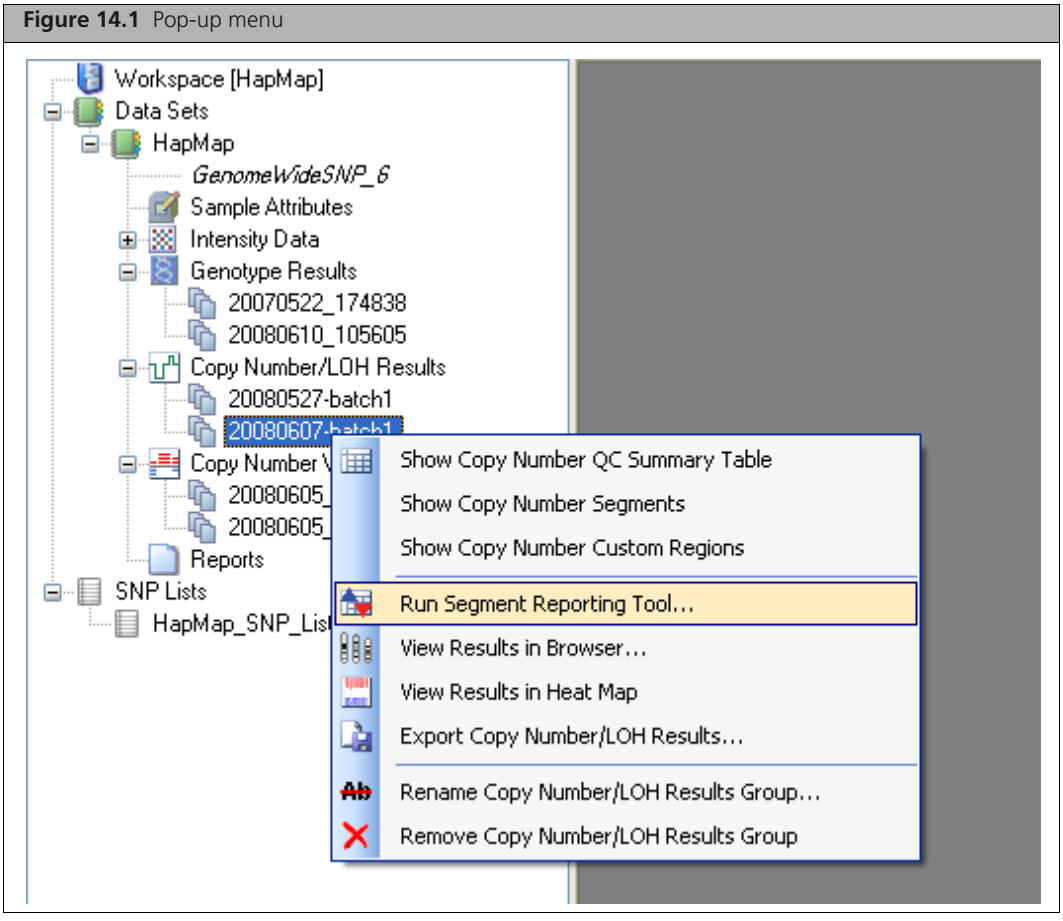
You can select from several options for using the Segment Reporting Tool:

- *Selecting CNV Map for Filtering Segments* on page 292
- *Selecting Filters* on page 293
- *Filter out segments that overlap with known CNV regions (optional)* on page 286
- *Adding a Suffix to the Segment Report Files* on page 294
- *Create Segment Summary Report File* on page 294
- *Using a Custom Regions File* on page 295
- *Create Custom Regions Summary report* on page 297

To create a segment report:

1. Select the results set you wish to generate a report for.
2. Do one of the following:
   - From the Workspace menu, select **Copy Number/LOH Results > Run Segment Reporting Tool**.
   - Right-click the Copy Number/LOH Results file set and select **Run Segment Reporting Tool** from the pop-up menu (Figure 14.1).

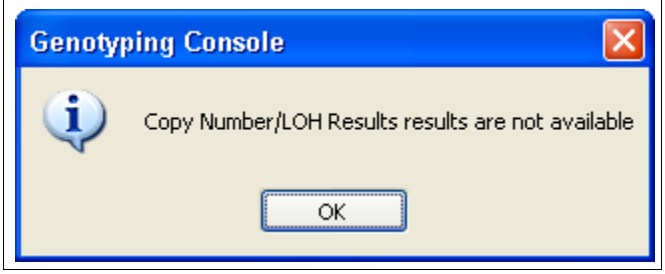
Figure 14.1 Pop-up menu



- Click the **Run Segment Reporting Tool** button in the tool bar.

  If you have selected a data set with no copy number files available, the following warning notice appears (Figure 14.2).

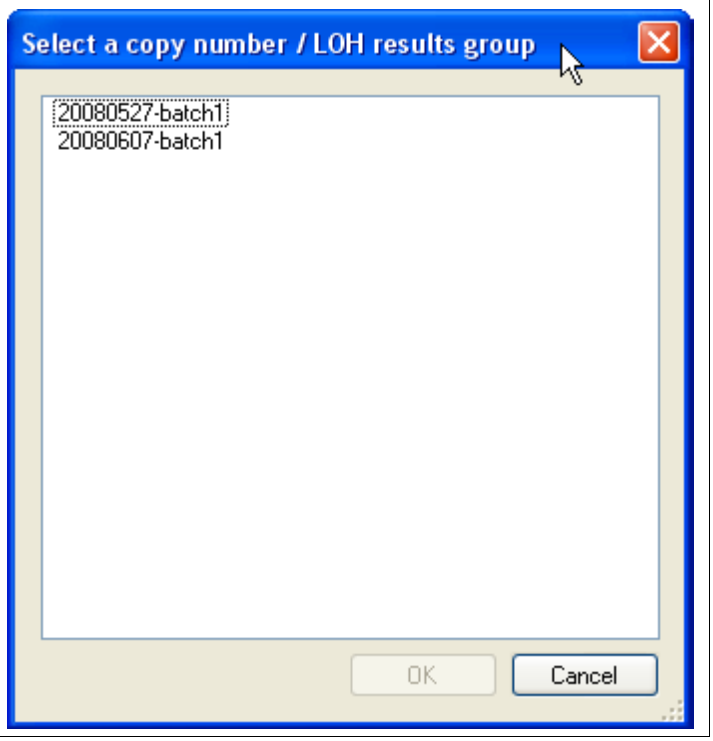
Figure 14.2 Warning notice



If you see this notice, click **OK** and then select a data set with copy number data.

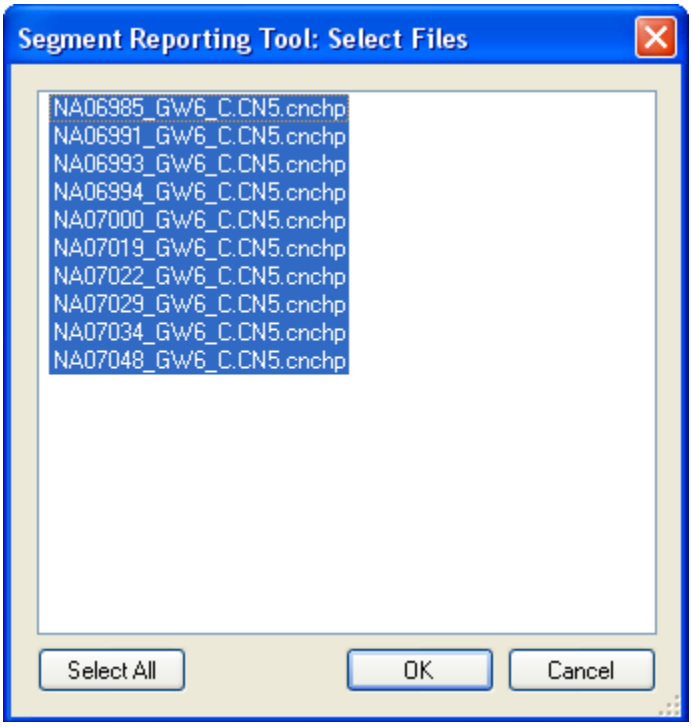
If you have selected a data set with more than one copy number result batch available, the following notice appears to ask you to choose a CN result batch ((Figure 14.3).

Figure 14.3 Select Results Group dialog box



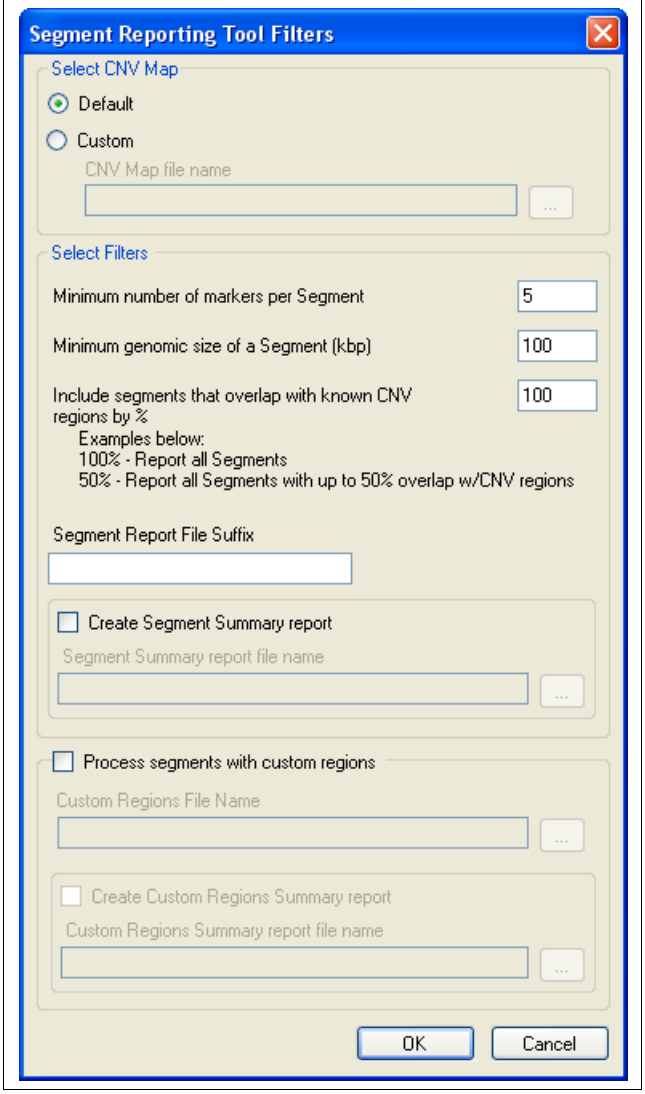
3. Select the group you wish to analyze and click **OK**.
   The Select Files dialog box opens (Figure 14.4).

Figure 14.4 Select files dialog box



4. Select the copy number data files you wish to analyze and click **OK**. The Segment Reporting Tool Filters dialog box opens (Figure 14.5)

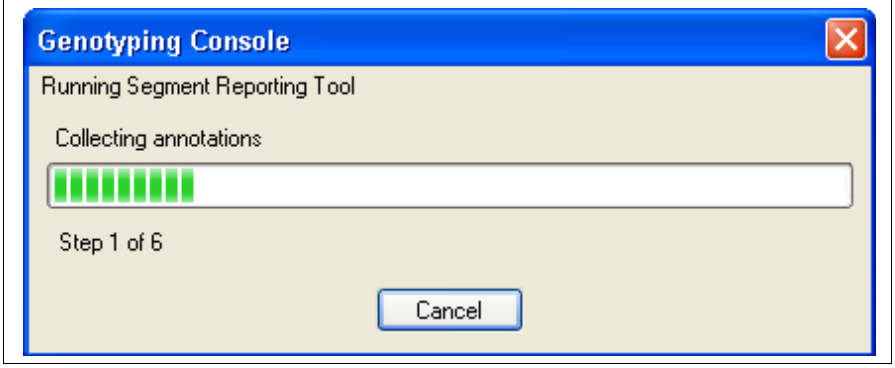
Figure 14.5 Segment Reporting Tool Filters dialog box



5. Select the options that will be used to generate the report (see below):
   - *Selecting CNV Map for Filtering Segments* on page 292
   - *Selecting Filters* on page 293
   - *Filter out segments that overlap with known CNV regions (optional)* on page 286
   - *Adding a Suffix to the Segment Report Files* on page 294
   - *Create Segment Summary Report File* on page 294
   - *Using a Custom Regions File* on page 295
   - *Create Custom Regions Summary report* on page 297
6. Click **OK** in the Segment Reporting Tool Filters dialog box.
   The Progress bar (Figure 14.6)displays the progress of the Segment Reporting Tool,

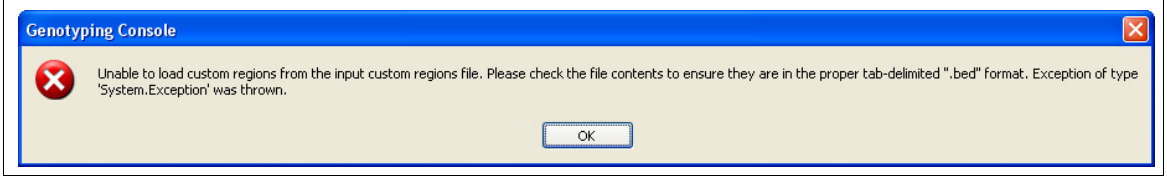
NOTE: An error message appears if you do not have24.annot.db for Human Mapping 100K or 500K arrays or the correct version of annot.db for SNP 6.0.

Figure 14.6 Progress bar



The following notice appears if the Custom Regions input file is not in the correct format (Figure 14.7).

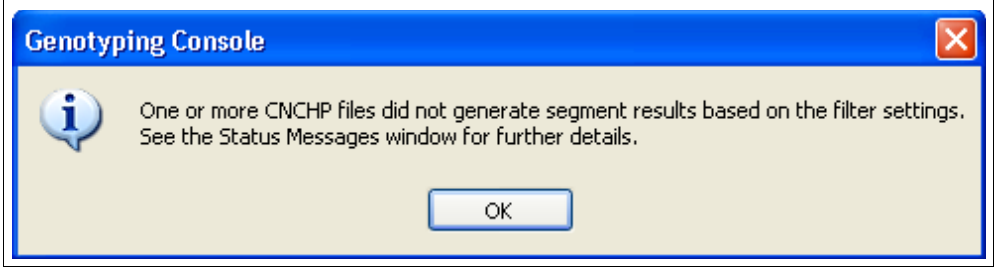
Figure 14.7 Notice of incorrect format



When the SRT has completed generating the segment report, then one of the following notices appears.

The following notice appears if none of the copy number data files had any Segments (Figure 14.8).

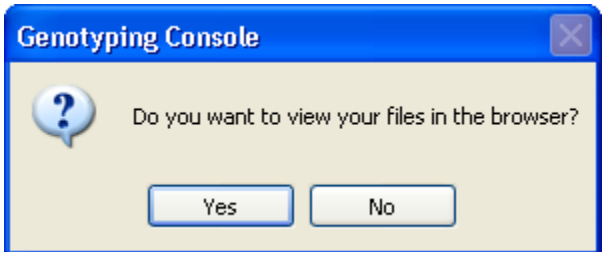
Figure 14.8 No Segments Results notice



The following notice (Figure 14.9)appears:

- If at least one of the copy number data files had Segments; or
- If you click **OK** in the notice above.

Figure 14.9 View Segment/CN/LOH data Notice



Click **Yes** to display the files in the GTC Browser.

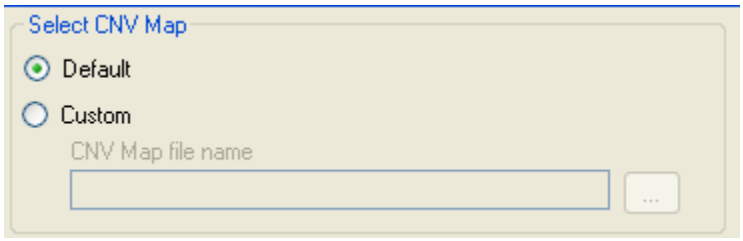
Selecting CNV Map for Filtering Segments

The SRT allows you to filter the detected copy number segments against a CNV Map of known copy number variation regions based on a specified percentage of overlap between the CN segments in the segment report and the user-selected CNV map. The SRT uses the Toronto CNV map as the default map for analysis. You can choose other CNV maps (e.g. Broad CNV map, or user-defined map) in the CMV Filters dialog box (Figure 14.10).

A CNV map template (Custom Regions template cn_input_regions.bed) is provided in the library folder.

When SNP and CN probe sets lose genome positions due to an annotation update, those SNP and CN probe sets are not included in the SRT to calculate % overlap.

Figure 14.10 CNV Map options in the SRT Filters dialog box (Figure 14.5)

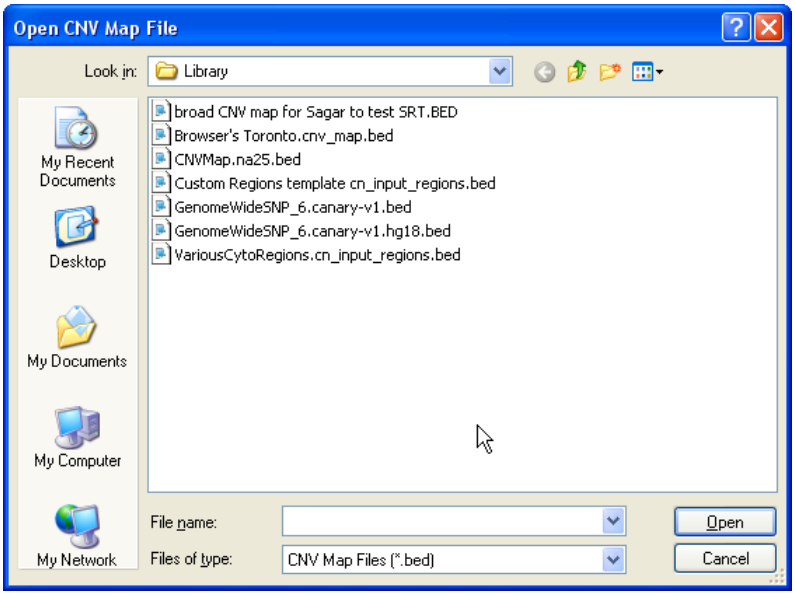


The CNV Map files use the BED file format.

To select a custom CNV map:

1. Select the Custom radio button.
2. Click the **Browse** button.
   The Open CNV Map File dialog box opens (Figure 14.11).

Figure 14.11 Open CNV Map File dialog box

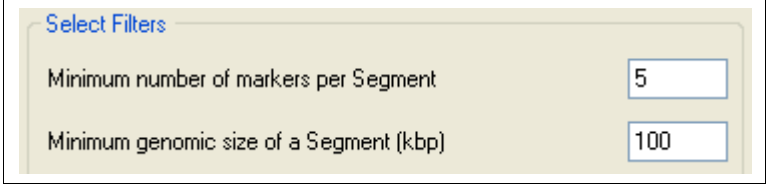


3. Select the CNV Map file from the list displayed in the dialog box and click **Open**.

Selecting Filters

You can define thresholds for the segment size and number of markers required to define segments (Figure 14.12).

Figure 14.12 Select Filters in the SRT Filters dialog box (Figure 14.5)



To set thresholds:

- Enter values for the filter parameters:

Minimum number of markers per segment — Minimum number of SNP and CNV probe sets that must be present to report the segment.

Minimum genomic size of a segment (kpb) — Minimum size of a segment in kilobase pairs.

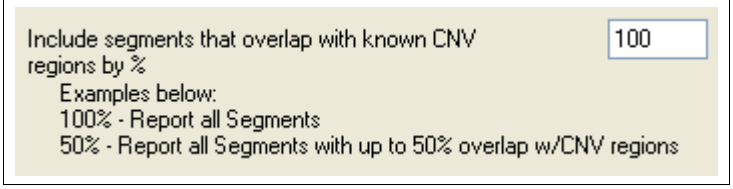
Include segments that overlap with known CNV regions by %

CNV regions from the CNV map files are the known (or user-defined) regions in the human genome identified as having copy number variants (CNVs), or copy number polymorphisms (CNPs). This data comes from the Toronto DGV database (or is user defined) and can be displayed in the Browser track.

To aid in the discovery of novel copy number variant regions, it is possible to exclude segments that overlap these regions with known copy number variants (Figure 14.13). The identification of segments to be excluded is based on the percentage of the markers (SNP+CN makers) that overlap the boundaries of the SNP and CN annotation in the database.

If the percentage of markers in a copy number changed segment which overlap with known CNV regions in the Toronto DGV database (or users defined CNV regions) exceeds the selected percentage, then the Segment Reporting Tool will not report that segment.

Figure 14.13 Overlap Filter setting in the SRT filters dialog box (Figure 14.5)



To set the threshold:

- Enter values for the % Overlap:

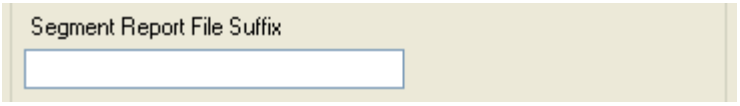
  If the percent value is set to 25%, segments with up to 25% of their markers overlapping known CNV regions **will be reported as part of the Segment Report**. Segments with more than 25% of their markers overlapping known CNV regions **will be excluded** from the Segment Report.

NOTE: 100% means that all of the segments will be reported.

Adding a Suffix to the Segment Report Files

A suffix can be added to keep the output files from overwriting the results of an earlier analysis (Figure 14.14). The suffix will be added to Segment Report files and Custom Region Report files. Suffixes are not added to Summary reports.

Figure 14.14 Add a segment report file suffix in the SRT Filter dialog box (Figure 14.5)



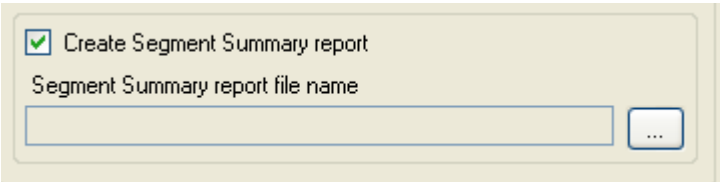
To add a suffix:

- Enter a suffix for the segment report file in the Segment Report File Suffix textbox.

Create Segment Summary Report File

This option (Figure 14.15) allows you to generate a summary segment report with information on changed regions in all the CNCHP files you are analyzing. This tab-delimited file contains the file extension .cn_segments_summary.

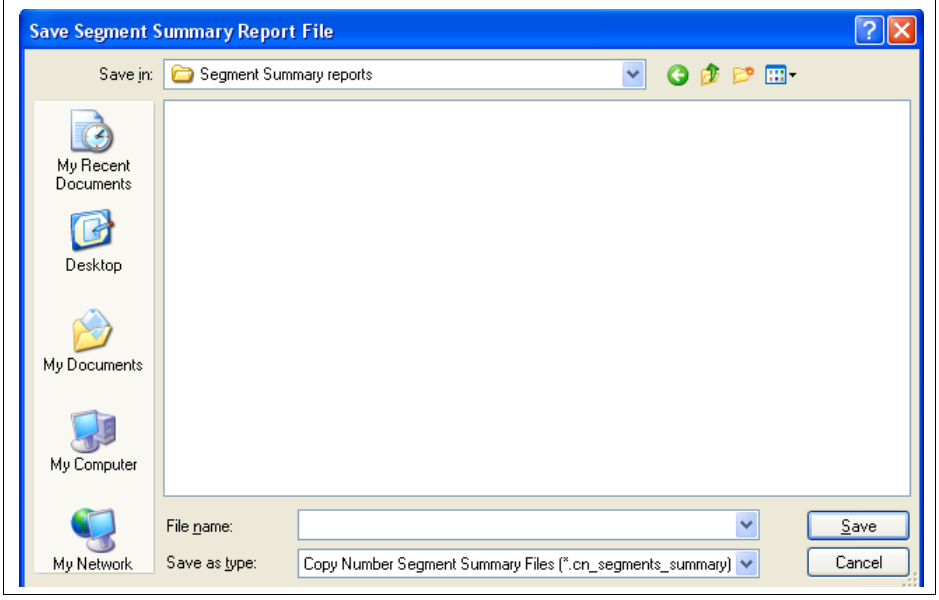
Figure 14.15 Segment Summary report file option in the SRT Filters dialog box (Figure 14.5)



To create a segment summary report file:

1. Select the Create segment report summary checkbox (Figure 14.15).
2. Click the **Browse** button ... .
   The Save Segment Summary Report File dialog box opens (Figure 14.16).

Figure 14.16 Save Segment Summary Report File dialog box



3. Select a location for the Segment Summary Report file and enter a name for the file. (Segment Report File suffixes are not automatically added to Summary files).
4. Click **Save** in the Save Segment Summary Report File dialog box.

Using a Custom Regions File

You can use a Custom Regions file (Figure 14.17) to look for copy number gain and loss in select regions of the genome. Custom Regions are defined in tab-delimited ".bed" format files.

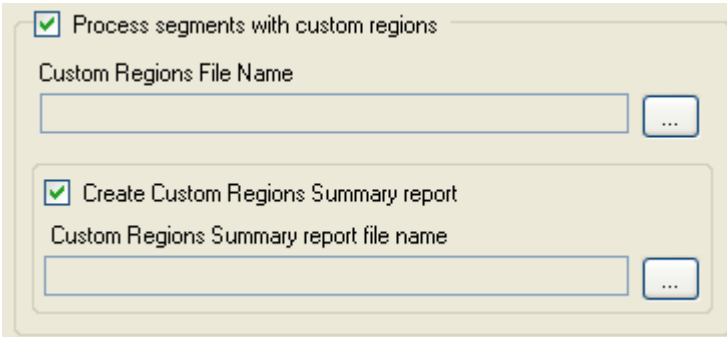
If a Custom Regions file has been selected, the report tool generates:

- Custom Regions Report file (.custom_regions) for each copy number file.
- Custom Regions Summary Report file (.custom_regions_summary) concatenating all the data for all files run at a time (optional).

The Custom Region Summary Report is the result of filtering the whole genome Copy Number Segment data generated by the Segment Reporting Tool for only the regions defined in the custom Regions file, during the same run for the same samples.

A sample template Custom Input Regions file (Custom Regions template cn_input_regions.bed) is located in the Library folder.

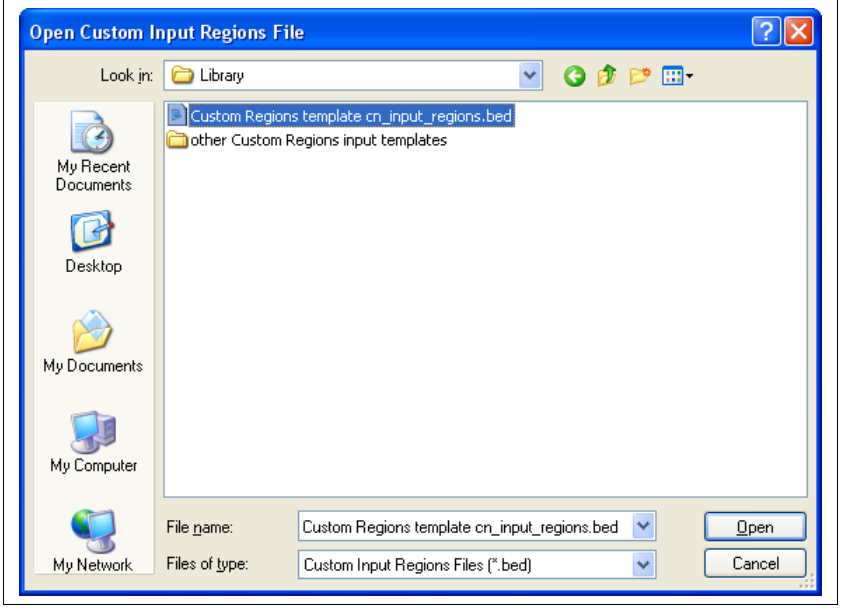
Figure 14.17 Options for using a custom regions file and creating a Custom Regions Summary report in the SRT Filters dialog box (Figure 14.5)



To select a custom regions file and generate custom regions reports:

1. Select the **Process segments with custom regions** checkbox.
2. Click the **Browse** button ... .
   The Open Custom Input Regions File dialog box opens (Figure 14.18).

Figure 14.18 Open Custom Input Regions File dialog box



3. Select the Custom Input Regions .bed file and click **Open**.
   You can proceed with the generation of the reports.
   Custom Region Reports are generated for each copy number file.

Custom Region File Format

The Segment Reporting Tool also allows generation of a Custom Regions Report (*.custom_regions). Custom Regions are any regions of the genome defined by coordinates entered into a text file in tab-delimited ".bed" format, as described in *http://genome.ucsc.edu/FAQ/FAQformat#format1*.

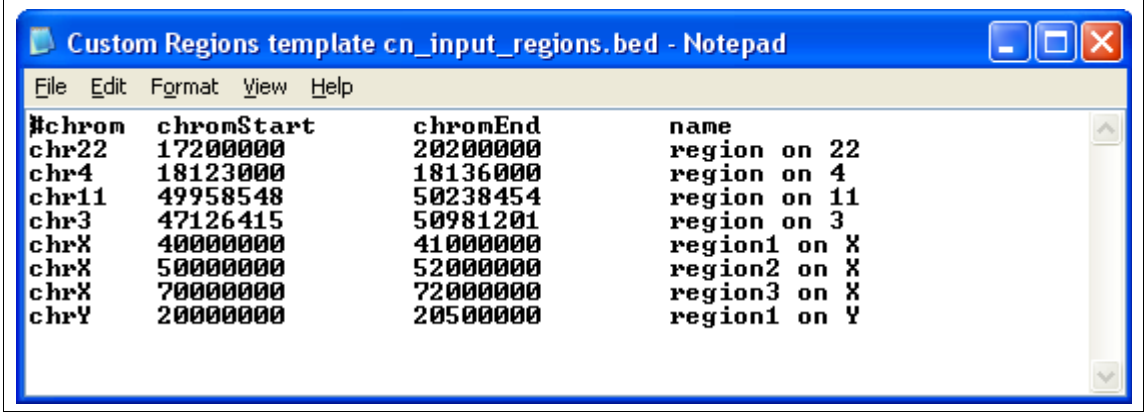
The Custom Regions Report that results from processing Segment for Custom Regions contains copy number gain and loss segment and CNV overlap information about just the defined regions.

You can use a Custom Regions Segments file (Figure 14.19) to look for copy number gain and loss in select regions of the genome. Custom Regions are defined in tab-delimited ".bed" file format, with columns for:

- chromosome
- custom region start position
- custom region stop position
- custom region name

The header lines marked with the # symbol are ignored.

Figure 14.19 Custom Input Regions File



Custom Regions template cn_input_regions.bed files can:

- Serve as custom region for SRT
- Serve as custom CNV map for SRT
- Be loaded into heat map viewer
- Be loaded into GTC browser
- Be loaded into other browsers such as USCS genome browser

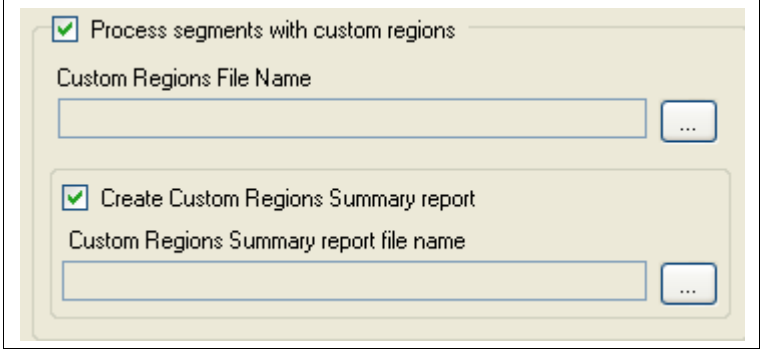
Create Custom Regions Summary report

This option allows you to generate a summary segment report with information on CN change segments for select regions in all the CNCHP files you are analyzing. This tab-delimited file uses the file extension `.custom_regions_summary`.

To create a custom regions summary report file:

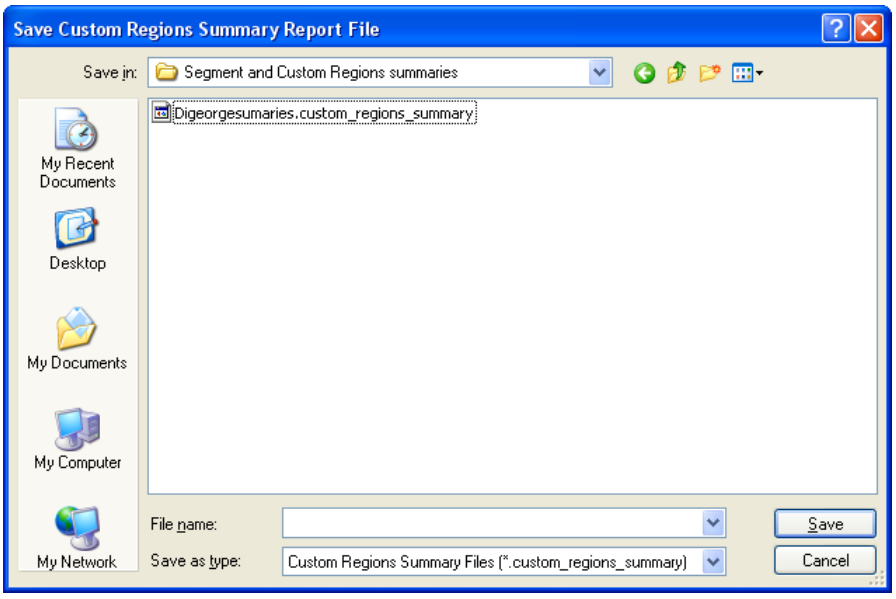
1. Select the **Create custom region summary report** checkbox (Figure 14.20).

Figure 14.20 Options for using a custom regions file and creating a Custom Regions Summary report in the SRT Filters dialog box (Figure 14.5)



2. Click the **Browse** button ... .
   The Save Custom Regions Summary Report File dialog box opens (Figure 14.21).

Figure 14.21 Save Custom Regions Summary Report File dialog box



3. Select a location for the Custom Regions Summary Report file and enter a name for the file. (Segment Report File suffixes are not automatically added to Summary files)
4. Click **Save** in the Save Custom Regions Summary Report File dialog box.
   You can proceed with the generation of the reports.
   Custom Region Reports are generated for each copy number file.

Segment Report Tool Results Files

The Segment Report Tool can produce the following types of report files:

- Segment Report file (.cn_segments) for each copy number file.
- Segment Summary Report file (.cn_segments_summary) concatenating all the data for all files run at a time (optional).

If a Custom Regions file has been selected, the report tool generates:

- Custom Regions Report file (.custom_regions) for each copy number file.
- Custom Regions Summary Report file (.custom_regions_summary) concatenating all the data for all files run at a time (optional).

CN segment and custom region files are automatically saved with CNCHP files in the same CN result folder; segment summary and custom region summary files can be saved manually.

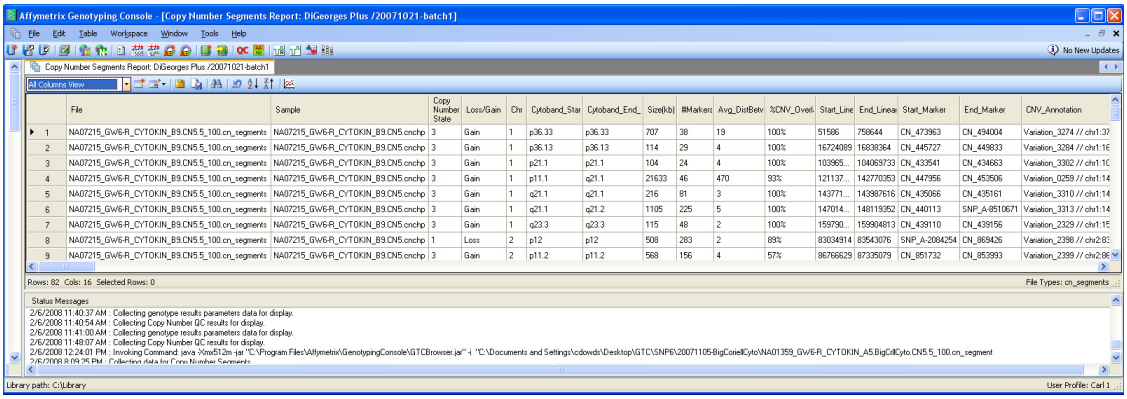
Segment Report File

The Segment Report files (Figure 14.22) contain information on the copy number segments detected in a given CNCHP file.

The Segment Report files can be displayed in the GTC Browser in the Karyoview and as an annotation track in the Chromosome View.

- Segment Data files for Human Mapping 100K/500K arrays have the CN4.cn_segments extension.
- Segment Data files for the Genome-Wide Human SNP Array 6.0 have the CN5.cn_segments extension.

Figure 14.22 Segment Reports

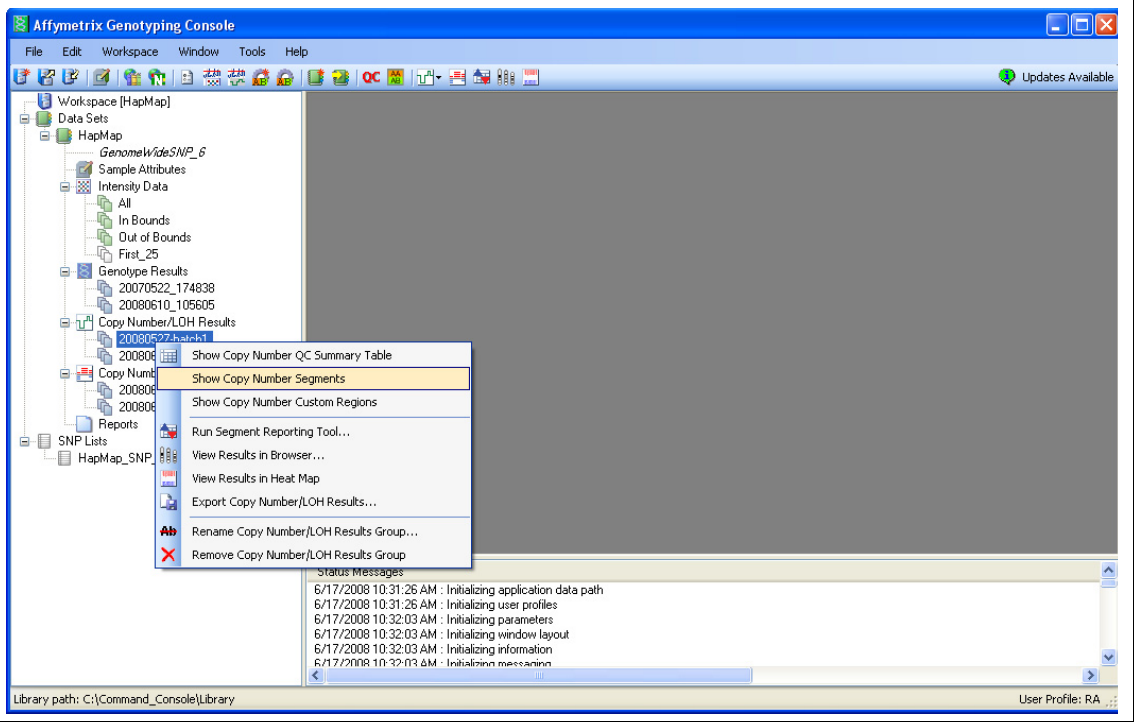


To View the Segment Report from Genotyping Console:

1. In the Genotyping Console data tree, select a Copy Number/LOH Results group for which you have previously generated Segment Reports using the Segment Reporting Tool.

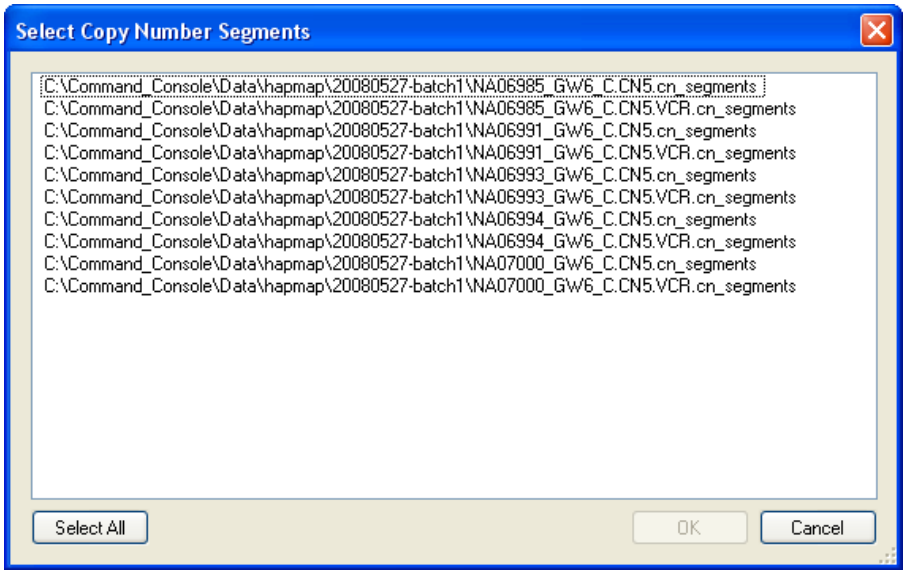
   To do this: Right-click on a Copy Number/LOH Results group and choose **Show Copy Number Segments** (Figure 14.23).

Figure 14.23 Show Copy Number Segments



The Select Copy Number Segments dialog box appear (Figure 14.24).

Figure 14.24 Select Copy Number Segments dialog box



2. Select Copy Number Segments files (*.cn_segments) from the list and Click **OK**.
   The Segment Reports for all chosen files open in a single list in the display area.

NOTE: Segment report information can also be viewed in the GTC Browser. See *Loading Data into the GTC Browser* on page 305.

The Copy Number Segments Report table content changes if you are using a custom map in the "Copy Number Segments Report" table.

Starting from GTC 3.0.1, "%CNV_Overlap" is replaced with the %CNV_Overlap numbers calculated from the custom map and "CNV_Annotation" is replaced with variations names from the custom map.

The table contains the following items:

File	Name of the segment data file (seen in GTC table view only).
Sample	CNCHP file name.
Copy Number State	Per marker CN as estimated by the HMM.
Loss/Gain	Whether the copy number change is a decrease or increase from the expected normal value.
Chr	Chromosome where the segment is located.
Cytoband_Start_Pos	The Chromosome's cytoband within which a copy number change segment begins.
Cytoband_End_Pos	The Chromosome's cytoband within which a copy number change segment ends.
Size (kb)	Size of the segment of copy number change.
#Markers	Number of SNPs+CNV markers within the segment.
Avg_DistBetweenMarkers(kb)	Length of segment divided by number of markers encompassed by that segment.

%CNV_Overlap	Percentage of markers in a segment which overlap the boundaries of a known CNV.
Start_Linear_Pos	Base pair position on the Chromosome at which the first marker in the segment begins (going from top of the p-arm to the bottom of the q-arm of the chromosome).
End_Linear_Position	Base pair position on the Chromosome at which the last marker in the segment begins (going from top of the p-arm to the bottom of the q-arm of the chromosome).
Start_Marker	Name of the first SNP or CN marker of a copy number change segment.
End_Marker	Name of the last SNP or CN marker of a copy number change segment.
CNV_Annotation	Information from the Toronto Database of Genomic Variants about the CNV variants which overlap the Copy Number change segment (or Genomic Variants annotation information from other database if a custom map is used).

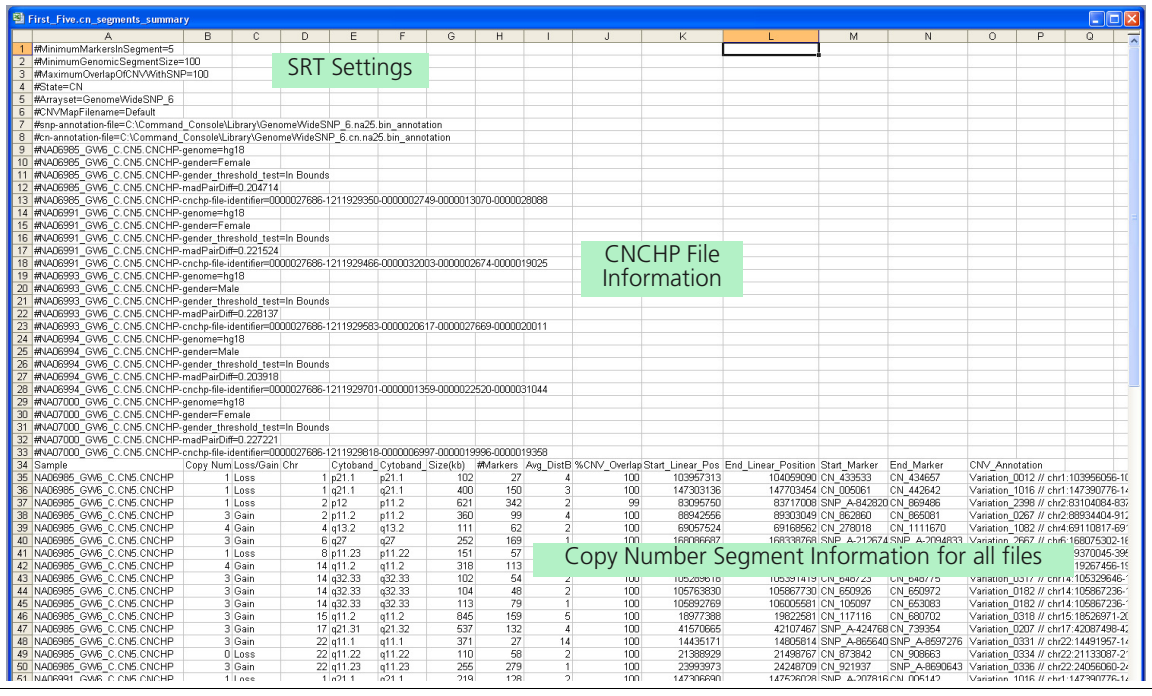
Segment Summary Report

The segment summary report (Figure 14.25) has information for every cn segment from the whole batch while the segment report only has cn segment info from one .CNCHP file. The header information is also concatenated to include data on all the .CNCHP files.

The summary report can't be displayed in the Browser; it can be viewed in a spreadsheet program.

You will be directed to specify a name and location for the Segment Summary Report file before performing the analysis.

Figure 14.25 Summary report displayed in spreadsheet



The Segment Summary report contains:

- Information on SRT settings and CNCHP files analyzed in the header.
- Copy Number Segment information (same as in Segment Report)

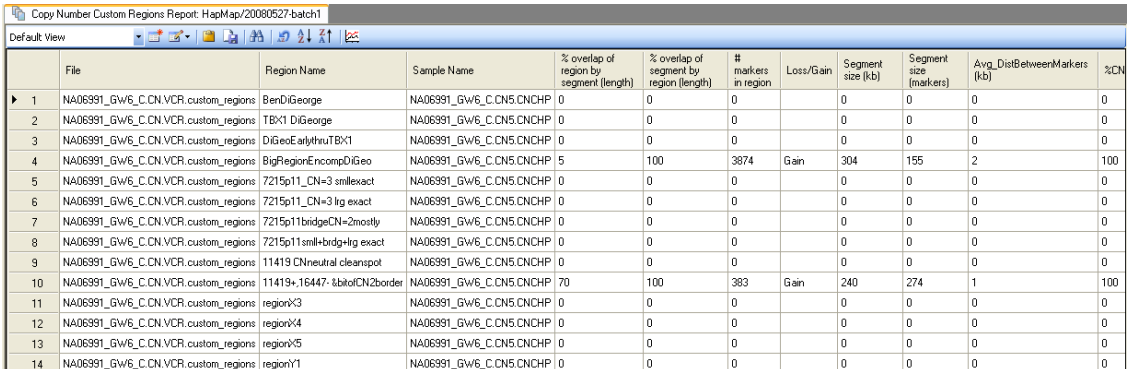
Custom Regions Report

The Custom Regions Report files (Figure 14.26) contain information on the copy number segments detected in the custom regions designated in the Custom Region file for a given CNCHP file. Each Segment overlapping a Region generates one row in the table. Regions with no overlapping Segments in a sample are represented as a single row in the table with the Loss/Gain column not populated.

- Custom Regions Segment Data files for Human Mapping 100K/500K arrays have the CN4.custom_regions extension

Custom Regions Segment Data files for the Genome-Wide Human SNP Array 6.0 have the CN5.custom_regions extension

Figure 14.26 Custom Regions table in displayed in GTC


Copy Number Custom Regions Report: HapMap/20080527-batch1

Default View

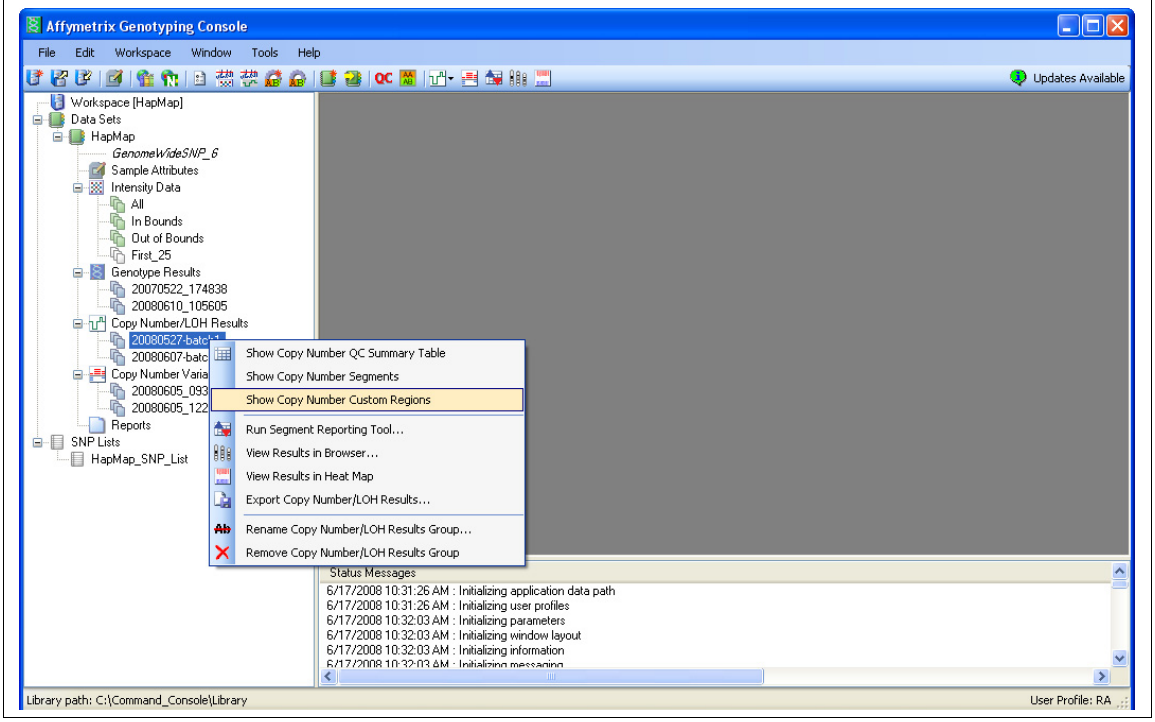
	File	Region Name	Sample Name	% overlap of region by segment (length)	% overlap of segment by region (length)	# markers in region	Loss/Gain	Segment size (kb)	Segment size (markers)	Avg_DistBetweenMarkers (kb)	%CN
1	NA06991_GW6_C.CN.VCR.custom_regions	BenDiGeorge	NA06991_GW6_C.CN5.CNCHP	0	0	0		0	0	0	0
2	NA06991_GW6_C.CN.VCR.custom_regions	TBX1 DiGeorge	NA06991_GW6_C.CN5.CNCHP	0	0	0		0	0	0	0
3	NA06991_GW6_C.CN.VCR.custom_regions	DiGeoEarlythruTBX1	NA06991_GW6_C.CN5.CNCHP	0	0	0		0	0	0	0
4	NA06991_GW6_C.CN.VCR.custom_regions	BigRegionEncompDiGeo	NA06991_GW6_C.CN5.CNCHP	5	100	3874	Gain	304	155	2	100
5	NA06991_GW6_C.CN.VCR.custom_regions	7215p11_CN=3 smllexact	NA06991_GW6_C.CN5.CNCHP	0	0	0		0	0	0	0
6	NA06991_GW6_C.CN.VCR.custom_regions	7215p11_CN=3 lrg exact	NA06991_GW6_C.CN5.CNCHP	0	0	0		0	0	0	0
7	NA06991_GW6_C.CN.VCR.custom_regions	7215p11bridgeCN=2mostly	NA06991_GW6_C.CN5.CNCHP	0	0	0		0	0	0	0
8	NA06991_GW6_C.CN.VCR.custom_regions	7215p11smll+brdg+lrg exact	NA06991_GW6_C.CN5.CNCHP	0	0	0		0	0	0	0
9	NA06991_GW6_C.CN.VCR.custom_regions	11419 CNneutral cleanspot	NA06991_GW6_C.CN5.CNCHP	0	0	0		0	0	0	0
10	NA06991_GW6_C.CN.VCR.custom_regions	11419+,16447- &bitofCN2border	NA06991_GW6_C.CN5.CNCHP	70	100	383	Gain	240	274	1	100
11	NA06991_GW6_C.CN.VCR.custom_regions	regionX3	NA06991_GW6_C.CN5.CNCHP	0	0	0		0	0	0	0
12	NA06991_GW6_C.CN.VCR.custom_regions	regionX4	NA06991_GW6_C.CN5.CNCHP	0	0	0		0	0	0	0
13	NA06991_GW6_C.CN.VCR.custom_regions	regionX5	NA06991_GW6_C.CN5.CNCHP	0	0	0		0	0	0	0
14	NA06991_GW6_C.CN.VCR.custom_regions	regionY1	NA06991_GW6_C.CN5.CNCHP	0	0	0		0	0	0	0

To View A Custom Region Report in Genotyping Console:

1. In the Genotyping Console data tree, select a Copy Number/LOH Results group for which you have previously generated Custom Region Reports using the Segment Reporting Tool.

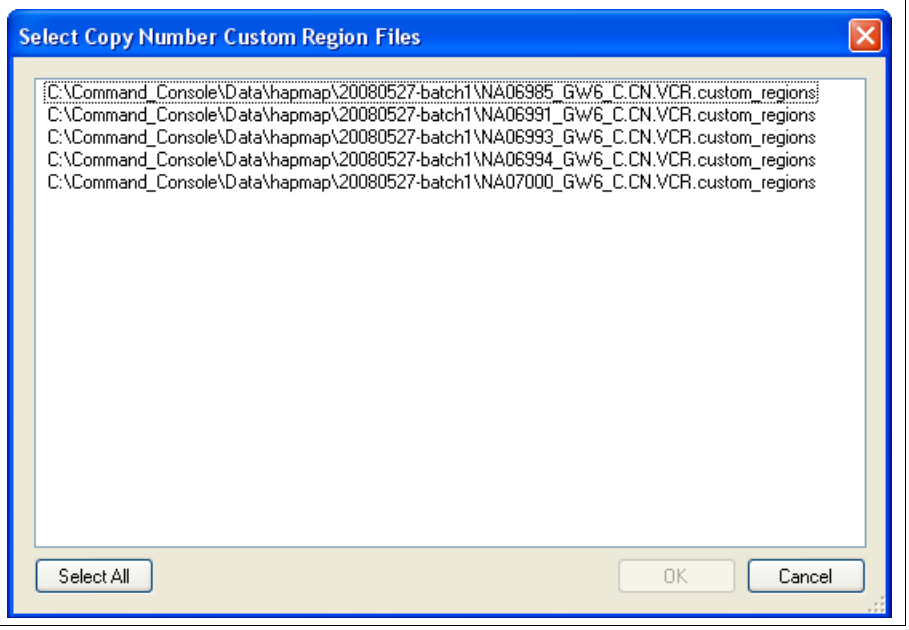
   Right-click on a Copy Number/LOH Results group and choose **Show Copy Number Custom Regions** (Figure 14.27).

Figure 14.27 Select Custom Regions



The Select Copy Number Custom Region Files dialog box opens (Figure 14.28).

Figure 14.28 Select Copy Number Custom Region Files



2. Select Copy Number Custom Regions files (*.custom_regions) from the list and click **OK.**

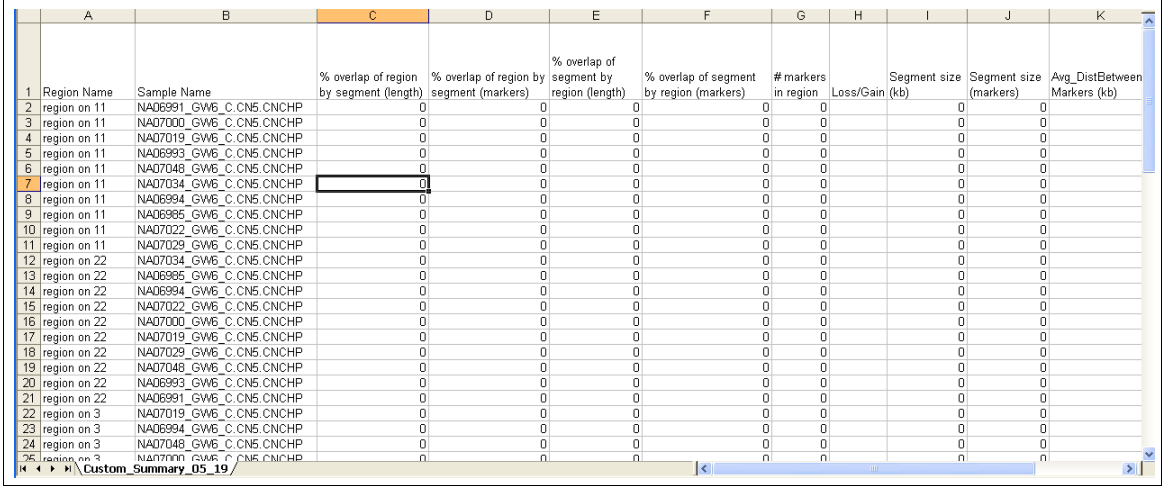
The Custom Regions for all selected files opens in a single list and displays the following information:

File	Custom Regions report file name.
Region Name	Region name from Custom Input Regions "*.bed" file.
Sample	CNCHP File name.
% overlap of region by segment (length)	Percentage of overlap of the Custom Region by any one segment in the region, as measured by length. Segments as large or larger than a Region will have a value of "100"
% overlap of segment by region (length)	Percentage of overlap of the Segment by the Region, as measured by length. Regions as large or larger than overlapping Segments will have a value of "100"
# markers in region	Number of SNPs+CNV markers within the region.
Loss/Gain	Whether the Copy number change is a decrease or increase from the expected normal value.
Segment size (kb)	Size of the segment of Copy Number change as measured in kilobase pairs.
Segment size (markers)	Size of the segment of Copy Number change as measured in total number of SNP + CN markers.
Avg_DistBetweenMarkers (kb)	Length of segment divided by number of markers encompassed by that segment.
%CNV_Overlap	Percentage of markers in the segment which overlap the boundaries of a known CNV.
Chromosome	Chromosome where the Region and Segment are located.
Cytoband_Start_Pos	The Chromosome's cytoband within which a Copy Number change segment begins.
Cytoband_End_Pos	The Chromosome's cytoband within which a Copy Number change segment ends.
Start_Linear_Pos	The base pair position on the Chromosome at which the first marker in the segment begins (going from top of the p-arm to the bottom of the q-arm of the chromosome).
End_Linear_Position	The base pair position on the Chromosome at which the last marker in the segment begins (going from top of the p-arm to the bottom of the q-arm of the chromosome).
Region start	The base pair position on the Chromosome at which the Custom Region begins (going from top of the p-arm to the bottom of the q-arm of the chromosome).
Region end	The base pair position on the Chromosome at which the Custom Region ends (going from top of the p-arm to the bottom of the q-arm of the chromosome).

Custom Regions Summary Report

The summary report cannot be displayed in the Browser; it must be viewed in a spreadsheet program, such as Excel (Figure 14.29). You will be directed to specify a name and location for the Segment Summary Report file before performing the analysis.

Figure 14.29 Custom Regions Summary Report in Excel



	A	B	C	D	E	F	G	H	I	J	K
1	Region Name	Sample Name	% overlap of region by segment (length)	% overlap of region by segment (markers)	% overlap of segment by region (length)	% overlap of segment by region (markers)	# markers in region	Loss/Gain	Segment size (kb)	Segment size (markers)	Avg_DistBetween Markers (kb)
2	region on 11	NA06991_GW6_C.CN5.CNCHP	0	0	0	0	0		0	0	
3	region on 11	NA07000_GW6_C.CN5.CNCHP	0	0	0	0	0		0	0	
4	region on 11	NA07019_GW6_C.CN5.CNCHP	0	0	0	0	0		0	0	
5	region on 11	NA06993_GW6_C.CN5.CNCHP	0	0	0	0	0		0	0	
6	region on 11	NA07048_GW6_C.CN5.CNCHP	0	0	0	0	0		0	0	
7	region on 11	NA07034_GW6_C.CN5.CNCHP	0	0	0	0	0		0	0	
8	region on 11	NA06994_GW6_C.CN5.CNCHP	0	0	0	0	0		0	0	
9	region on 11	NA06985_GW6_C.CN5.CNCHP	0	0	0	0	0		0	0	
10	region on 11	NA07022_GW6_C.CN5.CNCHP	0	0	0	0	0		0	0	
11	region on 11	NA07029_GW6_C.CN5.CNCHP	0	0	0	0	0		0	0	
12	region on 22	NA07034_GW6_C.CN5.CNCHP	0	0	0	0	0		0	0	
13	region on 22	NA06985_GW6_C.CN5.CNCHP	0	0	0	0	0		0	0	
14	region on 22	NA06994_GW6_C.CN5.CNCHP	0	0	0	0	0		0	0	
15	region on 22	NA07022_GW6_C.CN5.CNCHP	0	0	0	0	0		0	0	
16	region on 22	NA07000_GW6_C.CN5.CNCHP	0	0	0	0	0		0	0	
17	region on 22	NA07019_GW6_C.CN5.CNCHP	0	0	0	0	0		0	0	
18	region on 22	NA07029_GW6_C.CN5.CNCHP	0	0	0	0	0		0	0	
19	region on 22	NA07048_GW6_C.CN5.CNCHP	0	0	0	0	0		0	0	
20	region on 22	NA06993_GW6_C.CN5.CNCHP	0	0	0	0	0		0	0	
21	region on 22	NA06991_GW6_C.CN5.CNCHP	0	0	0	0	0		0	0	
22	region on 3	NA07019_GW6_C.CN5.CNCHP	0	0	0	0	0		0	0	
23	region on 3	NA06994_GW6_C.CN5.CNCHP	0	0	0	0	0		0	0	
24	region on 3	NA07048_GW6_C.CN5.CNCHP	0	0	0	0	0		0	0	
25	region on 3	NA07000_GW6_C.CN5.CNCHP	0	0	0	0	0		0	0	

Custom_Summary_05_19

The File contains Custom Regions Segment information, organized by Region Name, with the same information on regions as in the Segment Report.

Loading Data into the GTC Browser

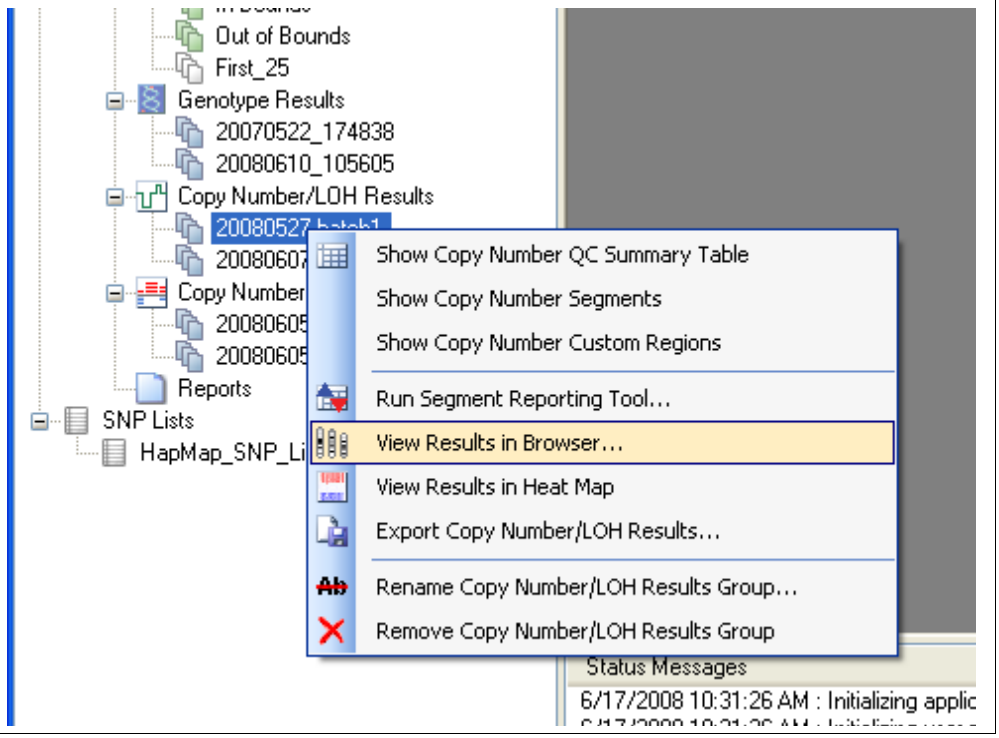
NOTE: When the Segment Reporting Tool is finished running the option to open the generated files in the Browser is provided.

NOTE: If you generated Custom Regions, you can load the cn_input_regions.bed file into the Browser using the **File > Open** menu to see your custom regions displayed as an Annotation track in the Chromosome View.

Displaying copy number data in the GTC browser:

1. In the Genotyping Console data tree, select the copy number data you wish to display (Figure 14.30).

Figure 14.30 Selecting results in the GTC data tree



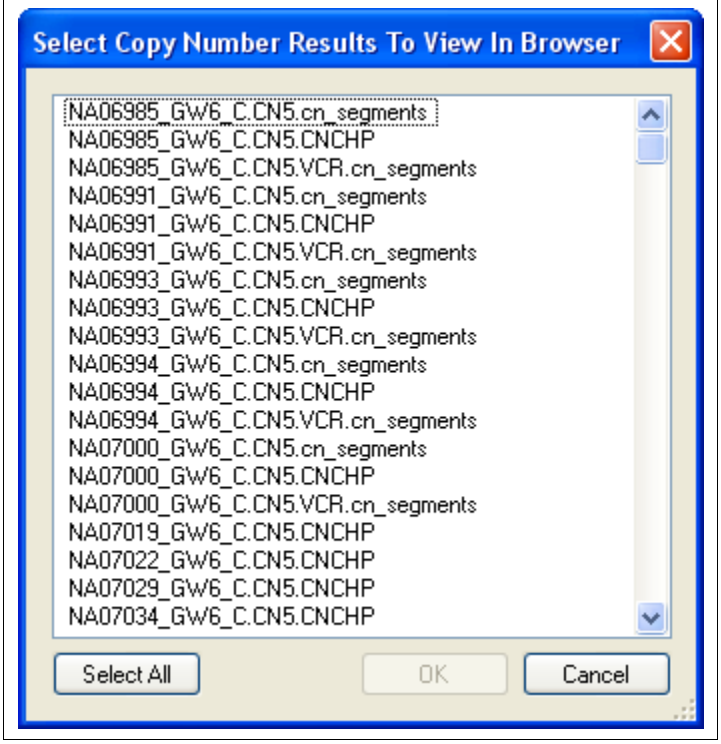
2. Right-click on the Results Group and select **View Results in Browser** from the context-sensitive menu; or

   From the Workspace menu, select **Copy Number/LOH Results > View Results in Browser**; or

   In the tool bar, click the **View Results in Browser** button.

   The Select Copy Number Results dialog box opens (Figure 14.31).

Figure 14.31 Select Copy Number Results dialog box



This dialog box displays a list of the results data available in the selected Results set.

You can select the following types of results for display:

- Segment Data files (.cn_segments)
- Copy Number Data files (.cnchp)
- LOH Data Files (.lohchp)

NOTE: Not all the file types may be available depending upon the type of array used.

3. Select the files you wish to view; or click **Select All**.
4. Click **OK**.

   The GTC browser opens and displays the data, along with the default annotation files.

   See the GTC Browser User Manual for more information.

NOTE: To compare results from different analysis runs, use the file open functionality with the Browser to open the files.

Export Copy Number/LOH data

The copy number/LOH data can be exported as tab-delimited text file that can be imported into other software.

NOTE: You can also export data in different formats in the GTC Browser (see the GTC Browser manual for more information).

NOTE: Annotation files (*.annot.db) are required to include dbSNP RS ID in the export option for CNCHP files generated in earlier versions of GTC. For Human Mapping 100K or 500K data, the export will require na24 version of the annotation file (*.na24.annot.db). For Genome-Wide SNP Array 6.0 data, the export will require na25 to na29 version of annotation files (*.annot.db), depending on the annotation version that was used to generated the CNCHP file.

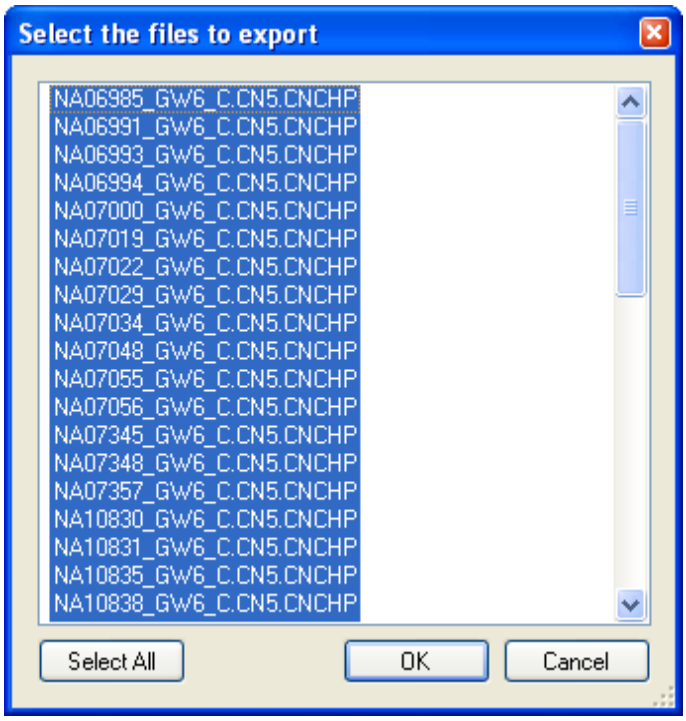
To export data:

1. Select the data set that you wish to export in the tree.
2. From the Workspace menu, select **Copy Number/LOH Results > Export Copy Number/LOH Results**; or

   Right-click the Copy Number/LOH data set and select **Export Copy Number/LOH Results** from the pop-up menu.

   The Select files for export dialog box opens (Figure 14.32).

Figure 14.32 Select files for export dialog box



3. Select the files to export from the list and click **OK**.

NOTE: You can click Select All to select all files in the list.

The Select Columns to Export dialog box opens.

The exact list of items will depend upon the array type used to generate the data (Figure 14.33, Figure 14.34).

Figure 14.33 Select columns for export (SNP6)

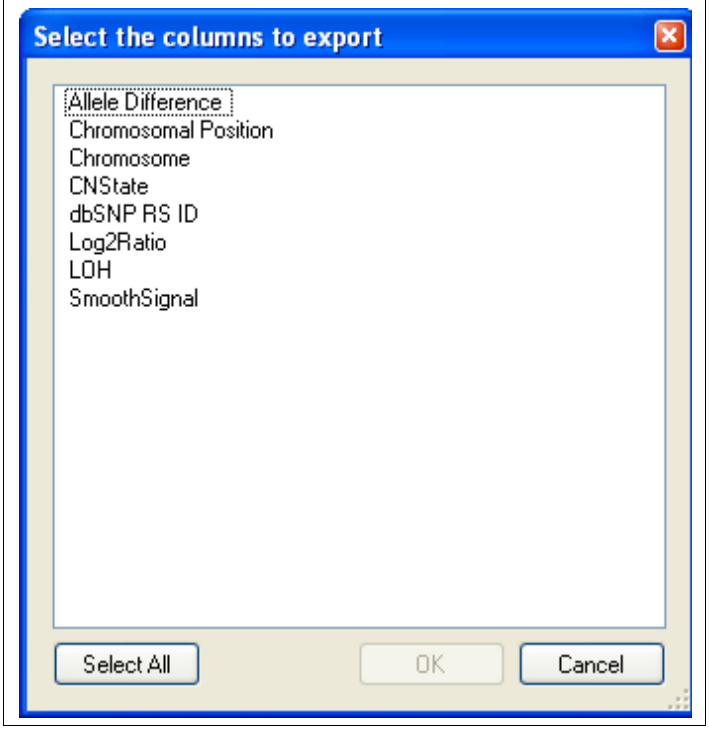
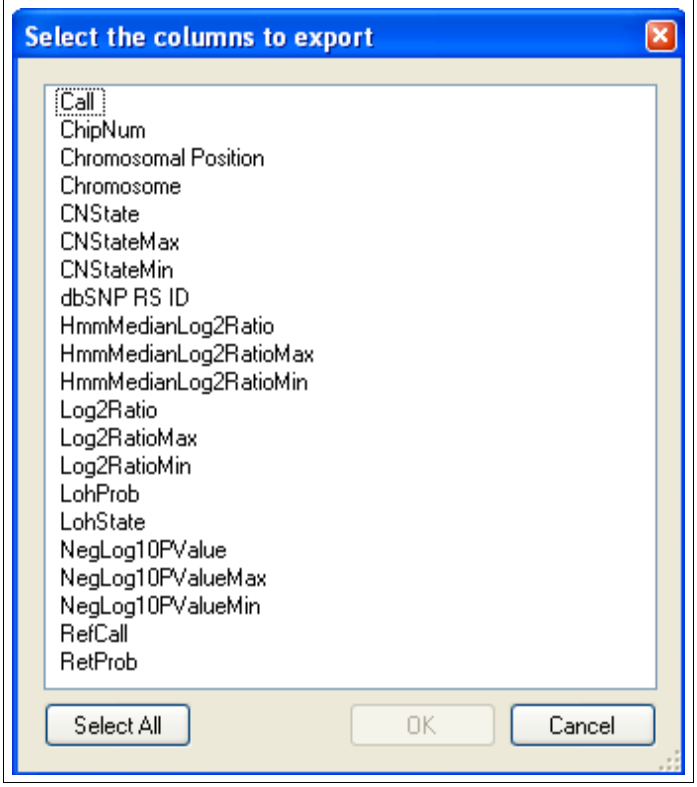


Figure 14.34 Select columns for CN/LOH export (100K/500K)



The data in these columns is described in:

- *Copy Number/LOH File Format for Human Mapping 100K/500K Array Data* on page 221
- *Copy Number/LOH Data File Format for Genome-Wide Human SNP Array 6.0 Data* on page 260

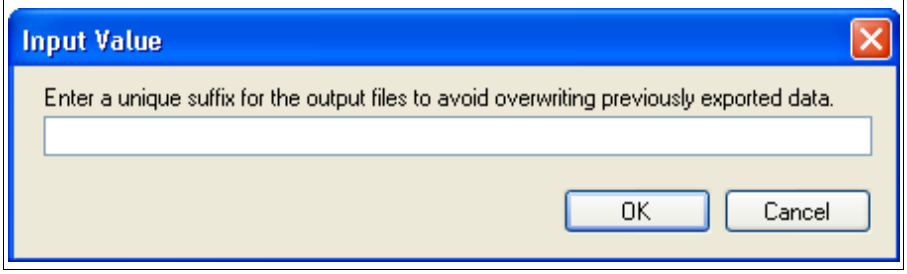
NOTE: not all of these columns may be available, depending upon whether or not you are exporting CN data, LOH data, or both.

4. Select the data to export and click **OK**.

NOTE: You can click Select All to select all data types in the list.

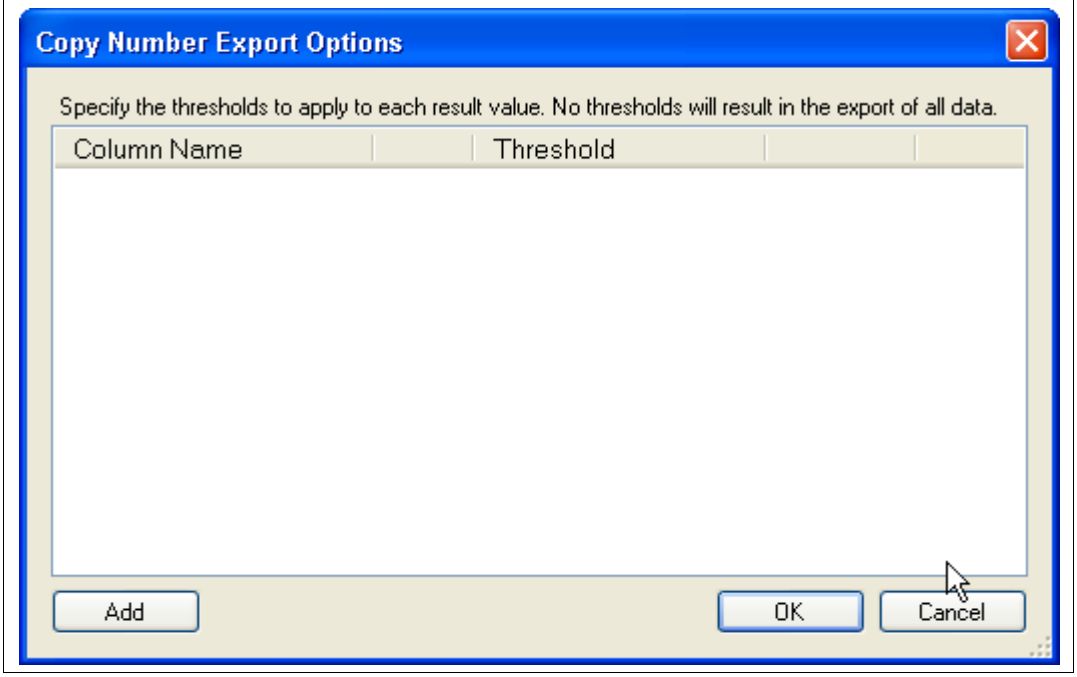
An input dialog box opens enabling you to enter a suffix to be applied to the default file name so that previously exported results will not be overwritten (Figure 14.35).

Figure 14.35 Input Value dialog box



5. Enter a suffix and select **OK**.
The Export Options dialog box opens (Figure 14.36).

Figure 14.36 Copy Number Export Options dialog box

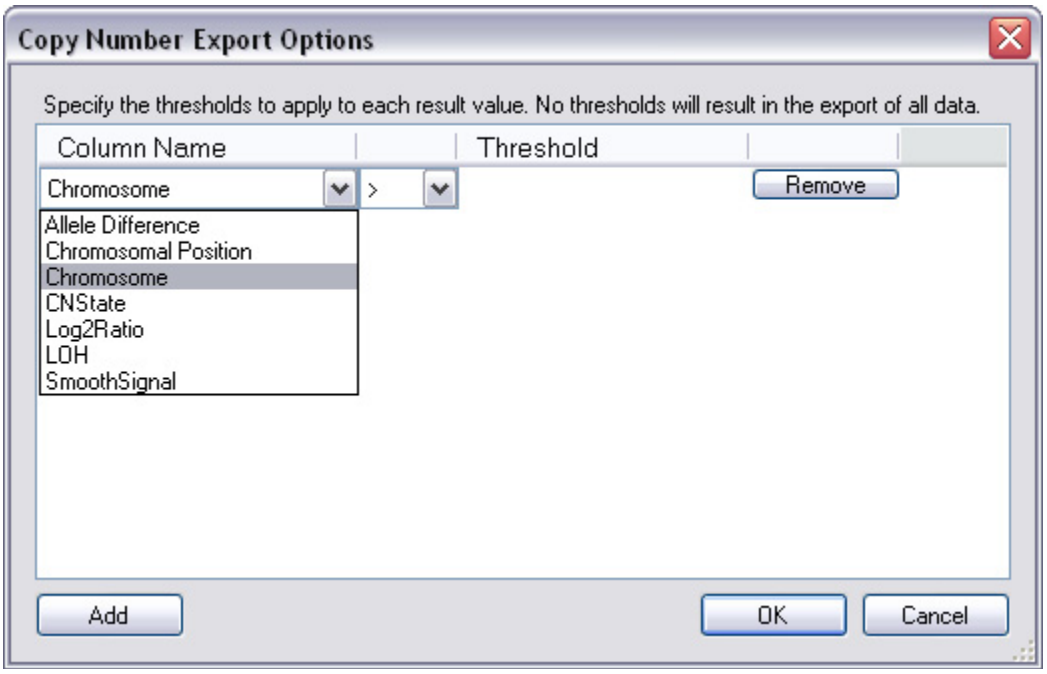


The export options dialog box allows you to filter the output using different parameters.

To add a threshold:

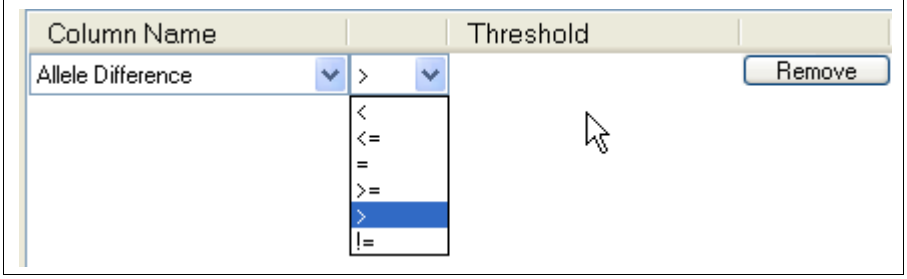
6. Click the Add button.
   A row appears in the table with drop-down lists.
7. Select the parameter that will be used to filter the output from the Column Name list (Figure 14.37)
   It is possible to filter on any of the exported columns, Chromosome and Position.

Figure 14.37 Selecting a parameter for filtering output



8. Select the comparison operator (Figure 14.38).
   - less than (<)
   - less than or equal to (?)
   - greater than (>)
   - greater than or equal to (?)
   - equal to (=)
   - not equal to (!=)

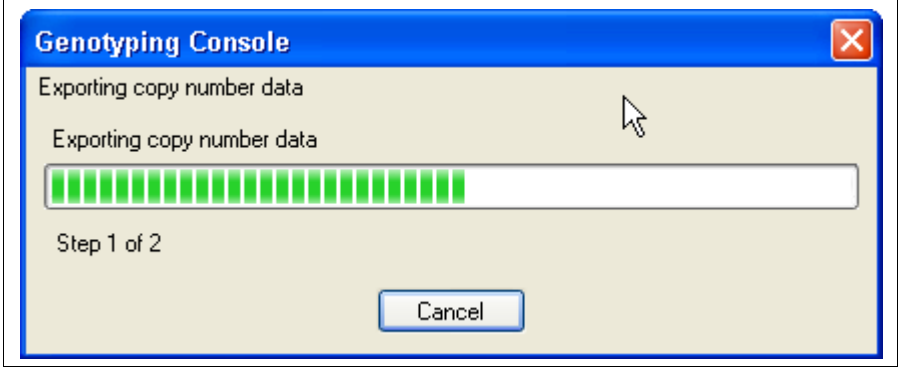
Figure 14.38 Selecting a comparison operator



9. Enter a value for the threshold parameter.
10. Repeat the above steps to filter on different parameters
11. Click **OK**.

The Progress bar displays the progress of the export (Figure 14.39).

Figure 14.39 Progress bar



The export process creates a text file using a name based on the .cnchp file names, with a .txt extension. The file is placed in the same directory used for the Copy Number/LOH Results group.

Setting QC Thresholds

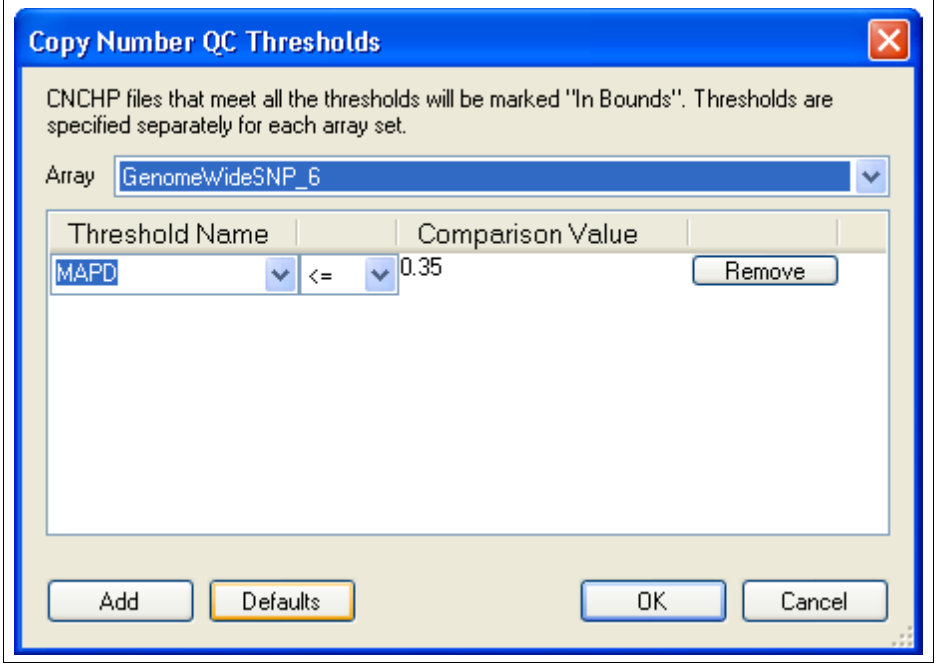
Files that exceed the QC thresholds set in this dialog box will be flagged in the Copy Number/LOH QC table as out of bounds.

Genotyping Console maintains default thresholds for copy number QC metrics, and will highlight in the copy number QC tables the metrics that are outside of the threshold values. You can modify the QC thresholds as needed.

To modify the QC threshold options:

1. Click on the **Copy Number QC Thresholds** button on the main tool bar, or
   From the Edit menu, select **Copy Number QC Thresholds**.
   The Copy Number QC Thresholds dialog box appears (Figure 14.40).

Figure 14.40 Copy Number QC Thresholds dialog box

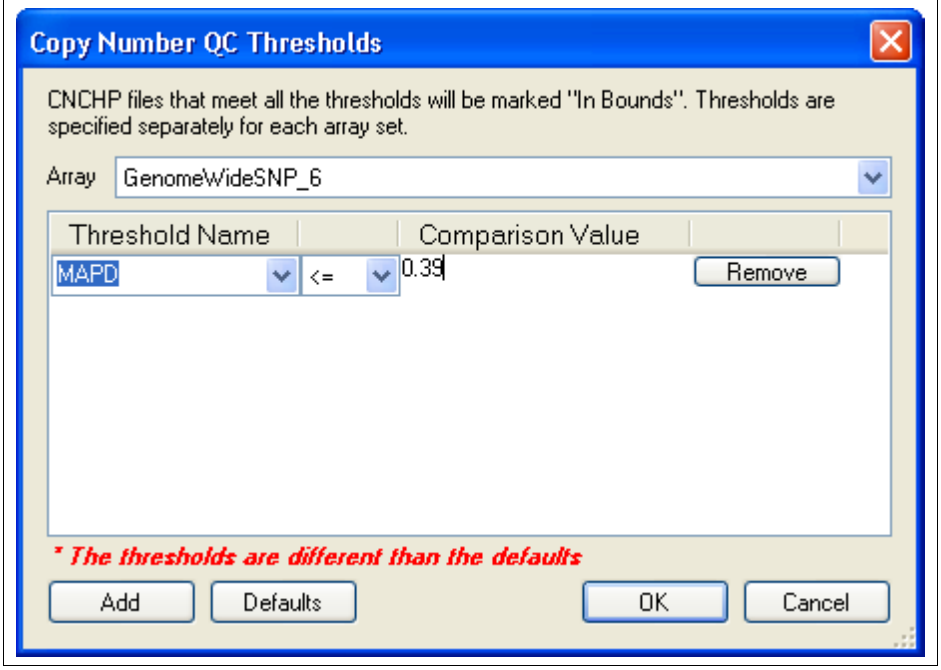


2. Select the array type to be modified from the Array dropdown list.
3. Enter the metric in the Threshold Name list in the table. The metrics are all listed in the Intensity QC Table (All Columns View).
4. Select the comparison operator:
   - less than (<)
   - less than or equal to (?)
   - greater than (>)
   - greater than or equal to (?)
   - equal to (=)
   - not equal to (!=)
5. Enter the Comparison value for the threshold.

   To delete a threshold item, click **Remove**.

NOTE: The default threshold for Genome-Wide Human SNP Array 6.0 is based on MAPD, while for Human Mapping 500K arrays it is based on IQR. The Human Mapping 100K array does not have a default threshold. When adjusting this value or adding additional metrics to threshold by, a flag will indicate that the thresholds are different from the defaults.

Figure 14.41 Values have been changed from default



NOTE: You can restore the Default threshold values by clicking Default.

If you wish to add another metric:

6. Select **Add**.
7. Type the exact name of this metric in the Threshold Name field, select a comparison, and enter a value.

   For additional metrics to be applied, they must exist in the Intensity QC Table (All Columns View).

For more information, see:

- *Copy Number QC Summary Table for 100K/500K* on page 231
- *CN/LOH QC Report Table for the Genome-Wide Human SNP Array 6.0* on page 261

8. Click **OK** to save the new copy number QC values.
   The new QC values will be used to filter results.

Copy Number Variation Analysis

Copy Number Variation (CNV) Analysis uses the Canary algorithm to make a CN state call (0, 1, 2, 3, 4) for previously identified regions with known copy number variations in the genome (also known as common copy number polymorphism, or CNP analysis). It uses a region file with a CNV region ID and a list of the CN/SNP probe sets in the region (a region with common copy number variation can contain a few too many CN/SNP probe sets).

NOTE: CNV analysis is only available for Genome-Wide Human SNP Array 6.0 data; it does not work with other arrays.

Performing Copy Number Variation Analysis

IMPORTANT: Always save your results folders with a unique batch name and location to make sure you can find your data later on. If you don't change the output root path, GTC will use the previous file path, which can belong to another data set or another hard disk. For more details on hard disk space requirements, see Appendix J, *Hard Disk Requirements* on page 363.

To perform CNV analysis:

1. Open the workspace and select the data set with the data for analysis.
2. Select the intensity data file set from the data tree.
3. From the workspace menu, select **Intensity Data > Perform Copy Number Variation Analysis…**; or
   Right-click the intensity data file set and select **Perform Copy Number Variation Analysis…** from the pop-up menu; or
   Click the **Perform Copy Number Variation Analysis…** button  in the tool bar.
   The Perform Copy Number Variation Analysis dialog box opens (Figure 15.1, Figure 15.2).

Figure 15.1 CNV Genotyping Options

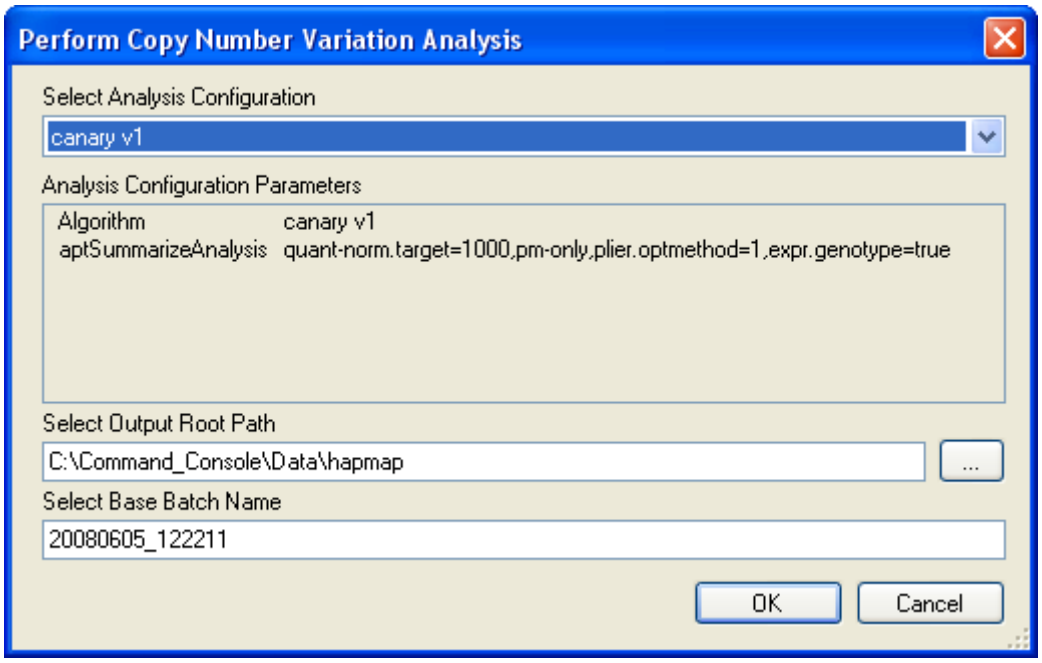
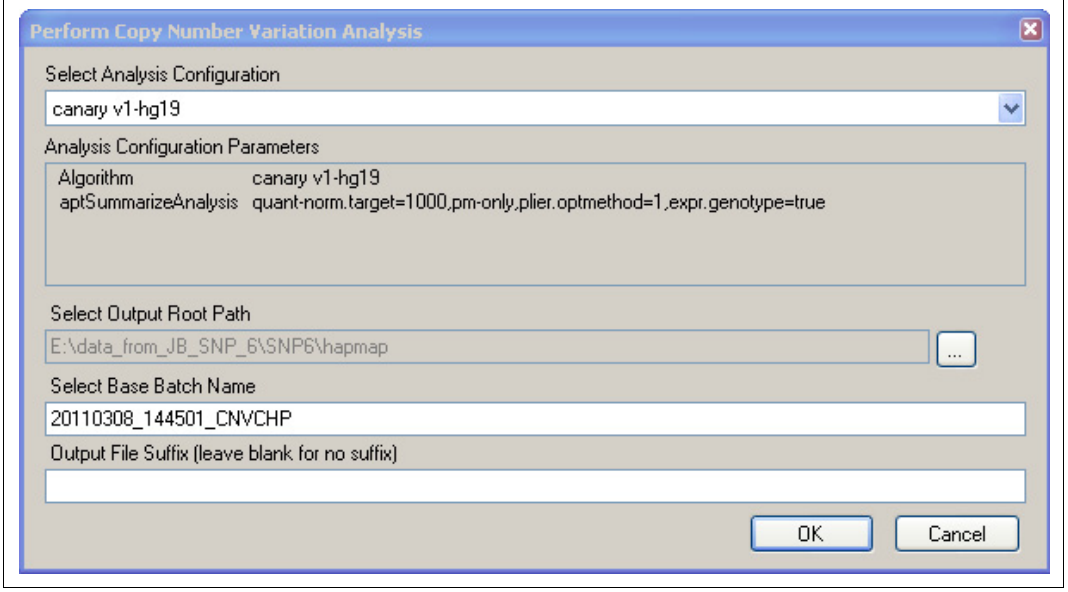


Figure 15.2 CNV Genotyping Options – Hg19



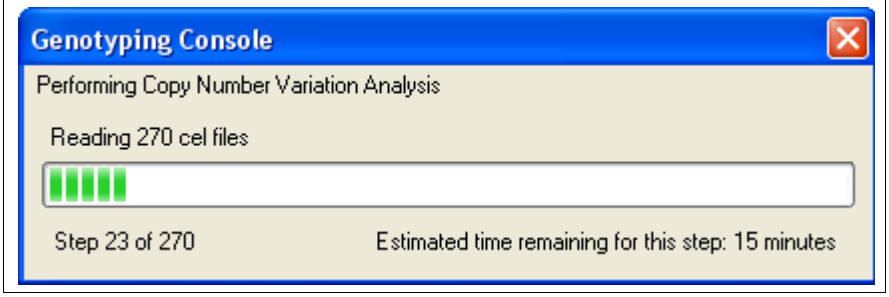
4. Select the appropriate analysis configuration for your data from the drop-down list:
   - canary v1 (hg18)
   - canary v1-hg19 (hg19)
5. Select the Output Root Path for the CNVCHP results set.

IMPORTANT: **Always save your results folders with a unique batch name and location to make sure you can find your data later on. If you don't change the output root path, GTC will use the previous file path, which can belong to another data set or another hard disk.**

6. Change the Base Batch (and folder) name if desired.
7. Click **OK**.

   Notices and a progress bar display the progress of the analysis (Figure 15.3).

Figure 15.3 CNV Progress notice



When the analysis is complete, the results are displayed in the CNV Table (see below).

The CNV call data can also be viewed in the Heat Map viewer if you have run Copy Number Analysis for the same CEL files; you cannot view CNVCHP data unless you have generated and loaded CNCHP data.

CNV Table Display

The CNV Results table displays the Call (CN State Call, Confidence Score, and sample attributes with All Column View) for each defined CNV Region on a selected chromosome, listed by CNVCHP file and CNV Region ID.

To open the CNV Table:

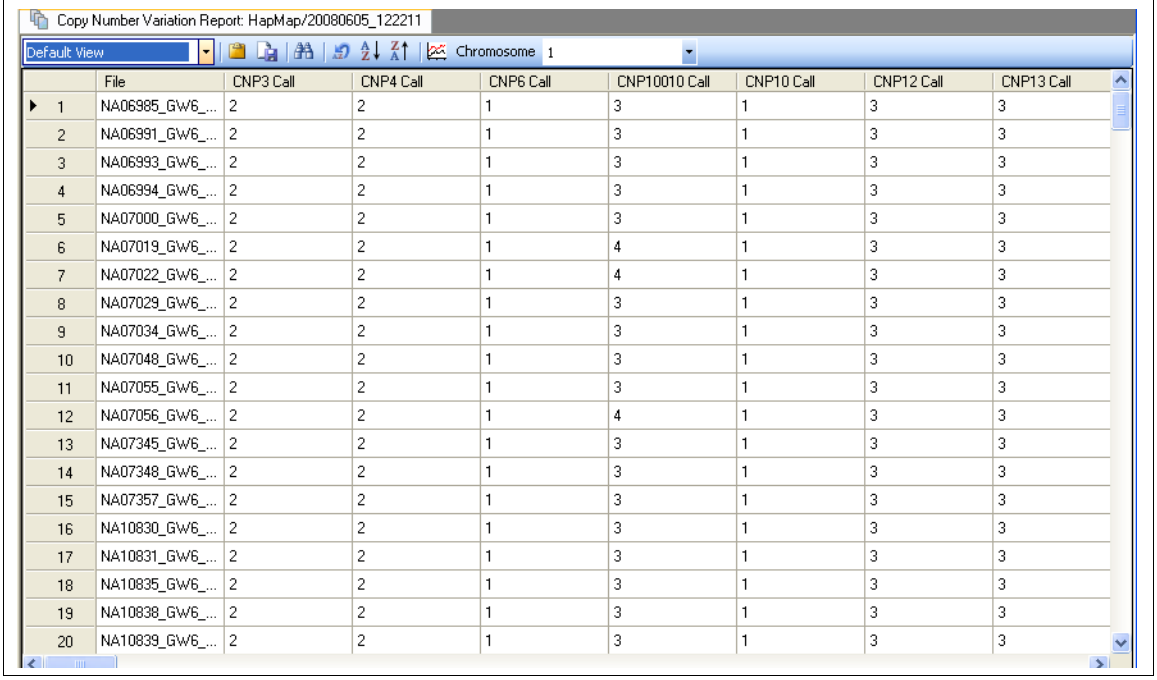
- Double-click on the CNV batch folder of interest; or

  Right-click on the Copy Number/LOH Results batch folder of interest and select **Show Copy Number Variation Results Table**; or

  From the Workspace menu, select **Copy Number Variation Results > Show Copy Number Variation Results Table**.

  The table opens and displays the results for Chromosome 1 (Figure 15.4).

Figure 15.4 CNV Results table, default view



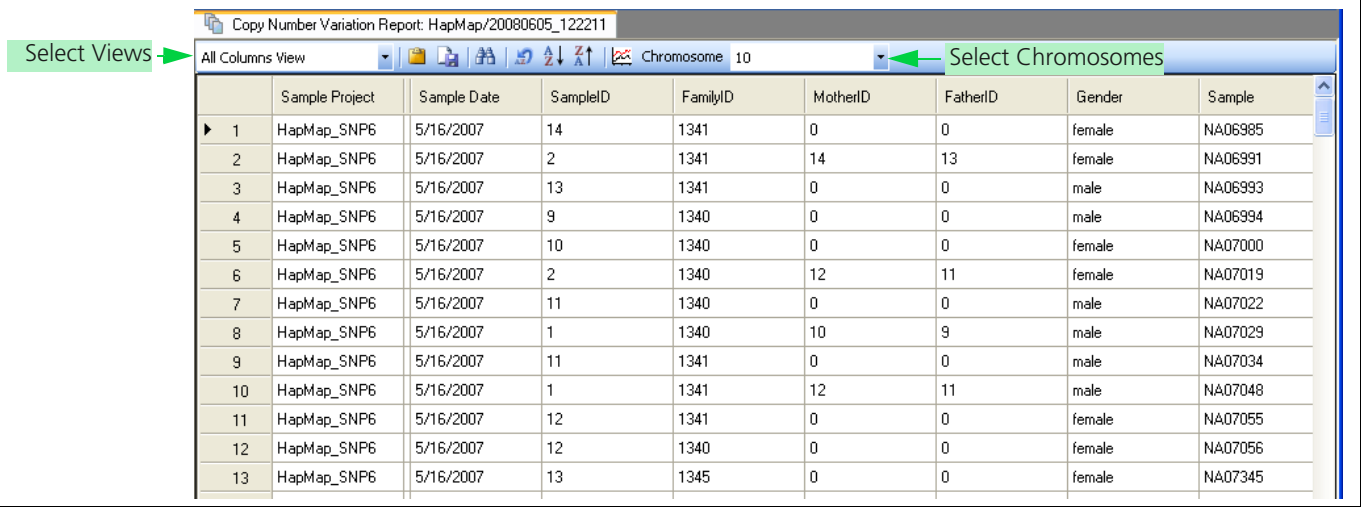
	File	CNP3 Call	CNP4 Call	CNP6 Call	CNP10010 Call	CNP10 Call	CNP12 Call	CNP13 Call
1	NA06985_GW6_...	2	2	1	3	1	3	3
2	NA06991_GW6_...	2	2	1	3	1	3	3
3	NA06993_GW6_...	2	2	1	3	1	3	3
4	NA06994_GW6_...	2	2	1	3	1	3	3
5	NA07000_GW6_...	2	2	1	3	1	3	3
6	NA07019_GW6_...	2	2	1	4	1	3	3
7	NA07022_GW6_...	2	2	1	4	1	3	3
8	NA07029_GW6_...	2	2	1	3	1	3	3
9	NA07034_GW6_...	2	2	1	3	1	3	3
10	NA07048_GW6_...	2	2	1	3	1	3	3
11	NA07055_GW6_...	2	2	1	3	1	3	3
12	NA07056_GW6_...	2	2	1	4	1	3	3
13	NA07345_GW6_...	2	2	1	3	1	3	3
14	NA07348_GW6_...	2	2	1	3	1	3	3
15	NA07357_GW6_...	2	2	1	3	1	3	3
16	NA10830_GW6_...	2	2	1	3	1	3	3
17	NA10831_GW6_...	2	2	1	3	1	3	3
18	NA10835_GW6_...	2	2	1	3	1	3	3
19	NA10838_GW6_...	2	2	1	3	1	3	3
20	NA10839_GW6_...	2	2	1	3	1	3	3

When you first open the table, it displays the Default view; if you change the view, the software remembers your choice and will open to the selected view the next time the table is opened.

If the samples have attributes, those attributes will be displayed at the far right of the table if you select All Columns View (Figure 15.5).

See *Table Features* on page 198 for more information on controlling the display of the table.

Figure 15.5 Sample attributes displayed in table



	Sample Project	Sample Date	SampleID	FamilyID	MotherID	FatherID	Gender	Sample
1	HapMap_SNP6	5/16/2007	14	1341	0	0	female	NA06985
2	HapMap_SNP6	5/16/2007	2	1341	14	13	female	NA06991
3	HapMap_SNP6	5/16/2007	13	1341	0	0	male	NA06993
4	HapMap_SNP6	5/16/2007	9	1340	0	0	male	NA06994
5	HapMap_SNP6	5/16/2007	10	1340	0	0	female	NA07000
6	HapMap_SNP6	5/16/2007	2	1340	12	11	female	NA07019
7	HapMap_SNP6	5/16/2007	11	1340	0	0	male	NA07022
8	HapMap_SNP6	5/16/2007	1	1340	10	9	male	NA07029
9	HapMap_SNP6	5/16/2007	11	1341	0	0	male	NA07034
10	HapMap_SNP6	5/16/2007	1	1341	12	11	male	NA07048
11	HapMap_SNP6	5/16/2007	12	1341	0	0	female	NA07055
12	HapMap_SNP6	5/16/2007	12	1340	0	0	female	NA07056
13	HapMap_SNP6	5/16/2007	13	1345	0	0	female	NA07345

For each chromosome, the CNV Results table displays:

- File Name: Name of the copy number variation CHP file.
- Call and Confidence Score for each defined CNV Region, by CNV Region ID. The CNV Region IDs are organized by genome position.

  Call - Copy number state estimated by the Canary algorithm

Confidence Score - Probability of Canary copy number state call given all possible CN State calls

To display results for a different chromosome, select the chromosome number from the Chromosome drop-down list in the table tool bar (Figure 15.5).

You can also scroll through chromosomes by clicking in the Chromosome dropdown list and:

- Using the mouse wheel
- Using the up/down arrow keys

Other table functions are described in *Table Features* on page 198.

You can also view CNV calls in the *Heat Map Viewer* on page 323.

Exporting CNV Data

You can export the CNV Results data in three different ways:

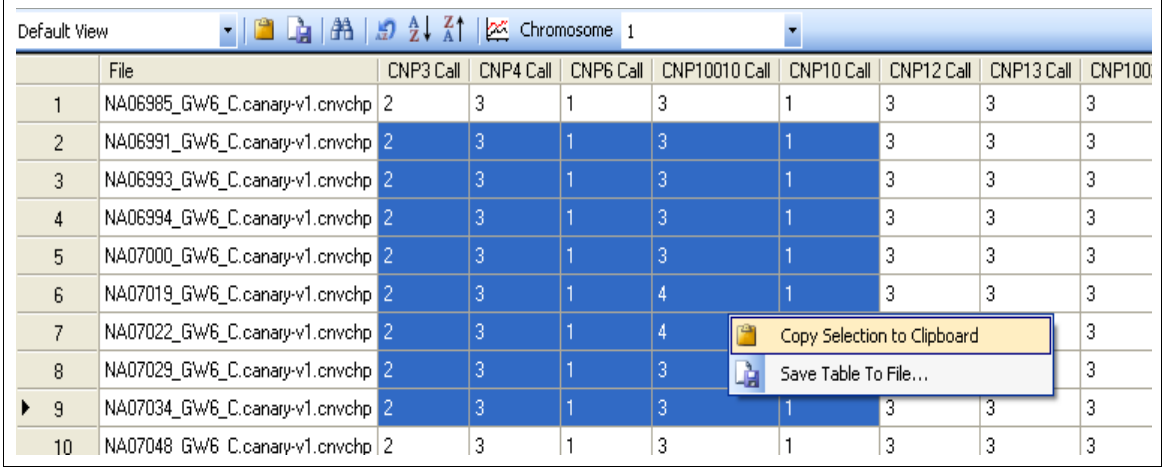
- Use the mouse to highlight and copy selected data in the table to the clipboard and paste it into a file
- Export the Table as a single text file with data for all CNVCHP files and the currently selected chromosome
- As a set of text files for the different CNV files, from the Batch Results, with data for all chromosomes in each file.

Exporting from the Table

To save selected data to the clipboard:

1. Select the cells you want to export in the table
2. Right-click in the table and select Copy Selection to Clipboard (Figure 15.6).

Figure 15.6 Right-click menu in table



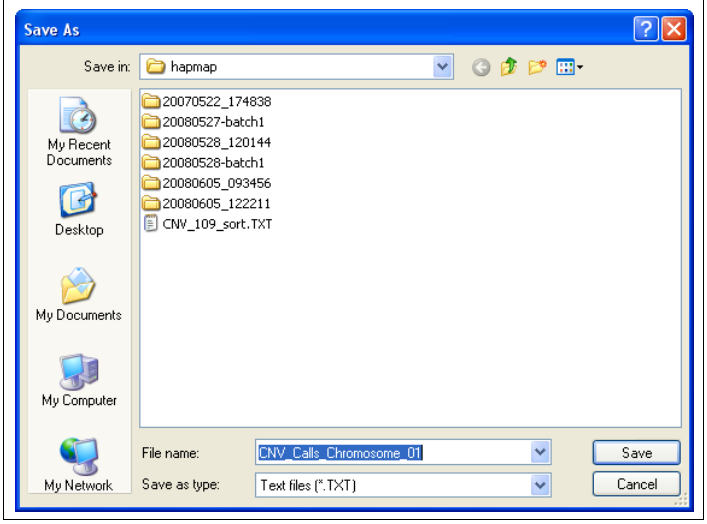
The selected data is copied to the clipboard and can be pasted into a new file.

When you export the data from the CNV Results table, you export the CNV data for the displayed chromosome and for all displayed CNVCHP files.

To save the table as a tab-delimited text file:

1. From the Table menu, select **Save Table to File…**; or
   Right-click in the table and **select Save Table to File…** from the pop-up menu; or
   Click the **Save Table to File** button in the table tool bar.
   The Save As dialog box opens (Figure 15.7).

Figure 15.7 Save As dialog box



2. Select a location, enter a name for the text file and click **Save**.

   The file is saved in the specified location.

   The file (Figure 15.8) contains a list of the files, chromosome regions and sample attribute information (if available) displayed in the table. It displays data only for the selected chromosome.

Figure 15.8 Text file (displayed in spreadsheet software)

CNV_Calls_Chromosome_01.TXT

	A	B	C	D	E	F	G	H	I	J
1	Index	File	CNP3 Call	CNP4 Call	CNP6 Call	CNP10010 Call	CNP10 Call	CNP12 Call	CNP13 Call	CNP10026 Call
2	1	NA06985_GW6_C.canary-v1.cnvchp	2	2	1	3	1	3	3	1
3	2	NA06991_GW6_C.canary-v1.cnvchp	2	2	1	3	1	3	3	1
4	3	NA06993_GW6_C.canary-v1.cnvchp	2	2	1	3	1	3	3	1
5	4	NA06994_GW6_C.canary-v1.cnvchp	2	2	1	3	1	3	3	1
6	5	NA07000_GW6_C.canary-v1.cnvchp	2	2	1	3	1	3	3	1
7	6	NA07019_GW6_C.canary-v1.cnvchp	2	2	1	4	1	3	3	1
8	7	NA07022_GW6_C.canary-v1.cnvchp	2	2	1	4	1	3	3	1
9	8	NA07029_GW6_C.canary-v1.cnvchp	2	2	1	3	1	3	3	1
10	9	NA07034_GW6_C.canary-v1.cnvchp	2	2	1	3	1	3	3	1
11	10	NA07048_GW6_C.canary-v1.cnvchp	2	2	1	3	1	3	3	1
12	11	NA07055_GW6_C.canary-v1.cnvchp	2	2	1	3	1	3	3	1
13	12	NA07056_GW6_C.canary-v1.cnvchp	2	2	1	4	1	3	3	1
14	13	NA07345_GW6_C.canary-v1.cnvchp	2	2	1	3	1	3	3	1
15	14	NA07348_GW6_C.canary-v1.cnvchp	2	2	1	3	1	3	3	1
16	15	NA07357_GW6_C.canary-v1.cnvchp	2	2	1	3	1	3	3	1
17	16	NA10830_GW6_C.canary-v1.cnvchp	2	2	1	3	1	3	3	1
18	17	NA10831_GW6_C.canary-v1.cnvchp	2	2	1	3	1	3	3	1

CNV_Calls_Chromosome_01

Exporting from the Batch Results

If you export data from the CNV batch folder, you create an individual text file for each CNVCHP file exported.

The file lists information about the CNV regions for every chromosome in each CNVCHP file in the results set.

To export CNV data from the Results set:

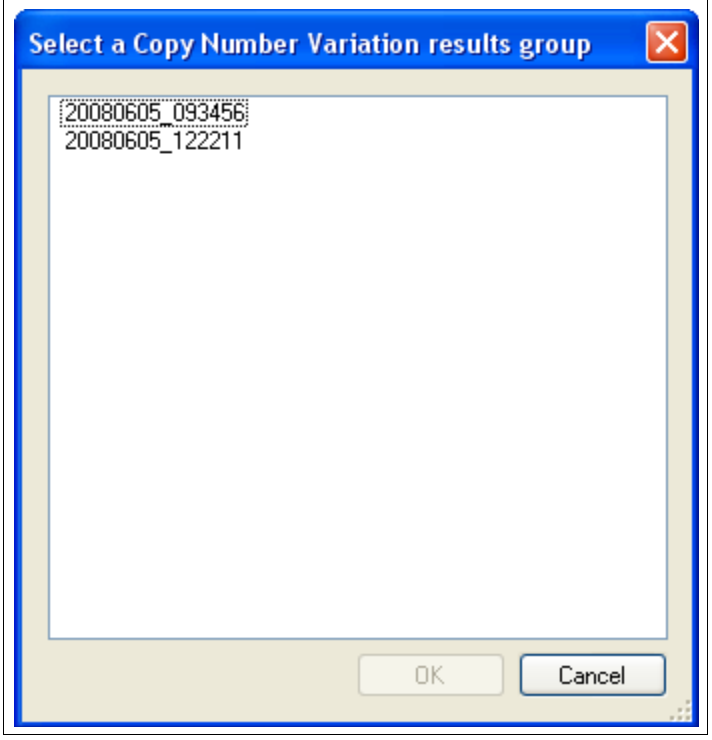
1. Select the Results set that you wish to export in the Data Tree.
2. From the Workspace menu, select **Copy Number Variation Results > Export Copy Number Variation Results**; or

Right-click the Copy Number/LOH data set and select **Export Copy Number Variation Results** from the pop-up menu.

If you have only one batch folder of CNVCHP files, these files will be automatically selected.

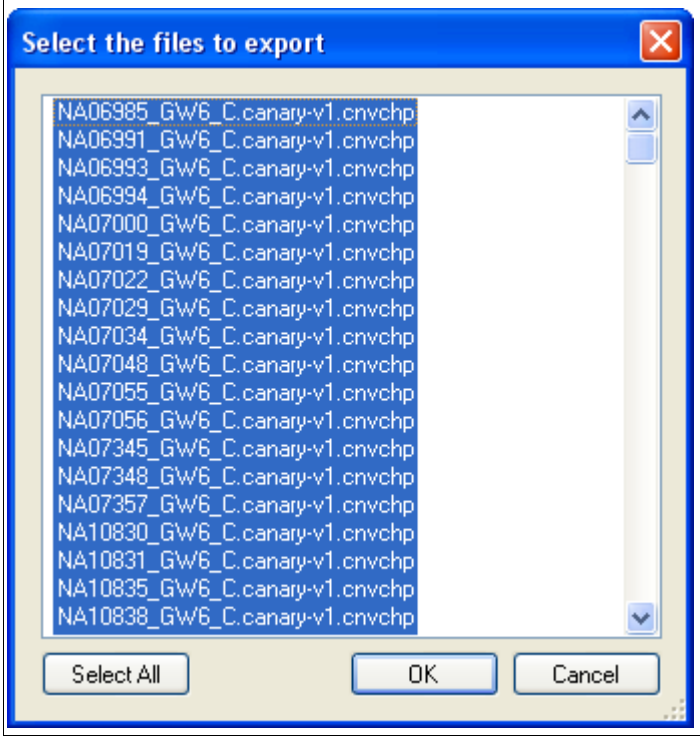
If you have not selected a batch results data set, the Select a Copy Number Variation results group dialog box opens (Figure 15.9).

Figure 15.9 Select a Copy Number Variation results group dialog box



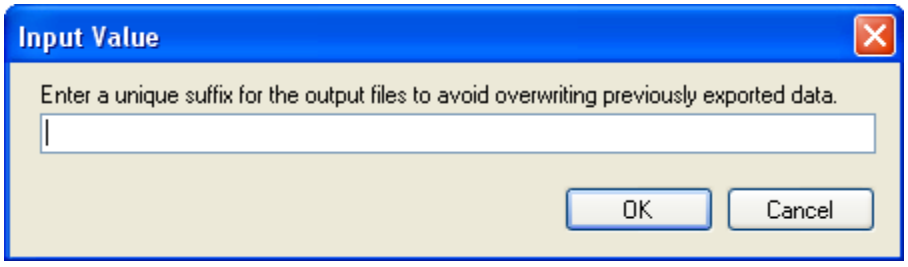
3. Select a results group for export and click **OK**.
   The Select the Files for Export dialog box opens (Figure 15.10).

Figure 15.10 Select the Files to Export dialog box



4. Select the Copy Number Variation Results files for export, or click **Select All**.
   Click **OK**.
   The Input Value dialog box opens (Figure 15.11).

Figure 15.11 Input Value dialog box



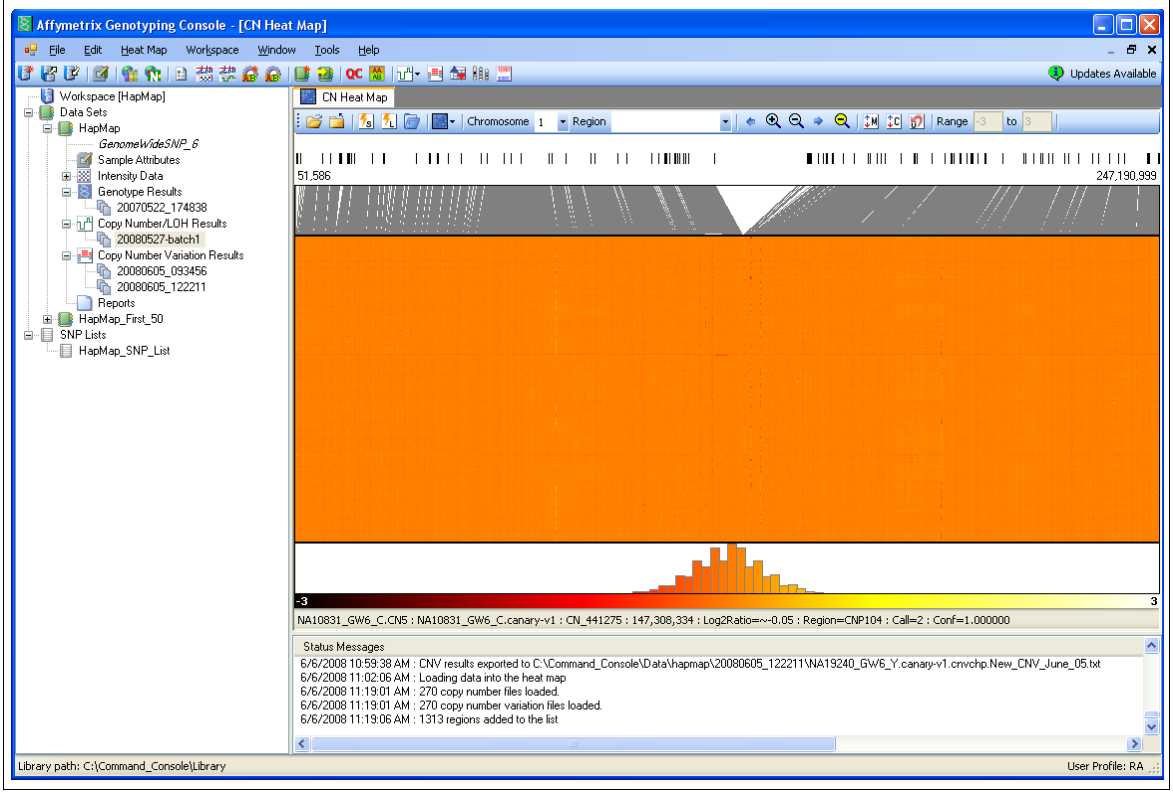
5. Enter a suffix for the output files (if desired) and click **OK**.
   Individual txt files are created for each CNVCHP file.
   Each file has header with information about CNV analysis, inherited from the CNVCHP files, and four columns, with:
   - Region
   - Signal
   - Call
   - Confidence

Heat Map Viewer

The Heat Map viewer (Figure 16.1) displays:

- Copy Number (CN) intensity values (~log2 ratios) from probe sets in the CNCHP files
- Copy number state calls for the copy number variations (CNV) in corresponding CNVCHP files, if available

Figure 16.1 Heat Map Viewer displayed in GTC



NOTE: You can view CN data with or without matching CNV data. If you change the default CNV map after data is loaded in the Heat Map viewer, all the associated CNVCNP files will be removed (CNVCHP files are map-specific). You must have CNCHP data available to load CNVCHP data.

The Heat Map Viewer enables users to:

- Compare the CNV calls from CNVCHP files using raw intensity values from individual probe sets within the CNV regions from CNCHP files.
- Survey large quantities of genomic data to detect de novo CNV regions.

The Heat Map viewer displays:

- CNV regions if CNVCHP files are loaded (the default CNV map file with BED format will automatically be loaded).
- Genomic positions currently displayed in the viewer.
- Log2ratio value data from CNCHP files (intensity value) for each SNP or CN probe displayed as a color value in a heat map with a pre-defined scale.

- Summary histogram to indicate the frequencies of probe sets with certain color values.

The Status Bar displays:

- Sample names: CNCHP file name, and CNVCHP file name (if available)
- SNP or CN probe set ID, with:
  - Chromosome Position
  - log2 ratio of a SNP or CN probe set
- CNV region ID if CNVCHP files are loaded, with:
  - CNV calls
  - Call confidence

Opening the Heat Map

To open the Heat Map viewer without loading data:

- Click the Heat Map button  in the GTC main tool bar.

  This allows you to *change the log2 Ratio Range* if desired before loading data (page 327). The recommended range should not exceed -10 to 10.

NOTE: You can view CN data with or without matching CNV data. If you change the default CNV map after data is loaded in the Heat Map viewer, all the associated CNVCNP files will be removed (CNVCHP files are map-specific). You must have CNCHP data available to load CNVCHP data.

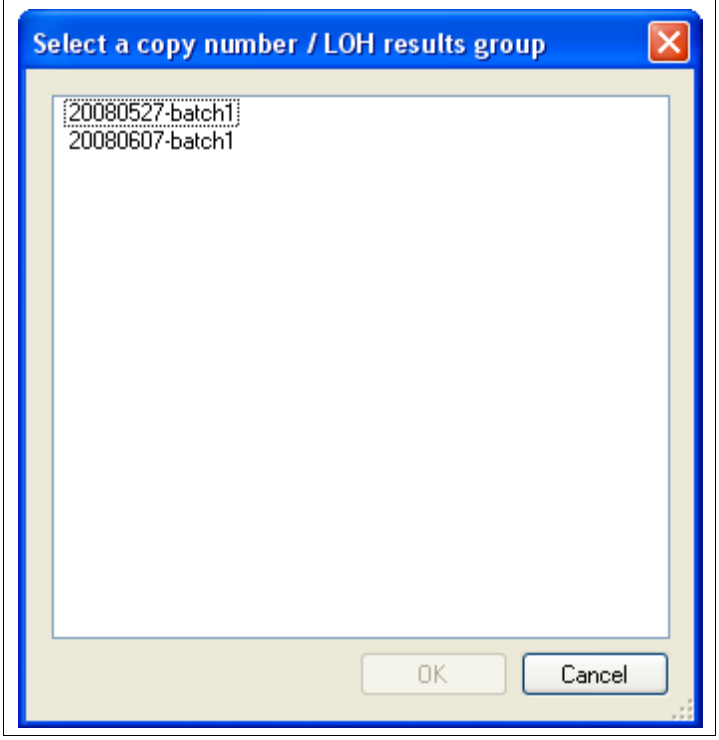
NOTE: You can only view CHCHP/CNVCHP files from one SNP 6.0 data set in the Heat Map viewer at a time.

NOTE: Loading data into the Heat Map may take a long time, especially with large results sets. You can use the Quick Load feature to save loaded data and reload it more quickly (see *Using the Quick Load Feature* on page 327.

To open the Heat Map Viewer and load data:

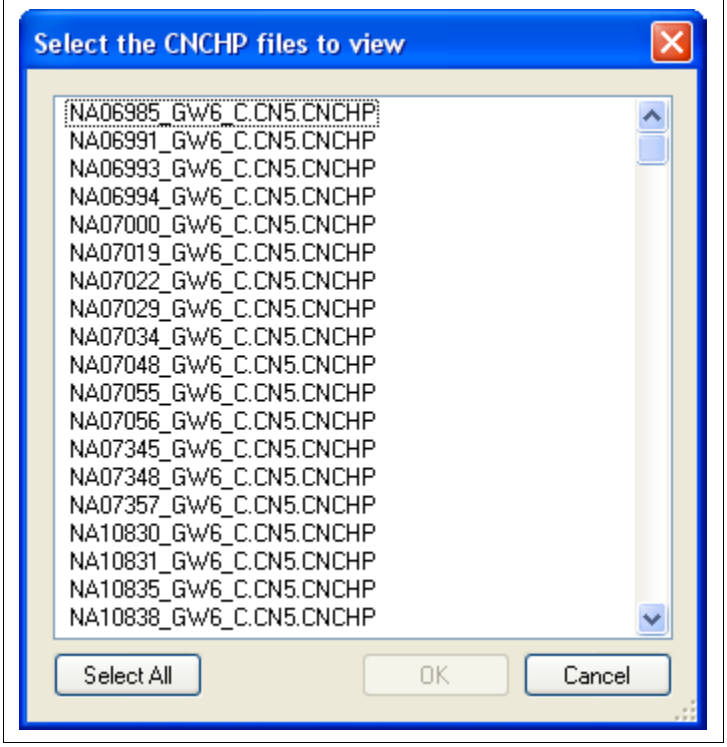
1. Right-click on the CN/LOH batch results you wish to view and select **View Results in Heat Map**; or
   From the Workspace menu, select **Copy Number/LOH Results > View Results in Heat Map**; or
   Click the Heat Map button  in the GTC main tool bar.
   If more than one batch of CN results is available, the Select a copy number.LOH results group dialog box opens (Figure 16.2).

Figure 16.2 Select Copy Number/LOH results group dialog box



2. Select the Results Group to be displayed and click **OK**.
   A list of the CNCHP files in the selected results group opens (Figure 16.3).

Figure 16.3 Select the CNCHP files to view dialog box



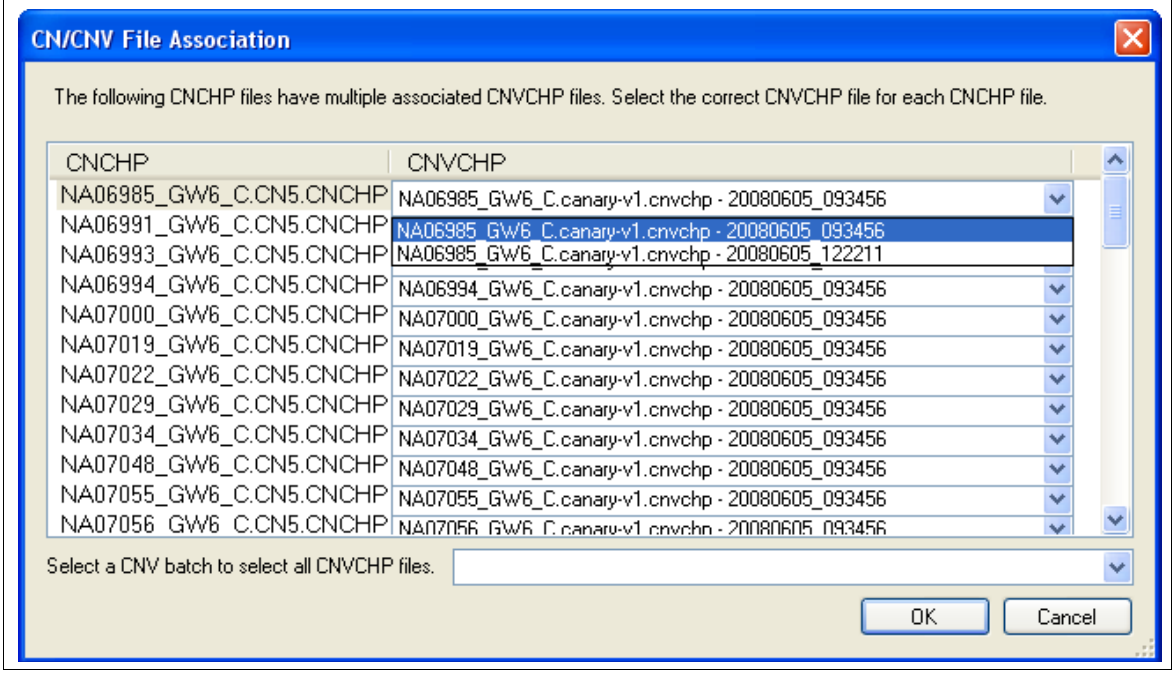
3. Select files you wish to load or click **Select All**.
4. Click **OK**.

   If only one CNVCHP file is associated with each selected CNCHP file, the data will automatically start loading both CNCHP and the matching CNVCHP files into the Heat Map.

   If some of the CNCHP files do not have matching CNVCHP files, only the CNCHP files are loaded for these samples.

   If multiple CNVCHP files are associated with the selected CNCHP file the CN/CNV File Association dialog box opens (Figure 16.4).

Figure 16.4 SCN/CNV File Association dialog box

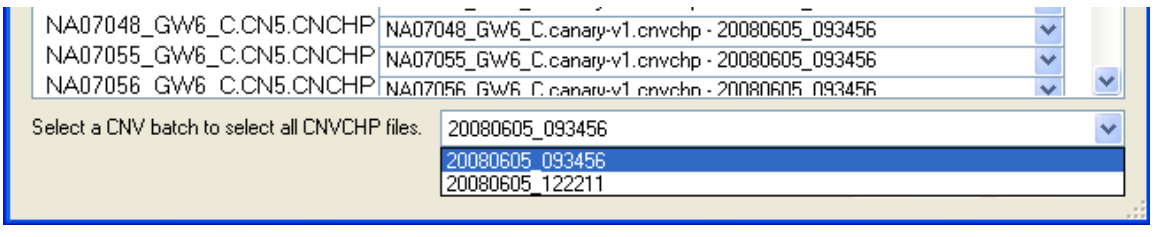


5. Use the CN/CNV File Association dialog box to select the CNV files to be displayed with the CNCHP files in the Heat Map.

   You can select all CNVCHP files from a given batch using the Select CNV Batch drop-down (Figure 16.5) and manually override your batch choice for any of chosen CNVCHP files

   The CNV data is automatically loaded if available, and the CNV map associated with these CNVCHP files will automatically follow.

Figure 16.5 Selecting CNV batch sets



6. Click **OK**.

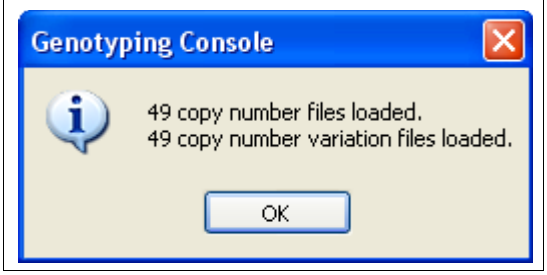
   Notices and progress bars display the progress of loading the data.

   The Heat Map Viewer opens and loads the results.

The CNV data is automatically loaded (if available) and the CNV map associated with these CNVCHP files is also automatically loaded.

When loading is finished, a notice (Figure 16.6) informs you of the number of copy number and copy number variation (if available) files loaded. Click **OK** to exit the confirmation window.

Figure 16.6 Load Notice



The Heat Map menu appears in the GTC main menu bar.

Changing the log2 Ratio Range

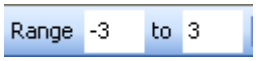
The range of log2 ratio values displayed on the selected heat map palette can be changed.

NOTE: Changing the ratio range must be done before loading data in the Heat Map Viewer.

To change the log2 Ratio range:

1. Open the Heat Map Viewer without loading data (Click the Heat Map button in the GTC main tool bar).
2. Enter the log2 ratio values in the Range boxes in the *Heat Map Viewer tool bar* (Figure 16.7, Figure 16.11).

Figure 16.7 Range boxes



3. Load data as described above.

Using the Quick Load Feature

Loading data into the Heat Map Viewer may take a long time, especially with large results sets.

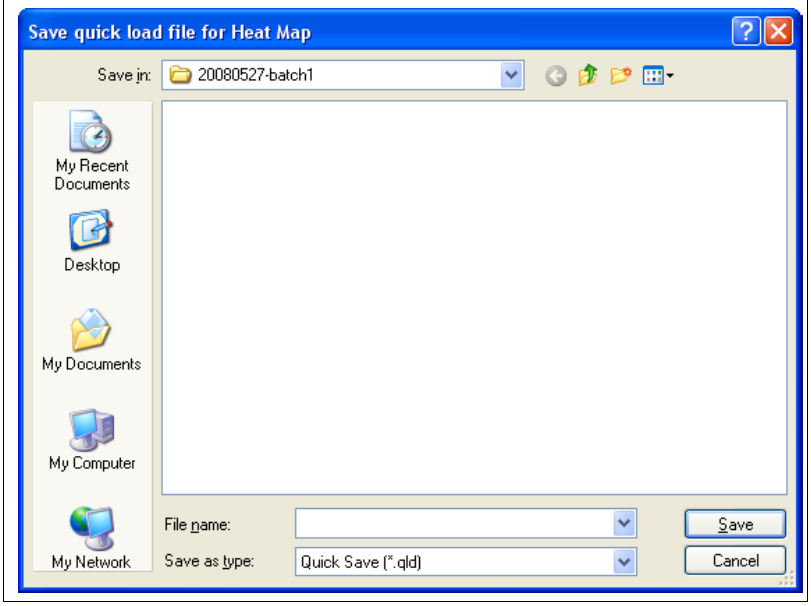
The Quick Load feature of the Heat Map Viewer enables you to save loaded data, so that you can reload the saved Quick Load data more quickly.

NOTE: You will not be able to add any more data to a quick load file, to change a CNV map, or to use the quick load feature again after a quick load file is open.

To save a data set as a Quick Load file after initially loading data in the Heat Map Viewer:

1. Click the **Quick Load** save button in the Heat Map Viewer toolbar.
   The Save Quick Load file for Heat map dialog box opens (Figure 16.8).

Figure 16.8 Save Quick Load file for Heat Map dialog box

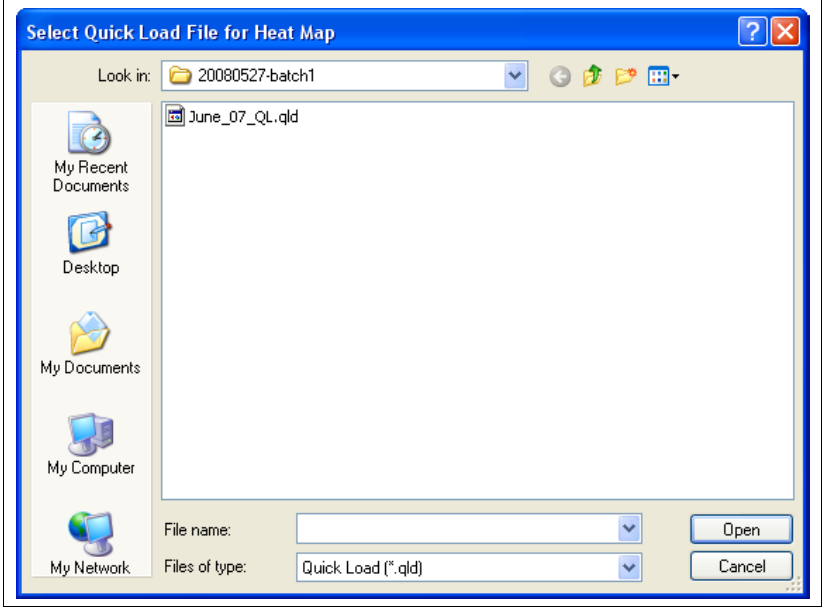


2. Select a location and enter a name for the Quick Load file.
3. Click **Save**.
   The Quick Load file is saved.

To reload a Quick Load file:

1. Click the **Quick Load** button . in the Heat Map Viewer toolbar.
   The Select Quick Load file for Heat Map dialog box opens (Figure 16.8).

Figure 16.9 Select Quick Load file for Heat Map dialog box



2. Select a previously created Quick Load file.
3. Click **Open**.
   The data is loaded into the Heat Map Viewer more quickly.

NOTE: You cannot add new data two an opened quick load file or open a second quick load file when one is already displayed in the Viewer.

Changing the CNV Map

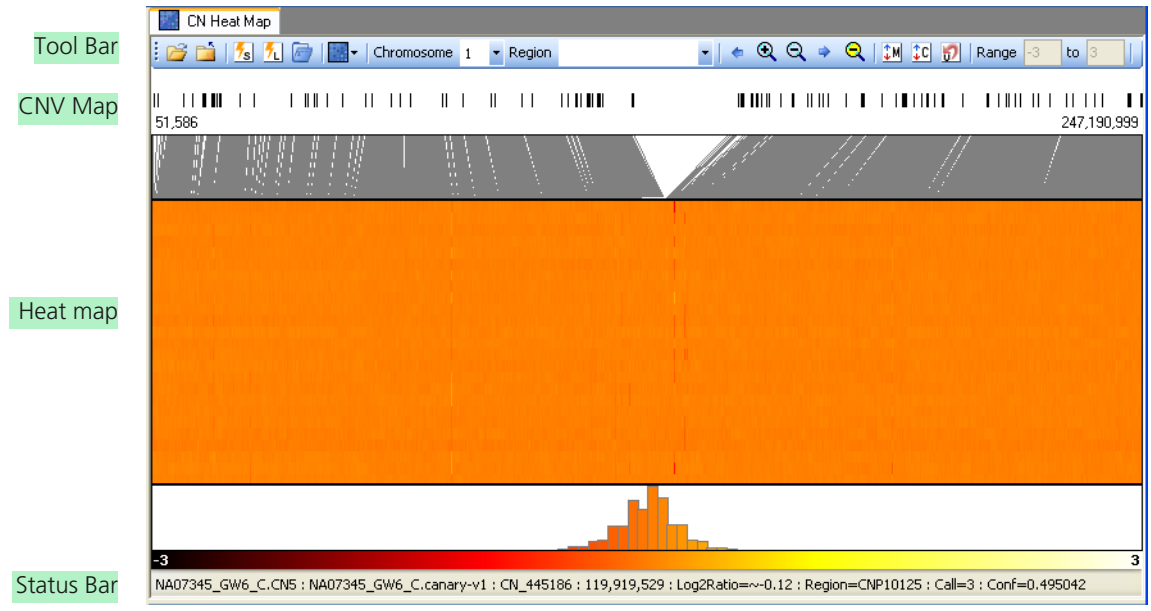
You can change a CNV map when you have added files to heat map; if you have both CNCHP and CNVCHP files loaded, changing a CNV map will flash out all the CNVCHP files because CNVCHP files are CNV map specific. Once you change your CNV map, you will no longer be able to see CNV calls and call confidences. You can still choose a CNV region or browse it in the heat map. Custom maps for CNV analysis are not supported at this time.

Overview of the Heat Map Display

The Heat Map viewer (Figure 16.10) displays:

- Log2ratio value data from CNCHP files (copy number data) for each SNP or CN probe set on the selected chromosome as a color value in a heat map scale.
- Genomic position of the SNP and CN probe sets and CNV regions for that chromosome
- Copy Number calls for the CNV regions from the CNVCHP files (if available) can be displayed in the status bar by mousing over the Heat Map.
- When first opened, the viewer displays the data for Chromosome 1.

Figure 16.10 Parts of the Heat Map



The Heat Map (Figure 16.10) has the following components:

- *Tool bar* on page 330
- *CNV Map* on page 331
- *Heat Map* on page 332
- *Heat Map Viewer Histogram* on page 334
- *Status Bar* on page 334

The Heat Map Viewer provides tools for:

Tool bar

The Tool bar (Figure 16.11) provides quick access to the functions of the Heat Map Viewer.

Figure 16.11 Heat Map Viewer tool bar

Table 16.1 Heat Map Viewer Tool bar functions

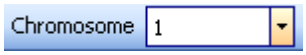
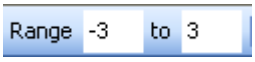
Button	Description
	Open
	Close files
	Save loaded data in Heat Map to disk
	Load a previously saved data from heat map
	Open CNV map
	Change color palette
	Select chromosome for display
	Select CNV region from CNV map
	Move left
	Zoom in
	Zoom out
	Move right

Table 16.1 Heat Map Viewer Tool bar functions

Button	Description
	Full zoom out
	Sort by median
	Sort by CNV call
	Resort sort
	Range display: Can only be set before loading data

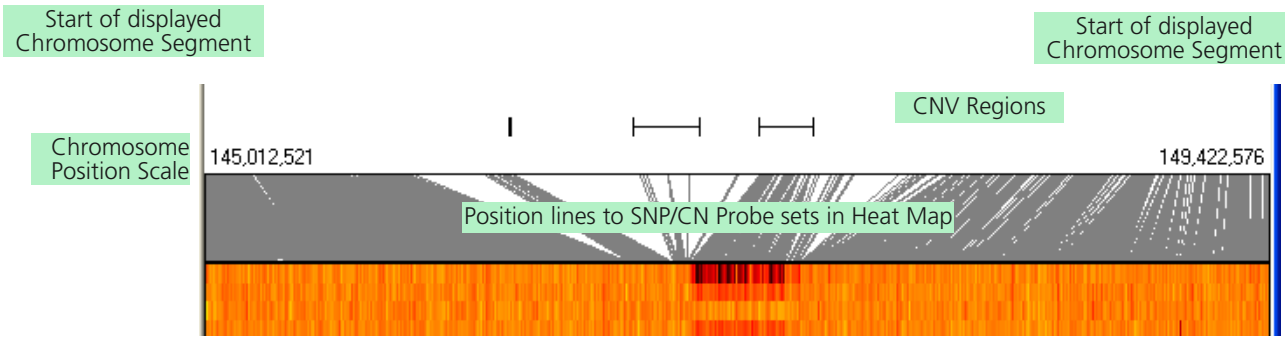
Many of these functions can also be accessed using the Heat Map Viewer menu when the Heat map is open. Some functions can also be accessed by right-clicking in the Heat Map and selecting the desired function from the popup menu.

CNV Map

The CNV Map (Figure 16.12) displays:

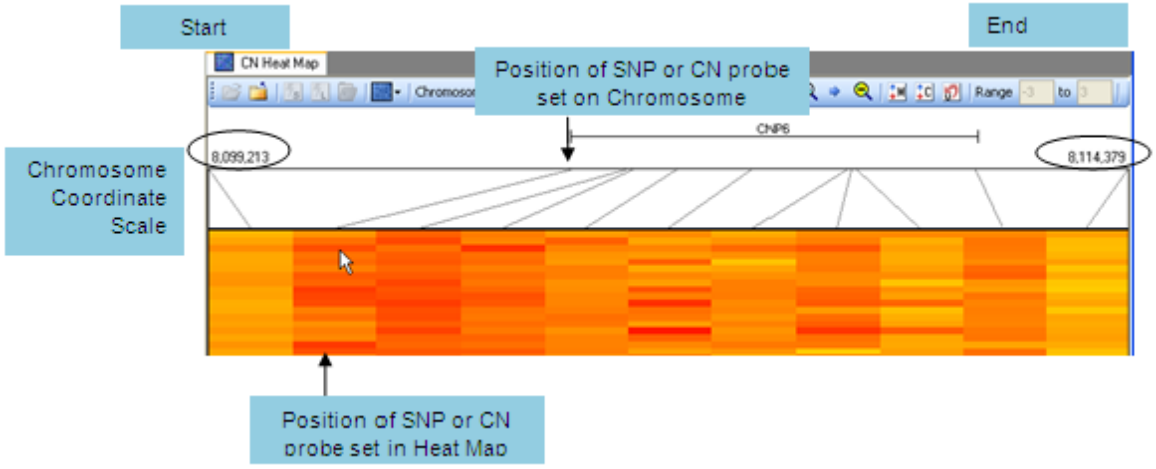
- CNV regions in the loaded CNV Map for the selected chromosome
- Chromosome position scale displaying the chromosome positions for CNV regions that contains the SNPs and CN probe sets displayed in the Heat Map Viewer.
- Position of the SNP and CN probe sets on the section of chromosome displayed in the current view of the Heat Map Viewer.

Figure 16.12 CNV Map (detail)



Since SNP and CN probe sets are not uniformly distributed along the chromosome, the relationship between the heat map and the chromosome map is not linear (Figure 16.13).

Figure 16.13 Non-alignment of SNP/CN Probe set in Heat Map with genomic position in CNV Map



Heat Map

The Heat Map displays the log2ratio values for the SNPs and CN probe sets using a heat range scale.

SNP/CN intensity values are displayed on the horizontal range, with the results files stacked vertically. (Figure 16.14)

Figure 16.14 CNV map, Heat map, histogram

When first loaded the data files are arranged from top to bottom by file name.

- You can sort by median intensity values for the SNP and CN probe sets displayed in the heat map within the current view window and unsort to go to the original order
- You can sort by the CNV calls from CNVCHP files in the heat map if your current viewing window has a CNV region in it and unsort to go to the original order

If some CNCHP files do not have CNVCHP file data, these files will be displayed at the bottom of the Heat Map after sorting on CNV Call values.

See *Sorting Data in the Heat Map* on page 339.

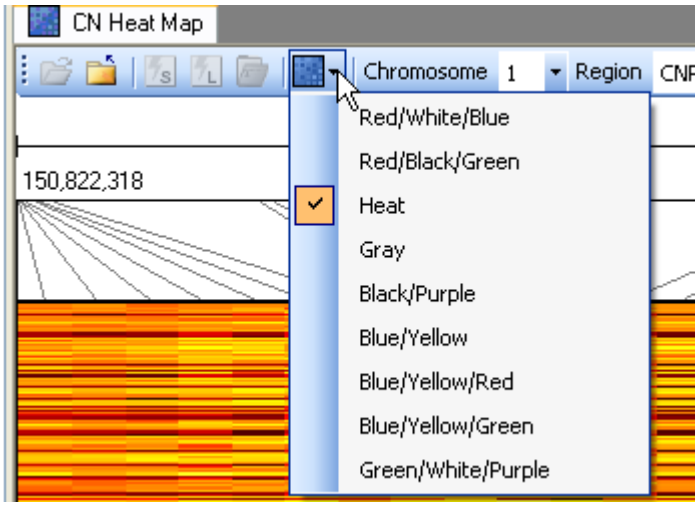
You can export a list of the CNCHP and CNVCHP file path and files names in their displayed order before and after sorting (see *Exporting a List of Data Files in Sorted Order* on page 340.

You can select different color palettes for the display (below) or *change the log2 Ratio rang*e (page 327).

To select different color palettes for the Viewer:

1. Click the Color Palettes button  in the viewer tool bar.
2. Select a palette choice (Figure 16.15).

Figure 16.15 Heat Map Viewer color palette options

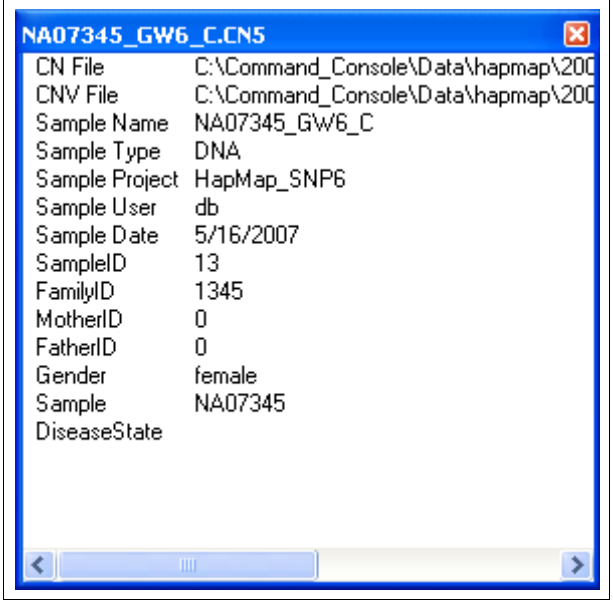


To display the attributes and other data, if available, for a data file:

- Double-click in the Heat Map in the file row you are interested in.

  A box (Figure 16.16) opens with sample data: CNCHP and CNVHP file path and file names (if available), and sample attribute data (if available).

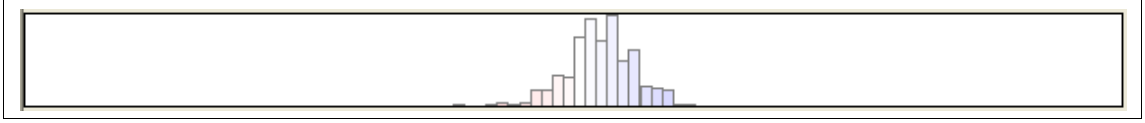
Figure 16.16 Sample data



Heat Map Viewer Histogram

The histogram (Figure 16.17) indicates the frequencies of probe sets with different intensity values.

Figure 16.17 Histogram



If the cursor is positioned over a specific region of the Heat Map, the histogram automatically adjusts and displays the frequencies of probe sets within that specific region.

Status Bar

You can display the following information in the Status bar (Figure 16.18) by putting the mouse arrow over a SNP or CN probe set position:

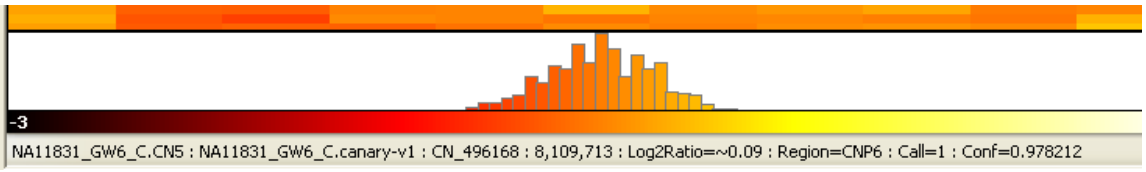
- CN and CNV file names (if CNV data available)
- SNP or CN probe set ID, with:
  - Chromosome Position
  - Log2Ratio

NOTE: The log2 ratios displayed in the status bar may not exactly match the log2 ratios for in the CNCHP files. The values in the CNCHP file are converted into a color value used for the heat map display; this color value is then translated into the log2 ratio value used for the status bar display.

- CNV region ID if CNVCHP files are loaded, with:
  - CNV calls
  - Call confidence

NOTE: Affymetrix recommends that you do not use long file names for the .CEL and .CHP files, since these long names can cause display problems in the Heat Map Viewer. The status bar in the Heat Map will not be able to display all the information if the CNCHP and CNVCHP file names (derived from the .CEL file names) are too long. If the data is truncated, you can increase the size of the Heat Map on the screen by dragging the vertical window split bar.

Figure 16.18 Information in Status bar



Navigating the Heat Map Viewer

The Heat Map Viewer provides several options for selecting data of interest:

- *Selecting the Chromosome for Display* (below)
- *Viewing CNV Regions* on page 335
- *Zooming In on a CNV Region* on page 337

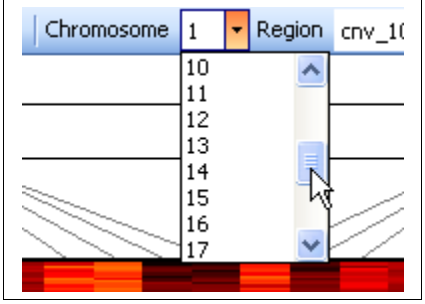
Selecting the Chromosome for Display

The Viewer displays the data for one chromosome at a time. When the Viewer is first opens, it displays all of chromosome 1 in the Chromosome Map and Heat Map.

To change the chromosome displayed in the Viewer:

- Select the Chromosome of interest from the Chromosome dropdown (Figure 16.19).

Figure 16.19 Chromosome list



- Users can scroll through the chromosomes (as viewed in the Heat Map) by using either the mouse scroll wheel or the up/down arrow keys to navigate the list of chromosomes in the dropdown menu.

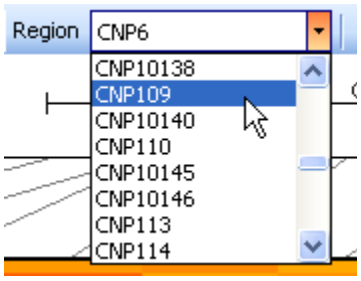
Viewing CNV Regions

The CNV regions in the loaded CNV map are displayed in the Chromosome Map.

To look at a specific CNV region

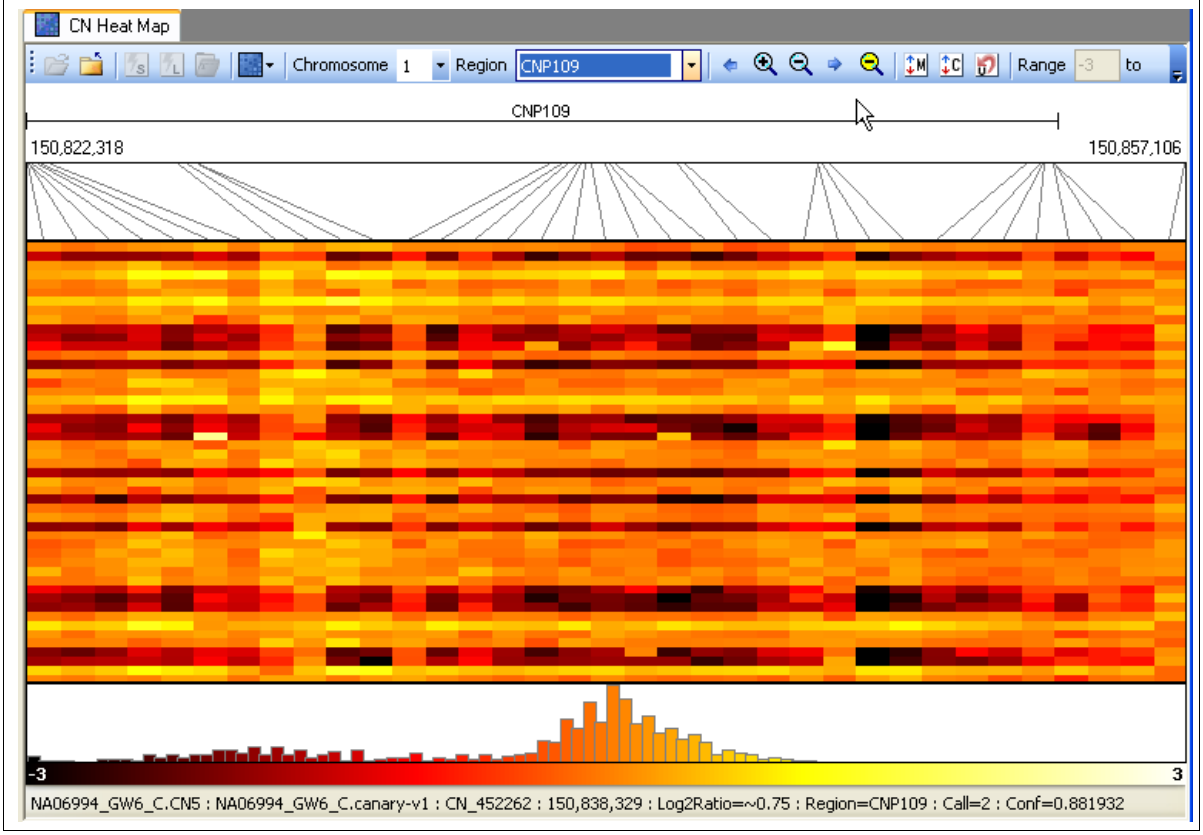
- Select the CNV region from the Region list in the viewer tool bar (Figure 16.20).

Figure 16.20 Regions dropdown list



The selected region is displayed in the Heat Map as default view (Figure 16.21).

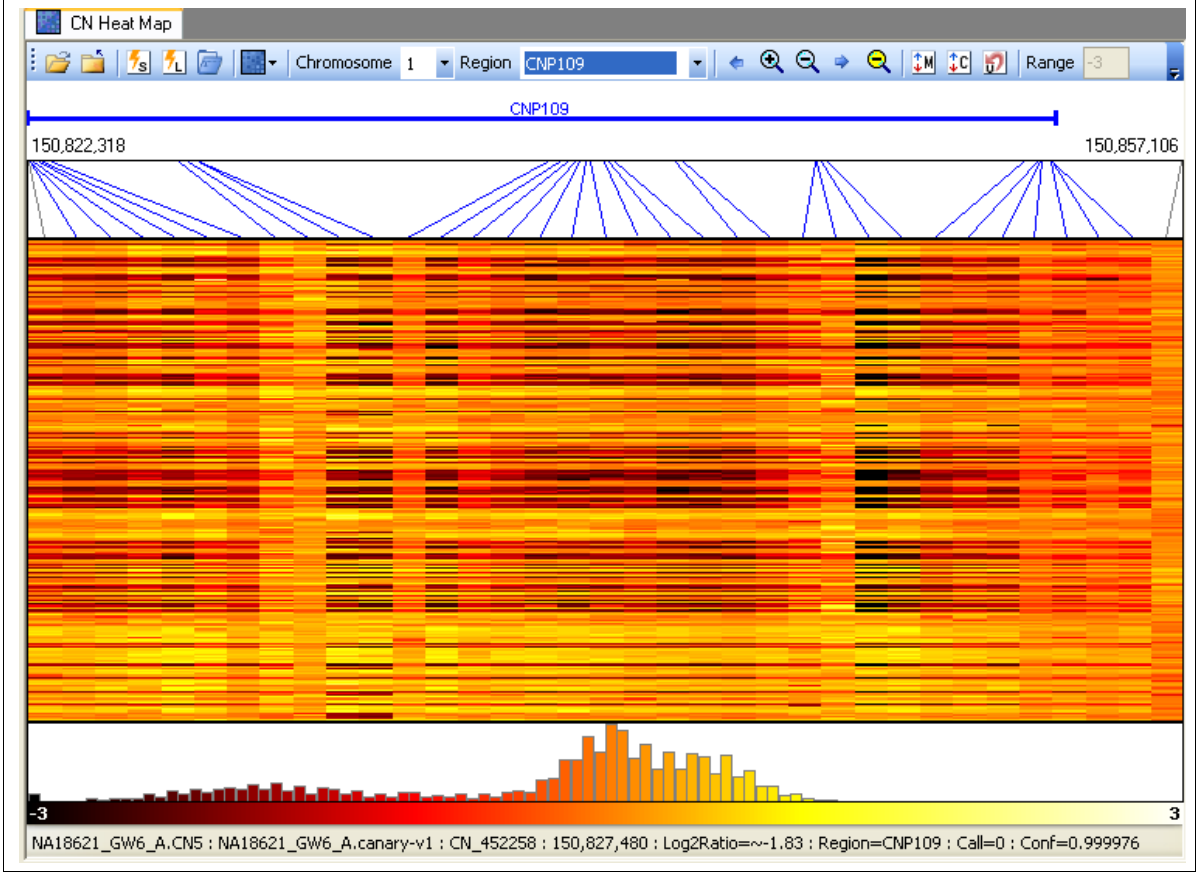
Figure 16.21 Selected Region displayed in the Heat map viewer



Users can scroll through the regions (as viewed in the Heat Map) by using either the mouse scroll wheel or the up/down arrow keys to navigate the list of regions in the dropdown menu.

Double-click a region in the CNV Map to highlight the markers and to zoom to that region (Figure 16.22).

Figure 16.22 Highlighted region and markers

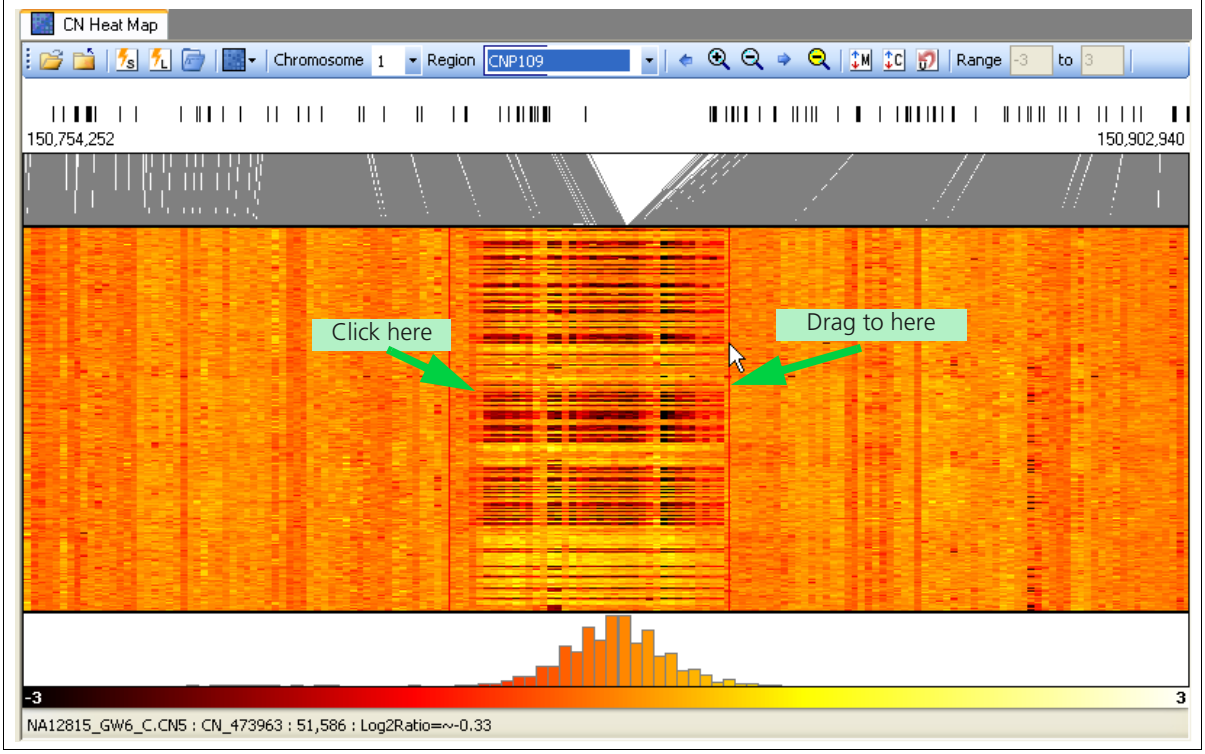


Zooming In on a CNV Region

You can zoom in on a section of the Heat Map by selecting the area in the Heat Map.

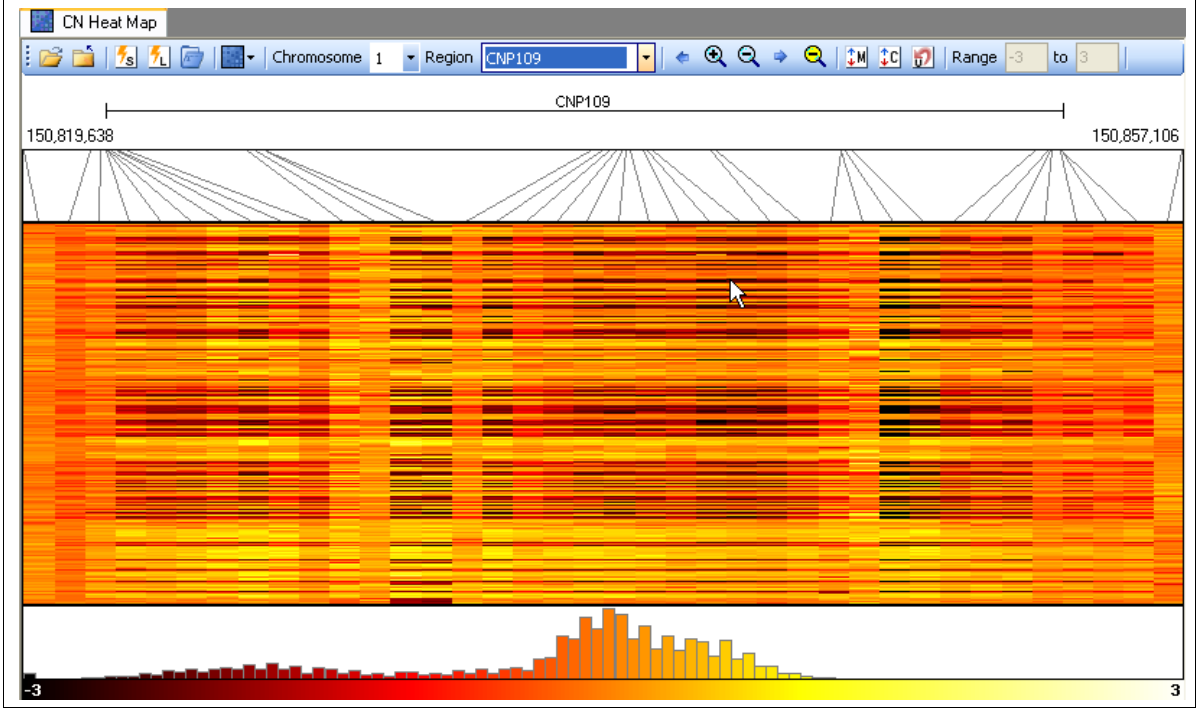
- Click at the start and release at the end of the area you wish to zoom in on (Figure 16.23).

Figure 16.23 Selecting the map region to zoom in on



The selected region is displayed in the Heat Map and the CNV map (Figure 16.24).

Figure 16.24 Zoomed in view of the region selected in the figure above



You can also use the buttons in the Heat Map Viewer tool bar (Table 16.2) and the commands in the Heat Map Viewer menu to change the view in the Heat Map.

Table 16.2 Heat Map Viewer tool bar navigation buttons

Button	Function
	Move left.
	Zoom in
	Zoom out
	Move right
	Full zoom out

Double click on a region in the CNV Map to highlight the markers and to zoom to that region.

Sorting Data in the Heat Map

You can sort the displayed SNP values by:

- Median Log2 ratio values for all the SNP and CN probe sets displayed in the current view in the Heat Map Viewer (Figure 16.25).
- CNV Call values for the CNV regions currently displayed in the Heat Map Viewer. If more than one CNV region is present, then the average of CNV calls for all the CNV regions is used to sort the CNV calls.

  If some CNCHP files do not have CNVCHP file data, these files will be displayed at the bottom of the Heat Map after sorting on CNV Call values.

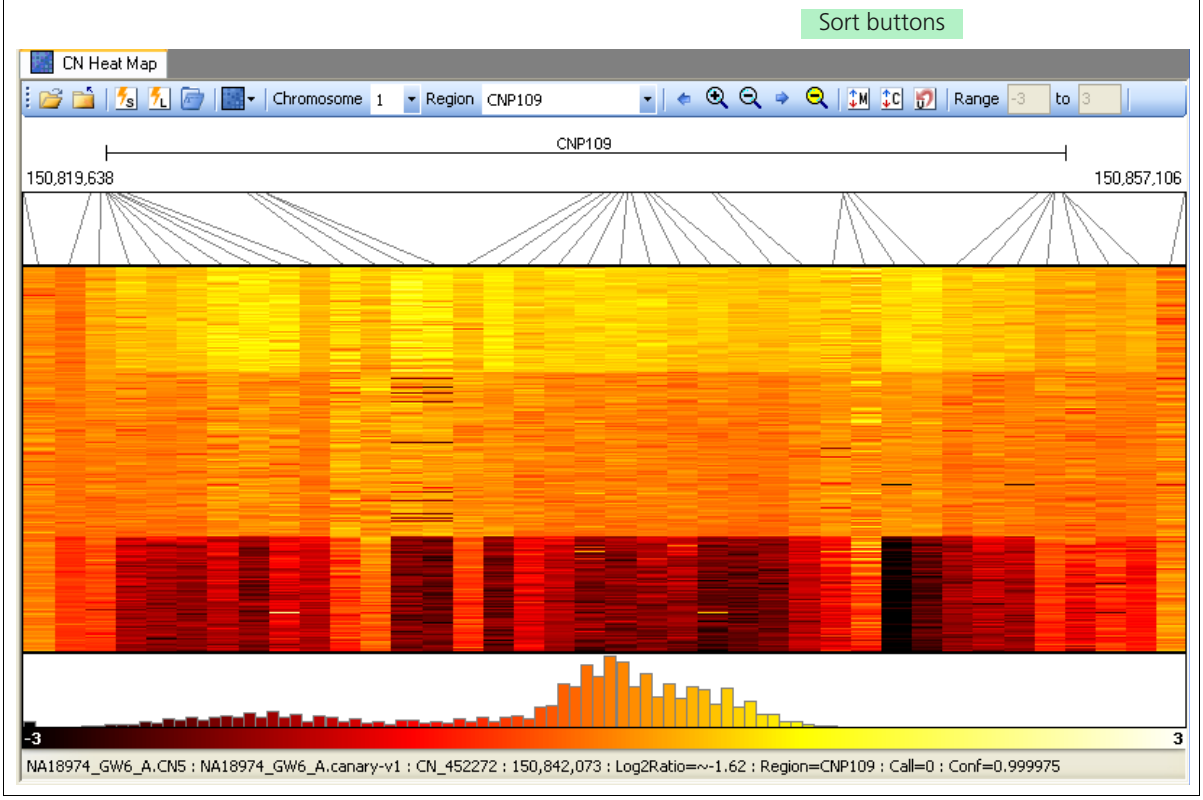
After sorting you can export a list of the files in their new sorted order.

To sort:

1. Zoom in on the region you wish to investigate.
2. Select the sort option from the Heat map menu, or click the button for the option:
   -  Sort by Median Log2 ratios
   -  Sort by CNV values

   The files will be sorted by the selected metric (Figure 16.25).

Figure 16.25 Sorted by median

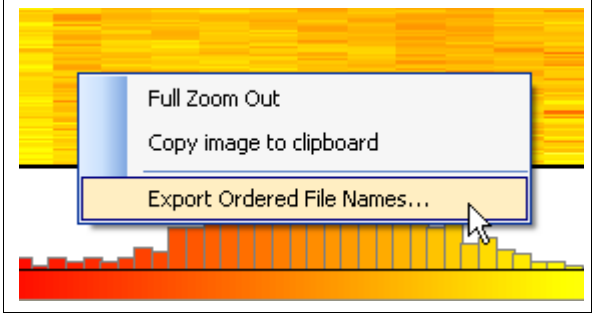


Exporting a List of Data Files in Sorted Order

To export a list of data files in their sorted order:

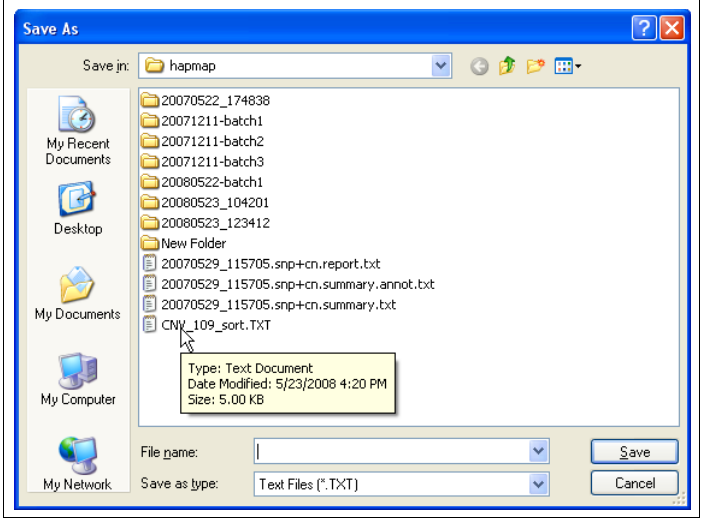
1. From the Heat Map menu, select **Export Ordered File Names..**; or
   Right-click in the heat map and select **Export Ordered File Names…** from the popup menu (Figure 16.26).

Figure 16.26 Heat Map shortcut menu



The Save As dialog box opens (Figure 16.27).

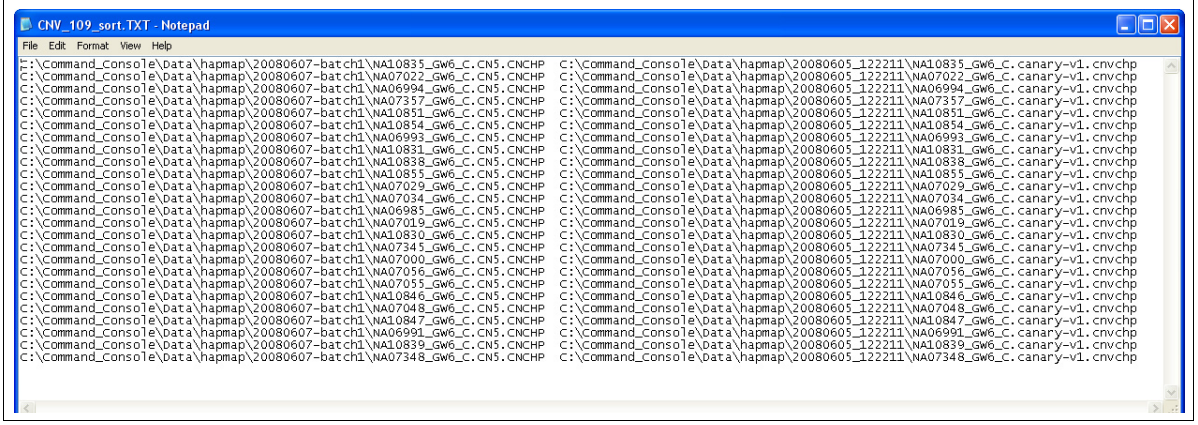
Figure 16.27 Save As dialog box



2. Select a location and enter a name for the file.
3. Click **Save** in the Save As dialog box.

   A text file (Figure 16.28) is created with the CNCHP and CNVCHP (if available) file path and file names.

Figure 16.28 List of sorted files in text format



Exporting Viewer Images

CN/LOH and CNV data cannot be exported directly from the Viewer, but can be exported using the export functions in GTC. See *Exporting CNV Data* on page 319 for more information.

GTC provides several ways to export a view of the Heat Map Views for use in a publication or to show other users.

You can:

- Print the Heat Map Viewer out.
- Export the image of the viewer to the clipboard.
- Export the image of the viewer to a PNG file

To print out the Heat Map viewer:

1. From the File Menu, select Print.
   The Print dialog box opens.
2. Select the printer and other options and click **OK** in the Print dialog box.

To export the image to the clipboard:

- Right-click in the heat map and select Copy image to clipboard from the popup menu; or
  From the Heat Map menu, select **Copy image to clipboard**.
  You can paste the image into a graphics file using software such as Paint.

To export the heat map image as a PNG file

1. From the Heat Map menu, select **Save image to file…**.
   A Save As dialog box opens.
2. Enter a name and location for the PNG file and click **Save**.
   The PNG file is created.

Viewing Regions in Public Data Sites

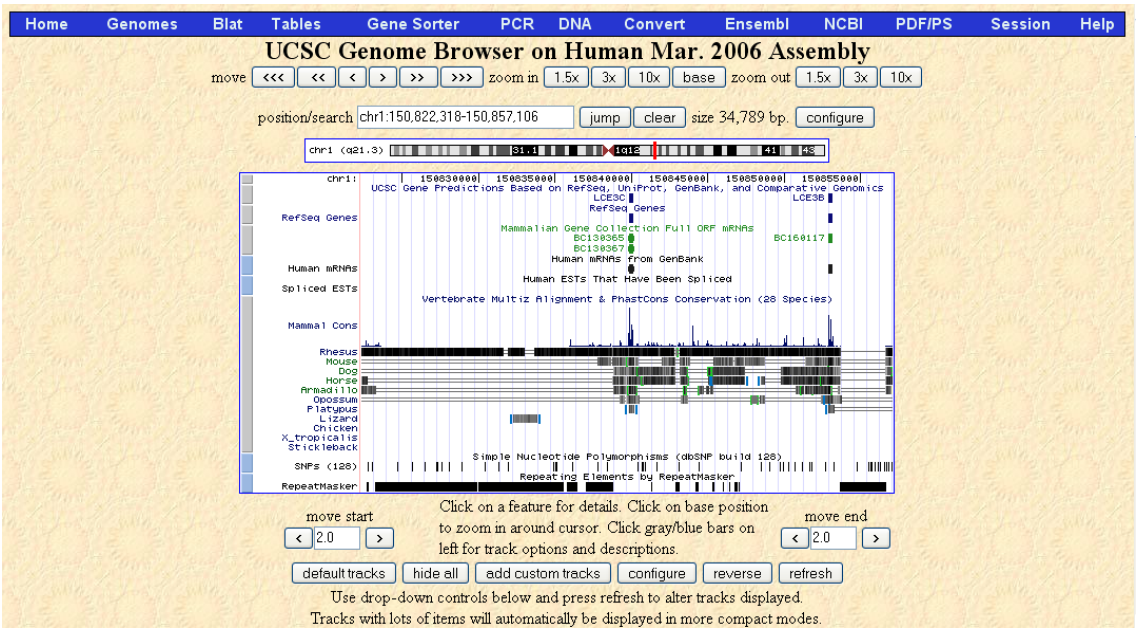
You can view the region selected in the Display area at one of the following public sites:

- UCSC Genome Browser
- Ensembl
- Toronto DGV

To view the selected region:

- From the Heat Map menu, select **External Links > [desired link]**.
  The external link will display the view using the genomic positions in the Heat Map viewer (Figure 16.29).

Figure 16.29 Display in the UCSC Genome Browser



Algorithms

The details of the algorithms used by GTC 4.2 and their typical performance are described in various white papers.

Genotyping

100K/500K BRLMM algorithm

http://www.affymetrix.com/support/technical/whitepapers/brlmm_whitepaper.pdf

SNP 5.0 arrays BRLMM-P algorithm

http://www.affymetrix.com/support/technical/whitepapers/brlmmp_whitepaper.pdf

SNP 6.0 Birdseed (v1) and Birdseed v2 genotyping algorithms

Genotyping Console 4.0 allows users to choose between genotyping SNP 6.0 array data with the Birdseed (v1) and the Birdseed v2 algorithms. Birdseed v2 uses EM to derive a maximum likelihood fit of a 2-dimensional Gaussian mixture model in A vs. B space.

A key difference between Birdseed (v1) and Birdseed v2 is that v1 uses SNP-specific models or priors only as an initial condition from which the EM fit is free to wander- on rare occasions this allows for mislabeling of the clusters. For Birdseed v2 the SNP-specific priors are used not only as initial conditions for EM, but are incorporated into the likelihood as Bayesian priors. This constrains the extent to which the EM fit can wander off. Correctly labeling SNP clusters, whose centers have shifted relative to the priors, is problematic for both Birdseed versions. However, given the additional constraint on the EM fit, Birdseed v2 is more likely than Birdseed to either correctly label the clusters or set genotypes to No Calls.

Birdseed v2 is usually more robust than Birdseed in the face of poor quality experiments, and increases accuracy with a small decrease in call rate in these cases. In high quality datasets, little performance difference between v1 and v2 is seen, while in low quality datasets large increases in concordance are seen with v2

Birdseed v2 clustering by plate is equivalent to clustering all samples, unlike Birdseed (v1) where clustering by plate increases False Discovery Rate. Because of this, use of Birdseed v2 allows clustering-by-plate or clustering all samples at once, which ever best fits with the laboratory's workflow.

See the Affymetrix.com website for information on Birdseed algorithms.

Axiom GT1 Algorithm

The Axiom GT1 method is a new genotyping procedure delivered in Genotyping Console 4.0 for use with the Axiom Genome-Wide Human array. The primary methodological change has been to incorporate multichannel processing into the APT workflow, supporting the ligation-based assay. In addition, Axiom GT1 incorporates substantial improvements and features in the areas of preprocessing and genotype calling over BRLMM-P which was used for the Genome-Wide SNP Array 5.0 (see *SNP 5.0 arrays BRLMM-P algorithm* on page 343). Many of the improvements in genotype calling were developed for the DMET Plus product, including 2-dimensional cluster modeling and outlier detection. Preprocessing has been improved by an artifact reduction layer which reduces the impact of spatially localized artifacts on genotyping performance. Together these changes allow for good genotyping performance on the ligation-based assay platform.

Multichannel processing allows the use of both traditional allelic differences, in which two different probes respond to the same region of sequence and distinguish alleles, as well as dye-based allele detection, in which the same probe is imaged in more than one channel to distinguish alleles. Both these workflows are handled in Genotyping Console 4.0 transparently to the user, and both types of probe strategy are used on the Axiom product.

The second area of improvement is in the genotype clustering and calling. Many of the improvements were developed in the course of the DMET Plus product and can be found described in the DMET Plus algorithm white paper:

http://www.affymetrix.com/support/technical/whitepapers/dmet_plus_algorithm_whitepaperv1.pdf

Briefly, clusters are now represented as 2-dimensional gaussians and resistance to non-gaussian cluster behavior has been improved. As usual, training data has been used to generate SNP-specific models which represent the cluster properties learned for each marker. Unlike DMET Plus which is designed to call in a single sample mode without adapting to the data, the default behavior is to use dynamic clustering to adapt the clusters to the observed data. Although a single sample can be run by itself, more samples allow more learning of any shifts from the training data.

Finally, the key advance in preprocessing is an "artifact reduction" layer that is designed to use information obtained from replicated probes to reduce the impact of small localized artifacts which sometimes occur. This method operates on the raw probe data using spatially distributed replicate probes to detect unusual differences between replicate intensities. Standard image processing operations (morphological transformations) are used to detect regions of the array where deviations occurring in both channels cluster, indicating a potential localized artifact. Once regions are marked as untrusted due to a potential artifact, intensities from trusted replicates are used to replace untrusted features for genotyping purposes. In the case where all replicates are marked untrusted for a given probe, the failsafe behavior is to leave the intensities unmodified and allow the genotyping method to evaluate whether the data is compatible with the clusters. This preprocessing layer improves the genotyping performance in the relatively rare case where localized artifacts occur on the image, while leaving typical arrays without artifacts unaffected.

Summarizing, Axiom GT1 handles multichannel data, incorporates improvements in genotype clustering and calling that have occurred in the development of other products, and introduces an artifact-reduction stage in preprocessing. These changes have been tuned to provide high performance on the ligation assay based genotyping platform and allow for flexible adaption of the method to future genotyping products.

Copy Number/LOH

100K/500K CN/LOH Algorithm

http://www.affymetrix.com/support/technical/whitepapers/cnat_4_algorithm_whitepaper.pdf

SNP 6.0 CN/LOH Algorithm

SNP 6.0 CN/LOH analysis uses the BRLMM-P+ algorithm, which is similar to BRLMM-P with some different parameters. See the existing documentation for BRLMM-P associated with SNP5 for more information.

SNP 6.0 CN GC waviness algorithm implemented into APT and since GTC 3.0.1

The summary of the algorithm correction is: for each sample, markers are divided into 25 different bins based on the equally spaced percentiles of the average GC count (GC content) in the upstream/downstream 250kb for a particular marker (500kb total). Within each of the 25 bins, the markers are sub-divided based on their type: CN/SNP marker type, enzyme fragment type (Nsp, Sty, Nsp+Sty), which gives 5 sub-bins per major bin as there is no CN probes in Sty-only fragments, for a total of 5x25=125 bins. For the autosomal markers in each bin, the median log2 ratio of each bin is adjusted to zero and interquartile ranges (IQRs) are equalized across all the bins. Then the log2 ratios of all markers (including X and Y markers) in that bin are adjusted using the adjustment based on the autosomal markers on that bin. Finally, the IQRs of all the adjusted log2 ratios (including the X and Y chromosomes) is multiplied by a factor that makes the IQRs of the adjusted log2 ratios equal to the IQRs of the original log2 ratios.

SNP 6.0 Canary Algorithm

The Canary Algorithm is a clustering algorithm developed by the Broad Institute used to provide copy number state calls of a pre-determined set of genomic regions with copy number variation (CNV regions). The copy number state call is reported by an integer call of copy number. Each call is paired with a confidence score between 0 and 1 with 1 reflecting a high level of confidence that the call is correct. The CNV regions are polymorphic in the sense that their copy number is atypically variable in relation to the genome as a whole. The terms copy number variation (CNV) and copy number polymorphism (CNP) are each used to describe the same attribute of copy number variability of genomic regions.

Inputs to the Canary algorithm are:

1. A region file containing region names and sets of SNP and CN probe sets for each region
2. A prior file containing clustering information empirically derived from external training data
3. A normalization file containing a list of names of probe sets used for normalizing the data
4. A set of CEL files, one for each sample to be genotyped.

A CDF file is needed by the software running Canary in order to retrieve probe sets intensities recorded in the CEL files.

Output consists of a set of CHP files, one for each CEL file, with the suffix CNVCHP. Each CHP file contains region names, intensities, calls and confidences.

Forward Strand Translation

The convention in the genomic research field has become to map allele genotypes to the forward strand of the genome. The convention used to select the reference strand to define Affymetrix alleles for Mapping 100K, 500K, SNP 5.0 and SNP 6.0 is based on an algorithm that alphabetically sorts the flanking-sequences for SNPs. They may be on either forward strand or reverse strand of the current genome. However, the relationship between Affymetrix alleles and the forward strand of the genome is provided in the publicly available NetAffx annotation files. For Axiom Genome-Wide Human Array, all Affymetrix alleles have been mapped to the forward strand of the current genome.

NetAffx defines allele A and allele B based on following convention: For AT or CG SNPs (SNP alleles are A/T or C/G), the alleles coded are in alphabetical order on that strand (allele A is C, allele B is G; or allele A is A, allele B is T). For non-AT and non-CG SNPs, allele A is A or T, allele B is C or G. For Axiom insertion/deletion alleles, allele A is '-', allele B is the insertion..

Table B.1 Affymetrix allele call codes defined by NetAffx convention

	CG SNP		AT SNP		Non-AT & Non-CG SNP		Insertion or Deletion **(Axiom™ Genome-Wide Human Array Only)**	
Base, Insertion or Deletion	C	G	A	T	A or T	C or G	Deletion (-)	Insertion (+)
Allele	A	B	A	B	A	B	A	B

For example, rs4607103 (SNP_A-2091752 on the Genome-Wide SNP Array 6.0) is a non-AT and non-CG SNP oriented on the reverse strand at position 64686944 on chromosome 3 (build 36.1). GTC 4.2 uses this information to provide the forward strand base call (Table B.2).

Table B.2 Example forward strand translation for SNP_A-2091752 (Genome-Wide SNP Array 6.0)

SNP_A-2091752	Annotation File Reverse Strand	Forward Strand Translation	SNP_A-2091752			
Allele A	A	T	Affymetrix Allele Call Codes	AA	AB	BB
Allele B	G	C	Translated Forward Strand Base Calls	TT	TC	CC

Advanced Workflows

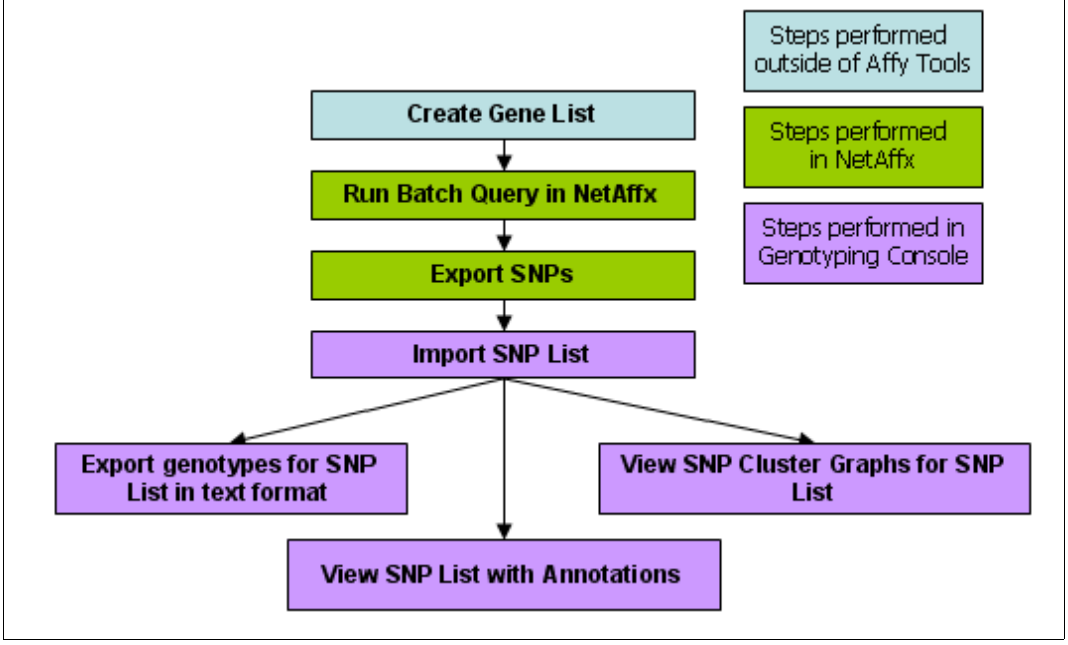
This Appendix describes the following Advanced Workflows:

- *Analyzing Genotyping Results of Specific Gene Lists*

Analyzing Genotyping Results of Specific Gene Lists

The figure below (Figure C.1) shows the basic steps on how to get SNP information for a specific set of genes and analyze those SNPs in Genotyping Console.

Figure C.1 Workflow to analyze specific gene lists



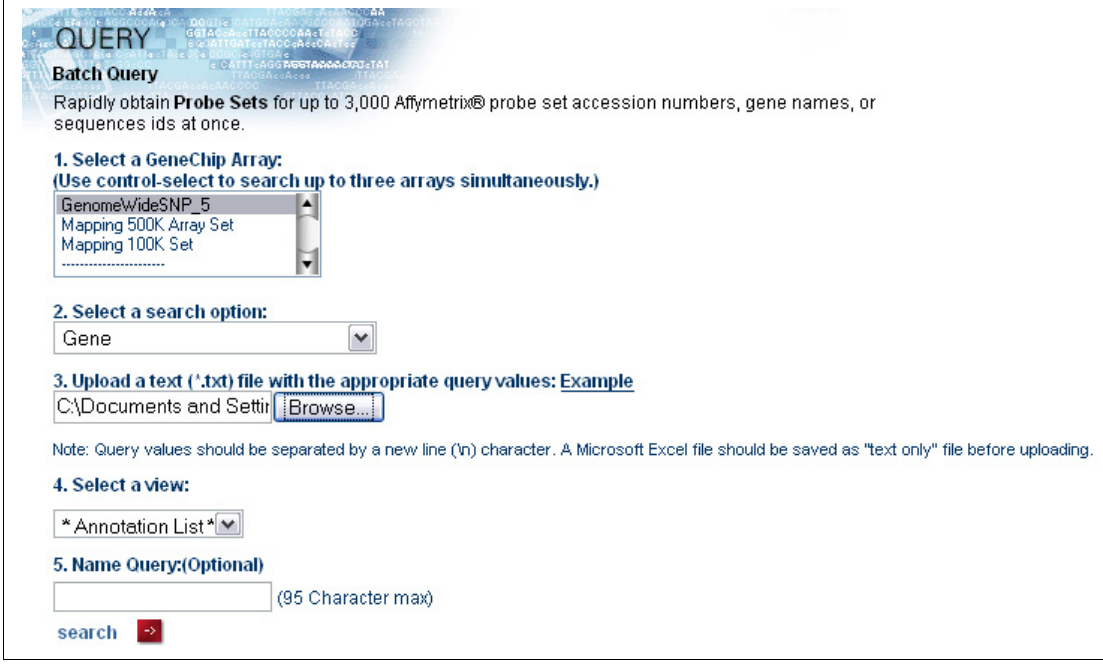
Step 1: A list of genes is generated. Perhaps the gene list contains a set of biologically relevant genes (e.g. kinases).

The list of genes must be contained in a text file where each gene ID is on a separate line.

Step 2: Using NetAffx, perform a batch query to identify SNPs which are mapped to the location of the specified genes in the list.

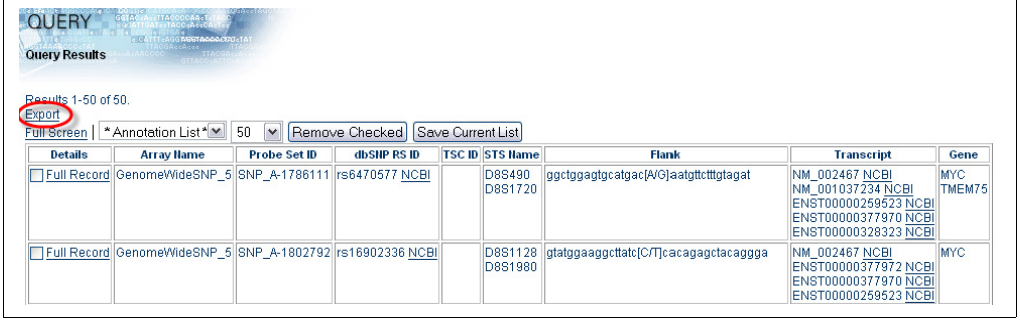
1. Login to NetAffx website (*http://www.affymetrix.com/analysis/index.affx*)
2. Select Genotyping Batch Query
3. Select the array type, search option, gene list file, and view.

Figure C.2 Query in NetAffx



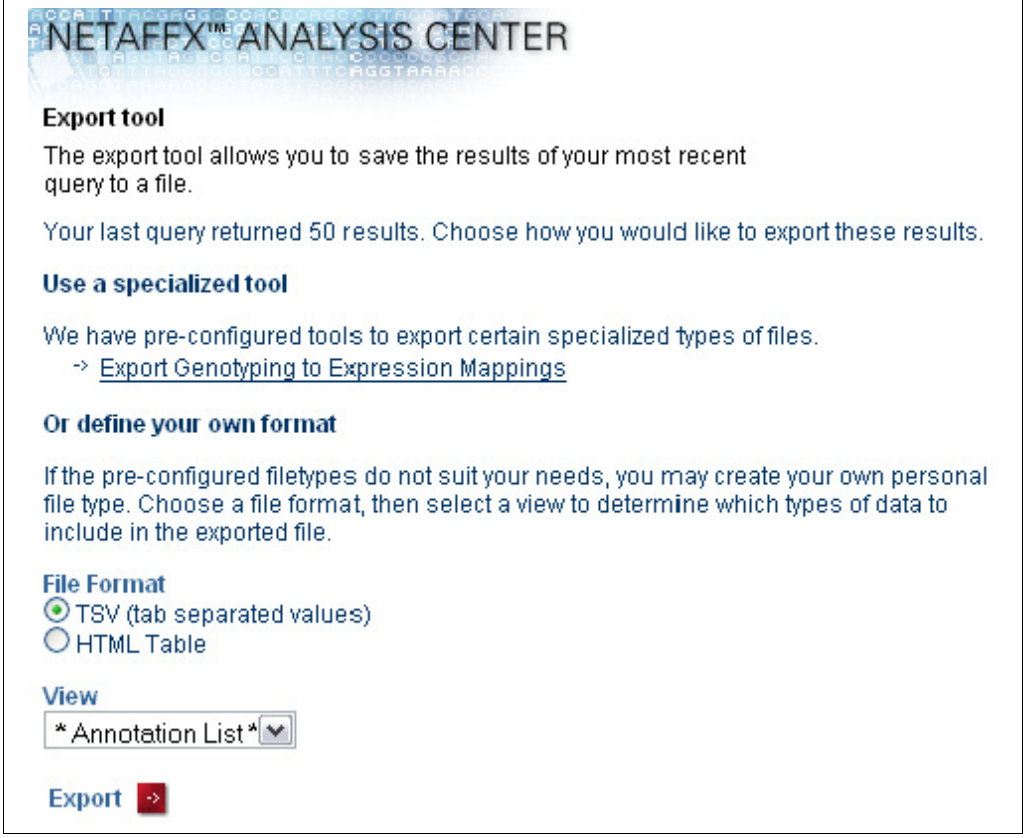
4. Click on search.
   NetAffx will identify all SNPs which are mapped to the specified genes.
5. Click on the Export button.

Figure C.3 Export button



6. Select the TSV export option.

Figure C.4 Export options



7. Click Export.

Step 3: Open Genotyping Console and import the SNP List generated by NetAffx.

8. Right-click on SNP Lists.
9. Select Import SNP List.
10. Migrate to the location of the TSV file generated by NetAffx and Select Open.
11. Provide a name for the SNP List to be displayed in Genotyping Console and Select **OK**.
    The SNP List will be displayed in the data tree.

Step 4: After the SNP List is imported in Genotyping Console, the SNP List can be used for many different functions:

- *View the SNP List* (page 129)
- *Exporting genotypes for SNP in the list* (page 184)
- *View the SNP Cluster Graph for SNPs in the list* (page 153)

Annotation Definitions

Table D.1 Annotation Definitions

Column Name	Description
Probe Set ID	The Affymetrix unique identifier for the set of probes used to detect a particular Single Nucleotide Polymorphism (SNP probe sets only).
Affx SNP ID	The Affymetrix unique identifier for the set of probes used to detect a particular Single Nucleotide Polymorphism (SNP). (SNP probe sets only, not available for Axiom™ Genome-Wide Human Array).
dbSNP RS ID	The dbSNP ID that corresponds to this probe set or SNP. The dbSNP at the National Center for Biotechnology Information (NCBI) attempts to maintain a unified and comprehensive view of known single nucleotide polymorphisms (SNPs), small scale insertions/deletions, polymorphic repetitive elements, and microsatellites from TSC and other sources. The dbSNP is updated periodically, and the dbSNP version used for mapping is given in the dbSNP version field. For more information, please see: *http://www.ncbi.nlm.nih.gov/SNP/* (SNP probe sets only).
Chromosome	The chromosome on which the SNP is located on the current Genome Version.
Chromosome Start	The nucleotide base start position where the SNP is found. The genomic coordinates given are in relation to the current genome version and may shift as subsequent genome builds are released.
Chromosome Stop	The nucleotide base stop position where the SNP is found. The genomic coordinates given are in relation to the current genome version and may shift as subsequent genome builds are released.
Strand	Genomic strand that the SNP resides on.
Cytoband	Cytoband location of the SNP derived from the SNP physical map and the chromosome band data provided by UCSC.
Strand Vs dbSNP	Indicates whether the SNP is on the same or reverse strand as compared to dbSNP (SNP probe sets only).
ChrX pseudo-autosomal region	SNPs on the X Chromosome which are mapped to the two pseudo-autosomal region have a value of 1 or 2 in this field. All other SNPs are indicated by 0. A value of "1" indicates that the marker maps to the PAR-1 region and a value of "2" indicates that the marker maps to the PAR-2 region. A value of "0" indicates that the marker does not map to either of the two PAR regions.
Probe Count	The total number of probes in the probe set.
Flank	The nucleotide sequence surrounding the SNP. This is a 33-mer sequence with 16 nucleotides on either end of the SNP position. The alleles at the SNP position are provided in the brackets (SNP probe sets only).
Allele A	The allele of the SNP that is in lower alphabetical order. When comparing the allele data on NetAffx to the allele data for the corresponding RefSNP record in dbSNP, the alleles reported here could be different from the alleles reported for the corresponding RefSNP on the dbSNP web site. This difference arises mainly from the reference genomic strand that was chosen to define the alleles by Affymetrix. To choose the reference genomic strand, we follow a convention based on the alphabetic ordering of the sequence surrounding the SNP. Sometimes the reference strand on the dbSNP is different from NetAffx, and the alleles could represent reverse complement of those provided on dbSNP (SNP probe sets only).

Table D.1 Annotation Definitions

Column Name	Description
Allele B	The allele of the SNP that is in higher alphabetical order. When comparing the allele data on NetAffx to the allele data for the corresponding RefSNP record in dbSNP, the alleles reported here could be different from the alleles reported for the corresponding RefSNP on the dbSNP web site. This difference arises mainly from the reference genomic strand that was chosen to define the alleles by Affymetrix. To choose the reference genomic strand, we follow a convention based on the alphabetic ordering of the sequence surrounding the SNP. Sometimes the reference strand on the dbSNP is different from NetAffx, and the alleles could represent reverse complement of those provided on dbSNP (SNP probe sets only).
Associated Gene	SNPs were associated with human genes by comparing the genomic locations of the SNPs to genomic alignments of human mRNA sequences. In cases where the SNP is within a known gene, NetAffx reports the association. Additionally, for genes with exon or CDS annotations, NetAffx reports whether or not the SNP is in an exon, and in the coding region. If the SNP is not within a known gene, NetAffx reports the closest genes in the genomic sequence, and the distance and relationship of the SNP relative to the genes. A SNP is upstream of a gene if it is located closer to the 5' end of the gene and is downstream of a gene if it is located closer to the 3' end of the gene.
Genetic Map	Describes the genetic location of the SNP derived from three separate linkage maps (deCODE, Marshfield, or SLM). The physical distance between the markers is assumed to be linear with their genetic distance. The genetic location is computed using the linkage maps from the latest physical location of the SNP and the neighboring microsatellite markers (SNP probe sets only).
Microsatellite	Describes the nearest microsatellite markers (upstream, downstream and overlapping) for the SNP.
Enzyme Fragment	Lists the enzyme, the restriction fragment containing the SNP and the fragment length. The Whole Genome Assay protocol detects SNPs that are contained within the genomic restriction fragments to simplify the sequence background for genotyping arrays (not available for Axiom Genome-Wide Human Array).
Copy Number Variation	When available, a description of Copy Number Variation Region (CN) probe sets as described by the Database of Genomic Variants (not available for Axiom Genome-Wide Human Array).
SNP Interference	This column is for Copy Number probe sets. It indicates whether or not a known SNP overlaps a copy number probe (CN probe sets only, not available for Axiom Genome-Wide Human Array).
In Final List	This column annotates extended content for genotyping arrays. A value of "1" indicates that the marker is included in the final version of the library file and a value of "0" indicates that the marker is not included in the final version of the library file (SNP probe sets only, not available for Axiom Genome-Wide Human Array).
% GC	The fraction of bases that are G or C in a window of 250,000 bases to each side of the SNP or CN position. All positions that are nearer to the end than 250,001 are set to the value of the position at 250,001 from that end. Position and chromosome values for SNPs and CN probes were mapped to the position of bases in the FASTA files for the build of the genome used in this release of NetAffx, and these bases were then used for all calculations (not available for Axiom Genome-Wide Human Array).
Heterozygous Allele Frequencies	Describes the heterozygous frequency of the allele from Yoruba, Japanese, Han Chinese and CEPH studies using the Affymetrix genotyping arrays. (SNP probe sets only)

Table D.1 Annotation Definitions

Column Name	Description
Allele Sample Size	Sample size used for Allele Frequency estimates (SNP probe sets only).
Allele Frequencies	Describes the major and minor frequency of the allele from Yoruba, Japanese, Han Chinese and CEPH studies using the Affymetrix genotyping arrays (SNP probe sets only).
Minor Allele	Indicates the Minor Allele of a SNP (SNP probe sets only).
Minor Allele Frequency	The Minor Allele Frequency of a SNP (SNP probe sets only).
OMIM ID	Furnishes OMIM and Morbid Map IDs and their respective gene titles. This database contains information from the Online Mendelian Inheritance in Man® (OMIM®) database, which has been obtained under a license from the Johns Hopkins University. This database/product does not represent the entire, unmodified OMIM® database, which is available in its entirety at www.ncbi.nlm.nih.gov/omim/.

Gender Calling in GTC

GTC 4.2 can generate gender calls from:

- Intensity QC
- Genotyping Analysis
- CN Segment Report (for SNP 6.0 only)

Copy number analysis for SNP 6.0 arrays provides information about calls for the X chromosomes and about calls for the Y chromosome based on signal intensity and allelic ratio, and provide a gender call (Female or Male) in the output table.

The processes used for gender calling differ depending upon:

- The type of array being analyzed.
- Step in the workflow being performed

Gender Calls in Intensity QC

See Chapter 6, *Intensity Quality Control for Genotyping Analysis* on page 74 for information on the algorithm used for the Intensity QC step.

QC analysis for genotyping uses DM algorithm to make SNP calls for Intensity QC purposes. It uses the following processes for making the gender call during this step.

Contrast QC is the recommended QC metric for the SNP 6.0 array in Genotyping Console 3.0.1. The default threshold is “greater than or equal to 0.4” for each sample. When adjusting this QC metric’s threshold value, or changing SNP 6.0 QC settings to another metric such as QC Call Rate, or adding additional metrics to threshold, a flag in the configuration setting dialog box will indicate that the thresholds are different than the defaults.

Contrast QC is a metric that captures the ability of an experiment to resolve SNP signals into three genotype clusters. It uses 10,000 random SNP 6.0 SNPs. See Appendix F, *Contrast QC for SNP 6.0 Intensity Data* on page 358 for more details.

Gender Calls in Intensity QC and Genotyping Analysis

The table below (Table E.1) summarizes the methods used for gender calls during Intensity QC and genotyping analysis.

Table E.1 Gender calling methods

Array Type	Gender Call Algorithm	Genotyping Algorithm Gender Call	Reference
Axiom™ Arrays	cn-probe-chrXY-ratio_gender	Yes: • Male • Female • Unknown	See below
Genome-Wide Human SNP Array 6.0	cn-probe-chrXY-ratio_gender	Yes: • Male • Female • Unknown	See below

Table E.1 Gender calling methods

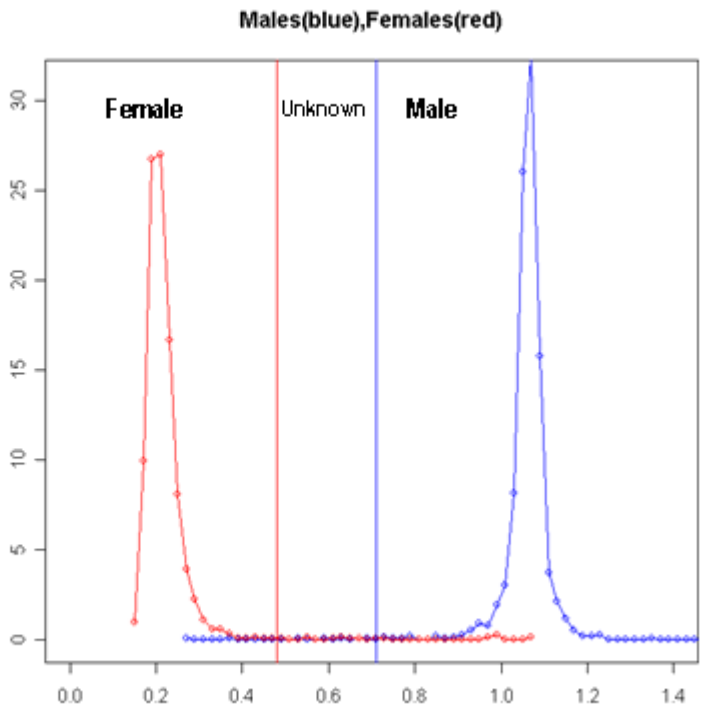
Array Type	Gender Call Algorithm	Genotyping Algorithm Gender Call	Reference
Genome-Wide Human SNP Array 5.0	em-cluster-chrX-het-contrast_gender	Yes: • Male • Female • Unknown	BRLMM-P white paper
Human Mapping 100K/500K Arrays	estimated heterozygosity rate on the X chromosome	Yes: • Male • Female	BRLMM white paper

Genotyping Gender Call Process: cn-probe-chrXY-ratio_gender

In GTC 4.2 the gender calling algorithm used to populate the "Computed Gender" call in the "Intensity QC Table" and the "CHP Summary Table" for SNP 6.0 and Axiom arrays is called cn-probe-chrXY-ratio_gender method from Affymetrix Power Tools (APT). The cn-probe-chrXY-ratio_gender method is more robust when dealing with lower quality samples. Optimal genotyping of sex chromosome SNPs requires use of the correct model type, haploid or diploid. Haploid models are used for X and Y chromosome SNPs, when the gender call is "male", while diploid models are used for X chromosome SNPs, when the gender call is "female". A "No Call" is made for Y chromosome SNPs when the gender call is female.

The cn-probe-chrXY-ratio_gender method determines gender based on the ratio (cn-probe-chrXY-ratio_gender_ratio) of the average probe intensity of nonpolymorphic probes on the Y chromosome (cn-probe-chrXY-ratio_gender_meanY) to the average probe intensity of nonpolymorphic probes on the X chromosome (cn-probe-chrXY-ratio_gender_meanX). The probe intensities are raw and untransformed for these calculations, and copy number probes within the pseudoautosomal regions (PAR region) of the X and Y chromosomes are excluded. For SNP 6.0 arrays, if the ratio is less than 0.48, the gender call is female; and if it is greater than 0.71, the gender call is male. If the ratio is between these values, the gender call is unknown. For Axiom™ Genome-Wide Human arrays, if the ratio is less than 0.54, the gender call is female, and if it is greater than 1.0, the gender call is male. If the ratio is between these values, the gender call is unknown.

Figure E.1 The SNP 6.0 frequency distribution of the Gender Y/X ratio for over 1500 male (blue) and 1500 female (red) samples without filtering based on QC callrate is shown here. The locations of the lower cutoff (red line) and upper cutoff (blue line) are shown, and regions corresponding to three possible gender calls are labeled Female, Unknown, and Male.



The cn-probe-chrXY-ratio_gender method produces "Unknown" gender calls for poor quality samples. However in extreme cases, where the sample has essentially no signal, the gender call will be male. Such experiments are easily identified by examining the QC CallRate.

The cn-probe-chrXY-ratio_gender method classifies genders considering only two possible cases, male: XY and female: XX. However, unusual genders such as XXX, XO, XXY, and XYY occur at low rates in populations along with X chromosome mosaicism, a variable loss or gain of the X chromosome known to happen sometimes both in vivo and in cell lines. To help detect and identify these unusual genders four additional gender columns can displayed in the CHP Summary Table by selecting "Show All Data". The four additional columns are:

em-cluster-chrX-het-contrast_gender_chrX_het_rate
The estimated heterozygosity rate (% AB genotypes) of SNPs on the X chromosome.

cn-probe-chrXY-ratio_gender_meanX
The average probe intensity (raw, untransformed) of X chromosome nonpolymorphic probes

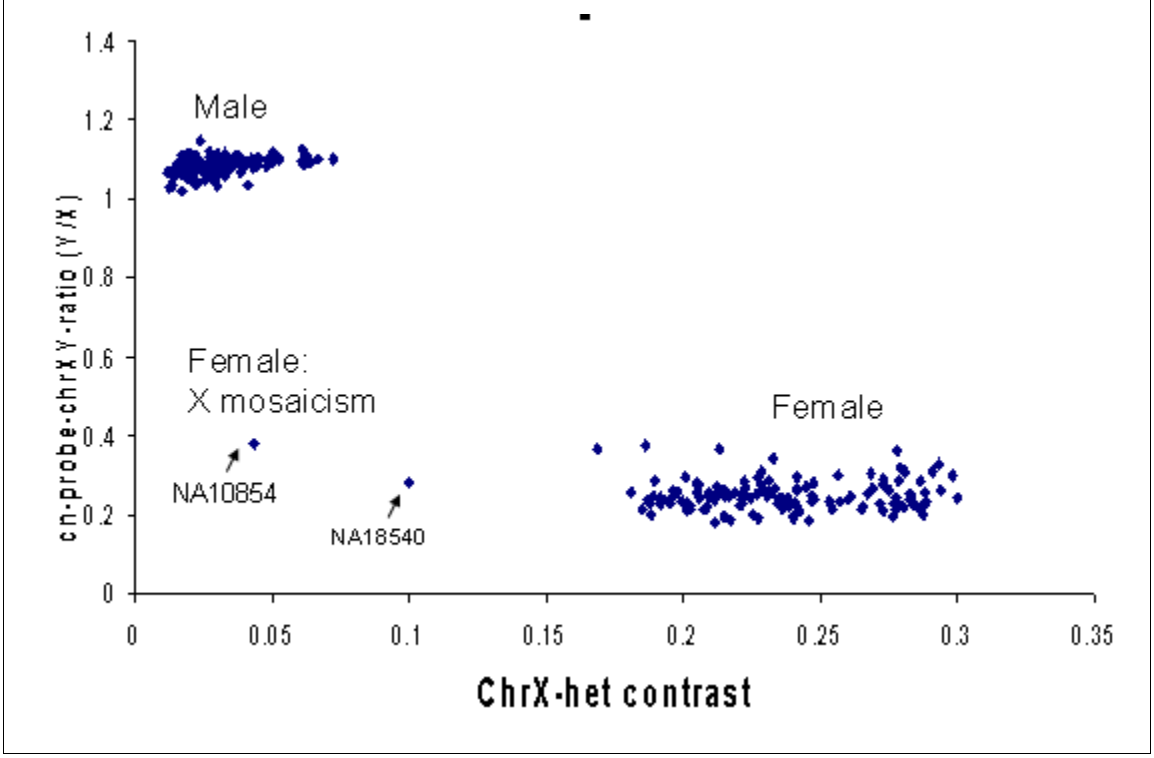
cn-probe-chrXY-ratio_gender_meanY
The average probe intensity (raw, untransformed) of Y chromosome nonpolymorphic probes

cn-probe-chrXY-ratio_gender_ratio
Gender ratio Y/X = cn-probe-chrXY-ratio_gender_meanY/ cn-probe-chrXY-ratio_gender_meanX

NOTE: **SNP 6.0 CHP files created with GTC 1.0 will not contain these data columns, one must genotype the files again using GTC 2.0 or above for them to be calculated.**

Scatter plots of em-cluster-chrX-het-contrast_gender_chrX_het_rate vs. cn-probe-chrXY-ratio_gender_ratio should contain two main clusters of points, one for males and one for females. Samples with unusual genders are expected to fall outside of the two main clusters indicating possible deviations from normal sex chromosome copy numbers. The figure below shows the this scatter plot for the 270 HapMap individuals. Sample NA10854 and NA18540 fall outside of the usual gender clusters. Previous work has demonstrated that NA10854 is known to have a significant degree of X mosaicism (BMC Bioinformatics 2006, 7:25) and that sample NA18540 has X chromosome mosaicism as well as aneuploidy in several other chromosomes (Am. J. Hum. Genet., 79:275-290, 2006)

Figure E.2 Gender Metrics for 270 HapMap samples on SNP 6.0



Gender Calls (Female or Male) in Copy Number Analysis (SNP 6.0 only)

Copy number analysis for SNP 6.0 data provides an actual gender call (Female or Male).

The gender is determined using the same method as in the SNP 6.0 genotyping gender call process described above, using the ratio of chrX to chrY nonpolymorphic probes.

CN Segment Report (SNP 6.0 only)

For SNP 6.0 Arrays the Segment Reporting Tool makes a gender determination for the sample, based on the detected copy number state for the X and Y chromosomes. Normal males and females are expected to have Copy Number State=2 for autosomes1-22. Females are expected to have Copy Number State =2 for the X chromosome, while normal males are expected to have Copy Number State=1 for the X chromosome and =1 for Y chromosome.

First the algorithm checks that the Copy Number QC metric MAPD is less than 0.5 to ensure the data is of sufficient quality. Next the mean copy number for the non-pseudo autosomal portion of the X chromosome and Y chromosome are used to assign gender. If the mean copy number for the X chromosome is between 0.8 to 1.3 and the mean copy number for Y is between 0.8 to 1.2, then a "male" is assigned. If the mean copy number for X is from 1.9 to 2.1 and Y is from 0 to 0.4, then a "female" is

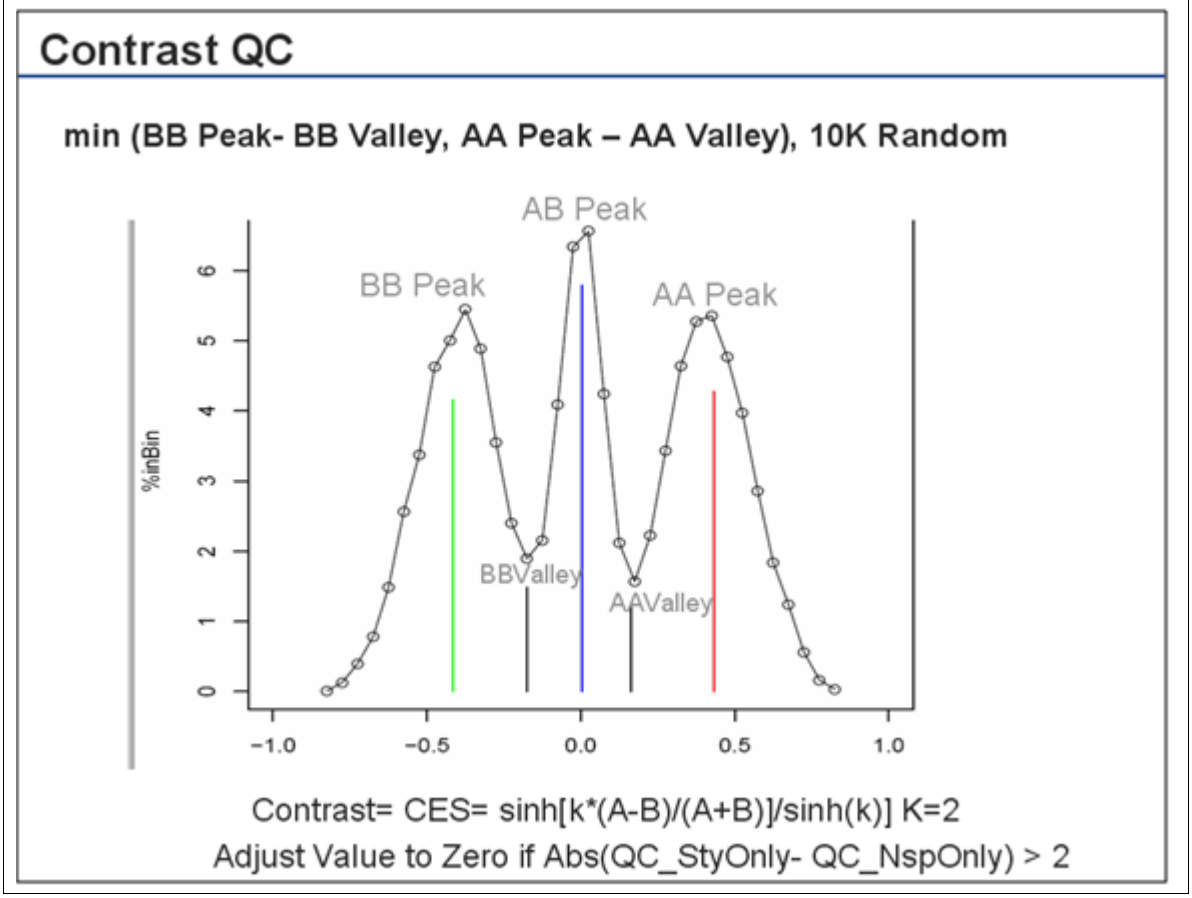
assigned. Finally, if neither of the above cases are true, then "Unknown" is assigned. Samples flagged “Unknown ” by the software and are assessed for Copy Number change as if they were female (CN State for X=2, and Y=0).

Contrast QC for SNP 6.0 Intensity Data

Contrast QC is the per sample Quality Control test metric for SNP 6.0 intensity data (CEL files). When all steps of the assay are working as expected, the Contrast QC is typically greater than 0.4. As an added flag for potentially problem data sets, check that the proportion of samples that fall below the 0.4 threshold are less than 10%, and the average of the samples that pass this 0.4 test are greater than or equal to 1.7. If the proportion falling below 0.4 is greater than 10%,or the average of the passing samples is at or below 1.7, then sample quality and process should be closely examined for possible issues.

The Contrast QC is a metric that captures the ability of an experiment to resolve SNP signals into three genotype clusters. It uses a static set of 10,000 randomly chosen SNP 6.0 SNPs, measuring the difference between peaks in “Contrast” distributions (Figure F.1) produced by homozygote genotypes, and the valleys they share with the heterozygote peak, and takes the smaller of the two values. In poor quality experiments the homozygote peaks are not well-resolved from the heterozygote peak and the difference values approach zero. Contrast QC values are also computed for Contrast distributions produced by a static set of 20K randomly chosen SNPs on Nsp fragments only and a static set of 20K randomly chosen SNPs on Sty fragments only. These are called Contrast QC (Nsp) and Contrast QC (Sty); respectively. If the absolute difference between these two values is greater than two, this is evidence that that a sample may have worked properly with one enzyme set, but not with the other, and the Contrast QC value is adjusted to zero to reflect this problem. These Contrast QC values are well correlated with the higher Call Rates and concordance achieved when calls are subsequently made with Birdseed (versions 1 or 2). The correlation between Birdseed accuracy and Birdseed Call Rate is also very high. As an extra guard against the inclusion of any outlier samples that pass through the Contrast QC filter, it is a good idea to reject samples that are notable outliers in terms of their Birdseed Call Rate. When using Birdseed (v1), clustering larger batches of samples will improve the performance of the algorithm. The algorithm improvements in Birdseed v2 allow you to cluster by plate with the same performance as clustering larger batches of samples.

Figure F.1 Distribution of Contrast Values. The X axis is the Contrast Value about which a bin of size 0.02 is centered. The Y axis is the %of SNPs (10000 random autosomal GW 6 SNPs) whose Contrast values fall within the bin. Contrast = sinh[K*(A-B)/(A+B)]/sinh(K)], K=2, A and B are the summary values for probes covering the A and B alleles; respectively (see http://www.affymetrix.com/support/technical/whitepapers/brlmm_whitepaper.pdf).



The Contrast QC is adjusted to zero if abs[Contrast QC (Nsp)- Contrast QC (Sty)] > 2

Best Practices SNP 6.0 Analysis Workflow

1. Study Design:
   - Where possible, randomization of cases and controls across sample plates is usually a good idea.
   - In studies involving trios, it is usually good to try to ensure that all three members of a trio are on the same sample plate.
2. Pre-Cluster Sample Quality Check
   - Reprocess samples with Contrast QC < 0.4
3. Pre-Cluster Plate or Dataset Check
4. Genotyping: Cluster Samples with Birdseed v2
   - Cluster by plate or cluster all together according to which process is most convenient for the lab workflow
   - Each cluster should contain a minimum of 44 samples with a least 15 female samples
5. Genotyping: Post-Cluster Sample Quality Check
   - Reject samples with outlier low Birdseed Callrates
   - Reject samples with excess predicted heterozygosity
6. Genotyping: Post-Genotyping SNP Filtration
   - Filter for SNPs with high SNP callrates over all samples in the study; somewhere in the range of 90-95%
   - The exception is Y chr SNPs- which are always NoCalls for Female samples
   - May also want to reject based on deviation from HW equilibrium, reproducibility, where possible and appropriate
7. Genotyping: Post-Association Study Analysis
   - Visually analyze all candidate SNPs
8. Copy Number: Reference Model File Creation
   - Set of samples used to create Reference Model File should contain a minimum of 44 samples with a least 15 female samples
9. Copy Number: CNCHP file Quality Check
   - Track CNCHP quality using MAPDs. Reprocess samples with MAPDs greater than 0.3 when using an intra-lab reference (Reference Model File made from lab's own samples) or greater than 0.35 when using an external reference (Reference generated elsewhere, such as the supplied 270HapMap Reference).
   - If MAPDs are consistently high when using an external reference, recalculate MAPDs with an intra-lab reference. If the MAPDs all drop significantly, then the high MAPD is an artifact introduced by a systematic difference between current samples and the samples that made up the reference rather than a quality issue.

Best Practices Axiom Analysis Workflow

1. Study Design
   - Where possible, randomization of cases and controls across sample plates is usually a good idea.
   - In studies involving trios, it is usually good to try ensure that all three members of a trio are on the sample plate.
2. Pre-Cluster Sample Quality Check:
   - Exclude/reprocess samples with Dish QC < 0.82
3. Genotyping, preliminary round: Cluster Samples with Axiom GT1
   - Cluster by 96 well plate or cluster all together according to which process is most convenient for the lab workflow
   - Each cluster should contain a minimum of 20 distinct samples with either zero females samples or at least 10 distinct female samples
   - Each cluster should contain a minimum of 90 distinct samples with either zero female samples or at least 30 distinct female samples when generic prior is used for Axiom myDesign™ arrays
4. Post-Cluster Sample Quality Check
   - Reject samples with clustering call rates less than 97%
   - Reject samples with excess predicted heterozygosity. What exactly constitutes an outlier will depend on the population. It is often useful to plot the heterozygosity against the sample call rate, often outlier samples will have unusual call rate/heterozygosity combinations. Note also that because Genotyping Console reports heterozygosity including chrX markers, females will generally have slightly higher heterozygosity than males.
5. Plate level quality check
   - For each plate, check the overall sample failure rate and the distribution of performance (DQC & call rate) for passing samples. Any plate with an unusually high number of failures or a striking shift in performance of passing samples should be considered carefully. The key goal would be to distinguish between the possibility of a plate-wide issue that may still affect even the passing samples as opposed to a sample-specific issue that affects just a specific subset of experiments.
6. Genotyping, final round
   - Repeat genotype clustering after rejection of any outlier samples identified in the preliminary round of clustering.
7. Post-Genotyping SNP Filtration
   - Exclude SNPs with low SNP call rates, evaluated over all passing samples in the study; somewhere in the range of 90-95% is typical
   - The exception is Y chr SNPs which are always NoCalls for Female samples
   - You may also want to reject based on deviation from HW equilibrium (in controls), reproducibility and Mendelian Inheritance errors where possible and appropriate
8. Post-Association Study Analysis
   - Visually inspect cluster plots for all candidate SNPs to ensure that there is nothing unusual about the clustering

Copy Number Variation Analysis

Copy Number Variation Analysis is performed using the Canary algorithm which was developed by the Broad Institute for the purpose of making copy number state calls for genomic regions with copy number variations. These genomic regions can be called regions with copy number variation (CNV regions) or regions with copy number polymorphism (CNP). These CNV regions are observed to be more variable in regard to copy number states than is typical of the genome as a whole. The specialized algorithm, Canary, was developed for these CNV regions because other copy number analysis methods assume a copy number of 2 to be the predominant copy number state in a sample of individuals. This frequency assumption is not reliable in the CNV regions and can lead to misled copy number state calls in the set of samples as a whole.

The Broad Institute first identified and made copy number state calls for the CNV regions in the population of HapMap samples. For each of these regions a set of probe sets, deemed to be “smart”, was assigned. Fidelity and robust response are two criteria attributed to smart probe sets. Within each CNV region selected by the Broad Institute, summaries of smart probe sets resulted in a clustering pattern consistent with copy number state. The frequency of HapMap individuals with a certain copy number state as well as cluster centers and means was recorded as empirical prior clustering estimates used by Canary. In GTC 4.2, the sets of smart probe sets mapping to CNV regions are stored a region file and the prior cluster information is stored in a prior file. All smart probe sets in the region file correspond to NCBI build 36.1 of the human genome. The CNV regions with their corresponding chromosomal positions are recorded in the CNV map file. This CNV map file is required for CNV result table display and also for the heat map viewer display.

GTC 4.2 uses a set of 1141 CNV regions derived from those identified by the Broad. To reduce sample-to-sample variability, these 1141 CNV regions are a subset filtered to ensure that each CNV region is mapped to by more than one smart probe set, reduced by restriction enzymes into more than one fragment and produces clustering results consistent in two full sets of HapMap samples independently processed at separate sites.

Hard Disk Requirements

This appendix provides example hard disk requirements for 450 CEL files from different types of arrays and analyses. The temp folder is required for analysis and the result folder is required to save data.

Table J.1 Hard disk (HD) requirements for 450 CEL files (temp folder and results folder on different hard disks)

Temp Folder HD			Result Folder HD	
Type of Analysis	**Per CEL File (MB)**	**Total GB Required (450 CEL files)**	**Per CEL File (MB)**	**Total GB Required (450 CEL files)**
SNP 6.0 Genotyping	83.54	~38	65.87	~30
SNP 6.0 CN/LOH	83.54	~38	78.10	~36
SNP 6.0 CNV	83.54	~38	0.046	<1
Axiom Genotyping	34.33	~16	22.58	~11

Table J.2 Hard disk (HD) requirements for 450 CEL files (temp folder and results folder on the same hard disk)

Temp Folder			Result Folder		HD (GB)
Type of Analysis	**Per CEL File (MB)**	**Total GB Required (450 CEL files)**	**Per CEL File (MB)**	**Total GB Required (450 CEL files)**	
SNP 6.0 Genotyping	83.54	~38	65.87	~30	~68
SNP 6.0 CN/LOH	83.54	~38	78.10	~36	~74
SNP 6.0 CNV	83.54	~38	0.046	<1	~39
Axiom Genotyping	34.33	~16	22.58	~11	~27

Axiom CNV Summary Tool and Viewer

The Axiom CNV Summary Tool generates input files for BioDiscovery Nexus using Axiom data.

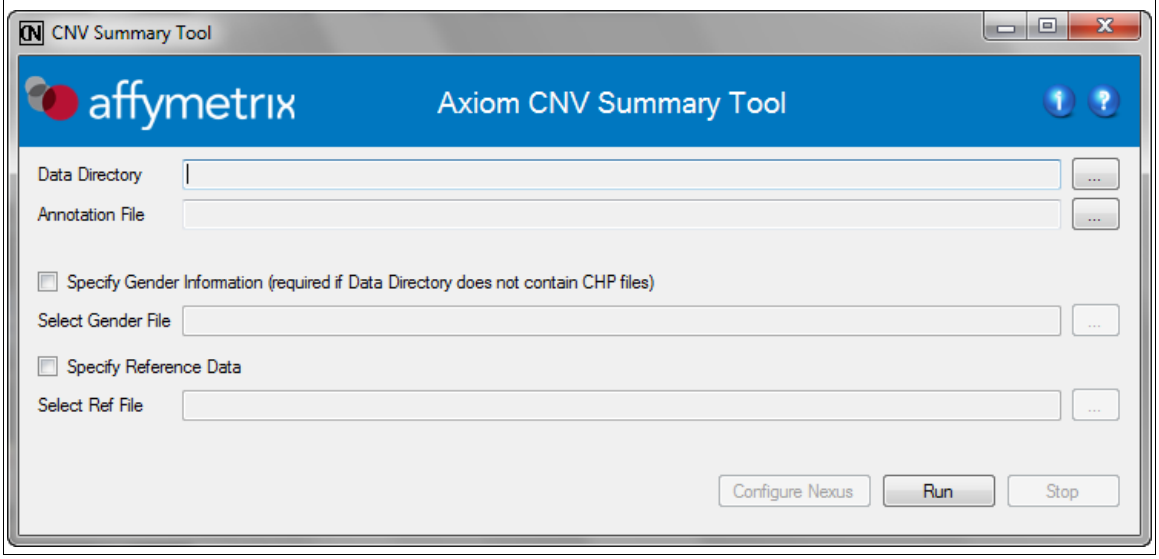
The included Axiom CNV Viewer allows you to view the data generated by the Axiom CNV Summary Tool. To use the Viewer, see *Using the Axiom CNV Viewer* on page 367.

Using the Axiom CNV Summary Tool

To start the Axiom CNV Summary Tool:

1. Click **Tools -> Axiom CNV Summary Tool**.
   The following window appears: (Figure K.1)

Figure K.1 CNV Summary Tool Main window



2. Click the **Data Directory** Browse button.
3. Navigate to the folder that contains your input data files (*AxiomGT1.CHP, AxiomGT1.calls.txt, and AxiomGT1.summary.txt files), then single click, Ctrl click, Shift click, or Ctrl-A (to select multiple files).
4. Click **OK**.
   The Data Directory path is now populated.
5. Click the **Annotation File** Browse button.
6. Navigate to the folder that contains the annotation file you downloaded earlier from www.affymetrix.com.

NOTE: Annotation files are array specific. If you are running an analysis for a specific array, make sure you use the appropriate annotation file.

Annotation files for *Axiom myDesign* arrays are provided directly to you from Affymetrix (for each custom array designed).

When you download an annotation file using GTC, the annotation file is automatically indexed for optimum processing.

7. Click to select the annotation file, then click **OK**.
   The Annotation File path is now populated

Select Gender File.

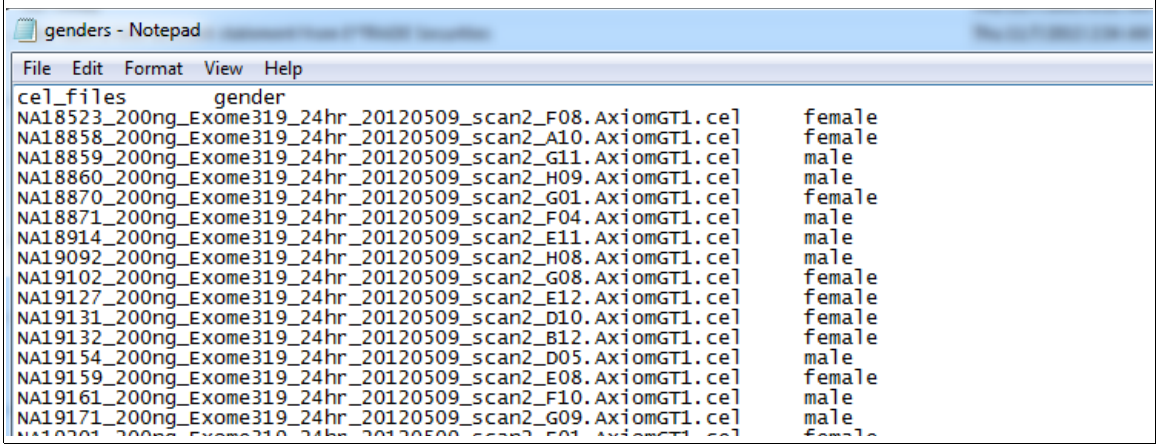
IMPORTANT: If there are no CHP files present in your Data Directory folder, you MUST include gender values for all samples. If you are using CHP files, skip this *Select Gender File* option.

To include gender values for all samples:

1. Click the ***Specify Gender information*** checkbox.
2. Click the **Gender File** Browse button.
3. Navigate to the folder that contains your Gender files.

NOTE: Your gender file must be a tab-delimited text file with 2 columns. Its first column header must be *cel_files.* The second column header must be *gender*, as shown in Figure K.2.

Figure K.2 Gender Text File contents example.



   The Gender column (far right) (Figure K.2) is not case-sensitive.
   - For *Female* type: **Female, female, F,** or **2**
   - For *Male* type: **Male, male, M,** or **1**
   - To specify an *unknown* gender type: **unknown** or **0**

   If no gender was specified or the gender was specified other than the required naming conventions stated above, the gender entry will be treated as *unknown*.
4. Click to select the *gender.txt* file you want to use, then click **OK**.
   The Gender File path is now populated.

Reference File (Optional)

The choice of samples to be used as a reference is critical for accurate CNV detection because the log2ratio at a marker is computed by dividing the intensity of the marker by the median intensity of that marker in the chosen reference set, in log space. The reference set, therefore, should represent the normal copy number state for each marker. One approach is to create the reference based on the individuals genotyped on the plate, provided that for each marker the vast majority of individuals on the plate are expected to have normal copy number states. Another approach is to create a separate reference based on individuals expected to have normal copy number states genotyped on different plates. If the latter

approach is chosen the number of samples used for the reference should be as large as possible, preferably at least 100. The analysis can be carried out with any number of samples but will be less accurate for smaller reference sets.

Do the following to specify reference data:

1. Click the ***Specify Reference Data*** checkbox.
2. Click the **Select Ref File** Browse button.
3. Navigate to the folder that contains your reference data file.
4. Click to select the file you want, then click **OK**.
   The Select Ref File path is now populated.

Running the Axiom CNV Summary Tool

1. After your Axiom CNV Summary Tool data paths are set, click **Run**.
   A green progress bar appears. Processing time varies depending on the amount of data you are processing, the number of SNPs on your array, and your system's hardware specifications.
   After the data has been successfully processed, a message appears.
2. Click **OK** to acknowledge the message.

Retrieving the Axiom CNV Summary Tool Data

The following files are produced and are stored in the **Data Directory** folder you assigned earlier:

- ***.cnv.txt** - Base name is the CEL file base name. (This file contains log2 ratio and BAF values for the sample associated with the CEL file.)
- **AxiomGT1.cnv.txt** - Contains log2 ratio and BAF values for all samples.
- **<annot>.probemappings.txt -** Required by BioDiscovery's Nexus software.
- **AxoimGT1.cnv.reference.txt** - Reference values for BAF and log2 ratio calculations.
- **AxiomGT1.cnv.params.txt** - Contains the parameters associated with the CNV analysis.

Ways to Use the Axiom CNV Summary Tool Data

Subsequent Analyses

Use the newly generated **AxoimGT1.cnv.reference.txt** for additional analysis.

1. Click the **Specify Reference Data** checkbox.
2. Click the Select Ref File Browse button, navigate to your Data Directory folder, then click to select the file: **AxoimGT1.cnv.reference.txt**
3. After your Axiom CNV Summary Tool data paths are set, click **Run**.
   A green progress bar appears. Allow time for your data to process.
   After the data has been successfully processed, a message appears.
4. Click **OK** to acknowledge the message.

Viewing Data in the Axiom CNV Viewer

View the newly generated ***.cnv.txt** for additional analysis in the included Axiom CNV Viewer.

Using the Axiom CNV Viewer

To start the Axiom CNV Viewer

1. Click **Tools -> Axiom CNV Viewer**.

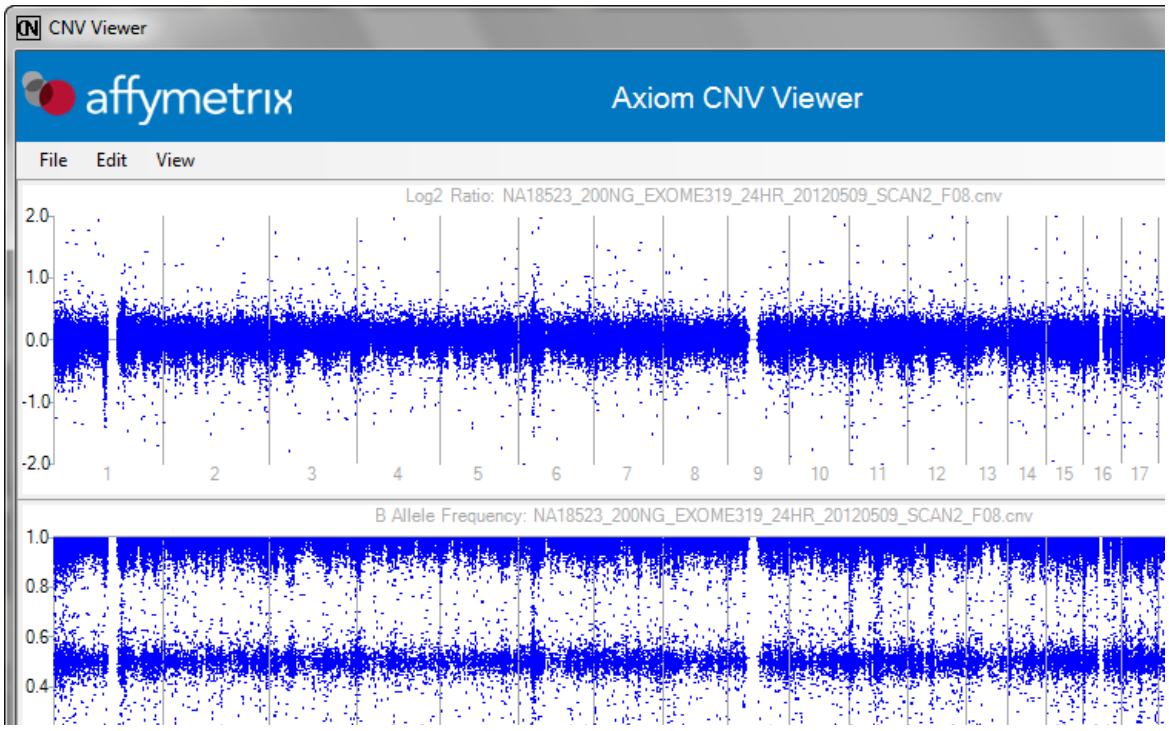
The following window appears: (Figure K.3)

Figure K.3 CNV Viewer - Startup window

2. Click **File** -> **Open**.
3. Navigate to your Data Directory folder, then select the ***.cnv.txt** file(s) you want to view.
4. Click **OK**.

The Viewer displays your data. (Figure K.4)

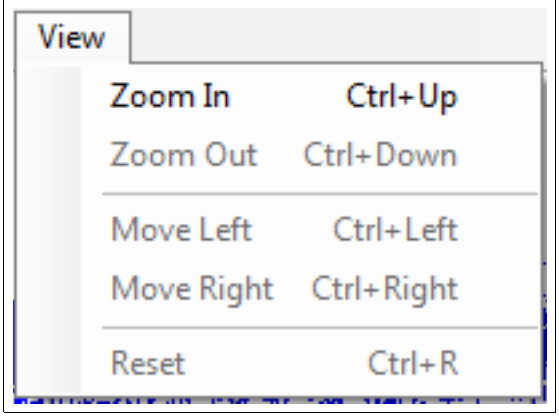
Figure K.4 CNV Summary Tool - Main window populated



To customize the display view:

5. Click **View**, then click to select one of the following viewing options or use the equivalent keyboard commands shown. (Figure K.5)
6. Repeat the viewing command as needed to reach the desired view.

Figure K.5 CNV Summary Tool - View drop-down menu



To reset your customized view back to the default whole genome view:

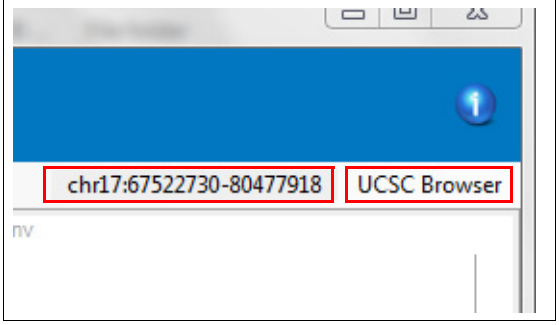
1. Click **Reset**.

To use the CNV Viewer to investigate your copy number changes:

Option #1

1. Use the *Zoom In* command or click, then drag your mouse cursor across a region of interest.
2. Once the Viewer has zoomed into a chromosome, a *UCSC Browser* button appears (upper right corner). (Figure K.6)

Figure K.6 UCSC Browser link and chr text filed.



3. Click on the **UCSC Browser** button.

   The UCSC website page appears (Figure K.7) and displays the current region based on the chromosome positions listed in the *chr* text box. (Figure K.6)

Option #2

1. Click inside the *chr* text box (Figure K.6), then manually enter your chromosome positions. You must use one of the following formats:

- chr17:67522730-80477918
- chr17:67,522,730-80,477,918

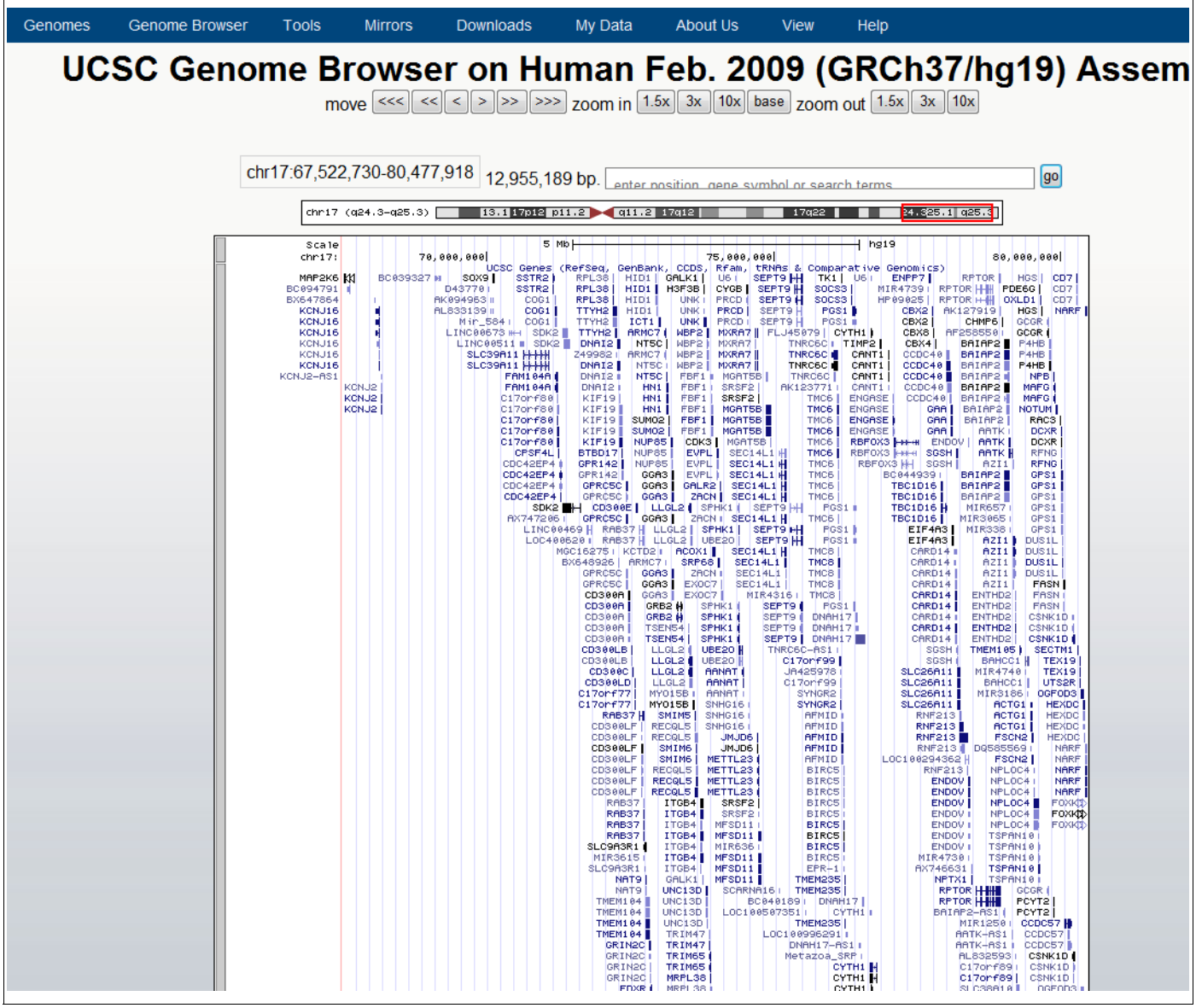
2. Press Enter.

NOTE: The region displayed in the CNV Viewer may be smaller than the chromosome positions you entered, because the CNV Viewer auto-adjusts your start and stop positions next to the nearest available start and stop markers.

3. (Optional) Click on the **UCSC Browser** button.

   The UCSC website page appears (Figure K.7) and displays the current region based on the chromosome positions listed in the *chr* text box. (Figure K.6)

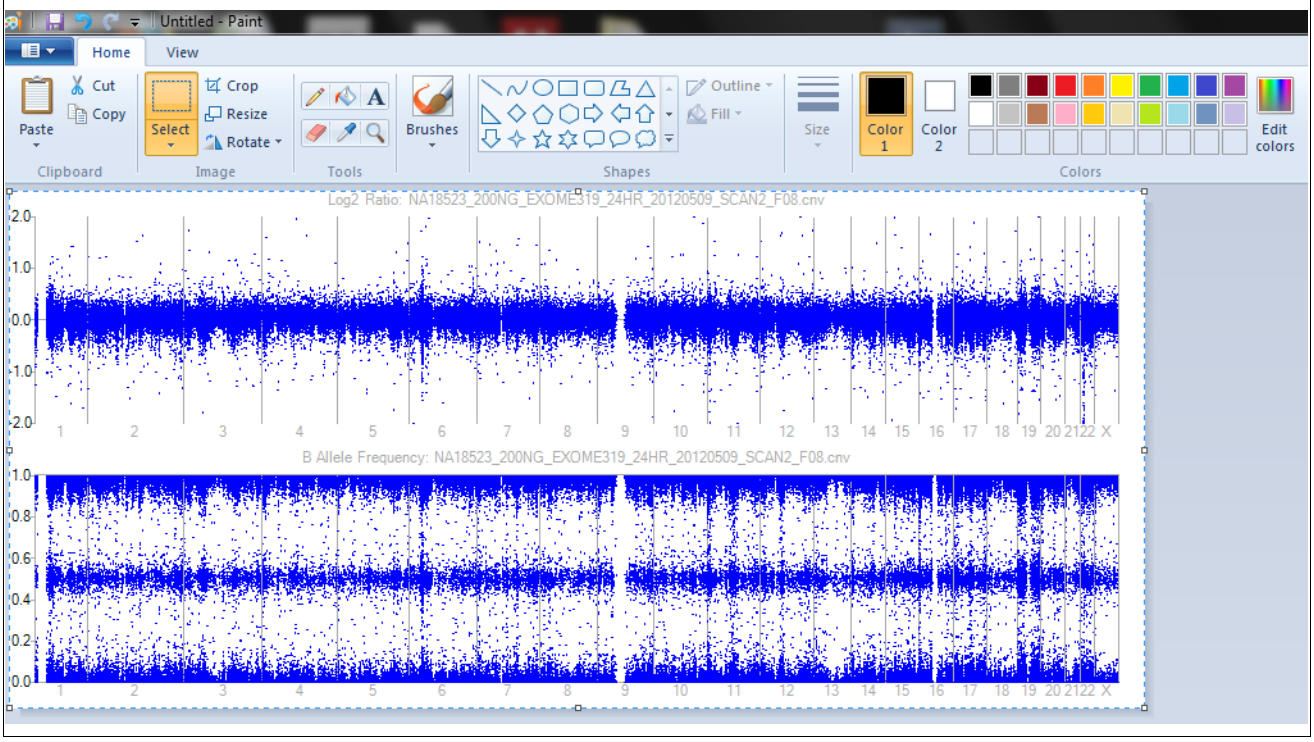
Figure K.7 UCSC website page



To copy the current view to your Clipboard:

1. Click **Edit -> Copy to clipboard**.
2. Use the paste command (**Ctrl-V**) to copy the current view into another software application, such as MS Paint. (Figure K.8)

Figure K.8 Paste to Clipboard - MS Paint Example



Further Copy Number Analysis Using BioDiscovery's Nexus Software

Use the newly generated **AxiomGT1.cnv.txt** and **<annot>.probemappings.txt** with BioDiscovery's Nexus software to perform copy number analysis.

Do the following to configure Axiom CNV Summary Tool output data to work with BioDiscovery's Nexus software:

1. Click the CNV Summary Tool's Configure Nexus button (bottom right).

NOTE: **If Nexus is not detected on your system, the *Configure Nexus* button is disabled.**

If multiple versions of Nexus are detected, a drop-down menu appears. Use this menu to select the appropriate version of Nexus. This menu does not appear if only one version of Nexus is detected.

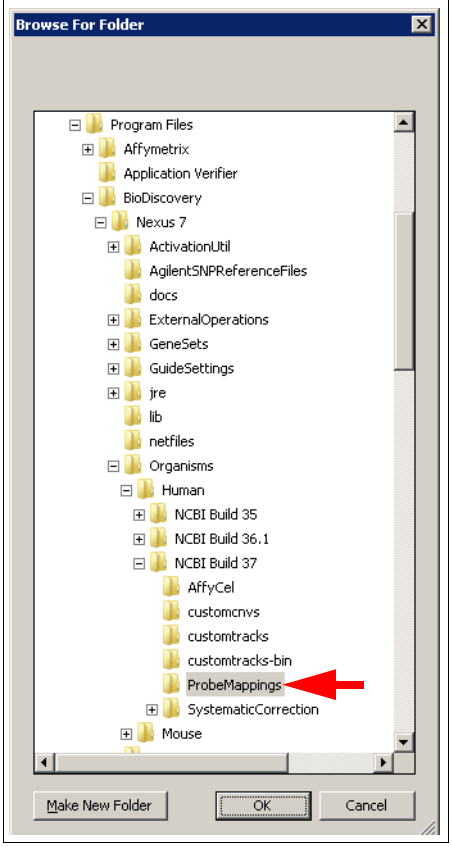
A file window appears.

2. Click to select the *probe mapping .txt* file. This file resides in your master Data Directory folder you setup earlier. See Step 2 on page 364.
3. Click **Open**.

An Explorer window appears. (Figure 9)

4. Navigate Nexus's *ProbeMappings* directory/folder, then click **OK**.

Figure 9 Nexus ProbeMapping folder



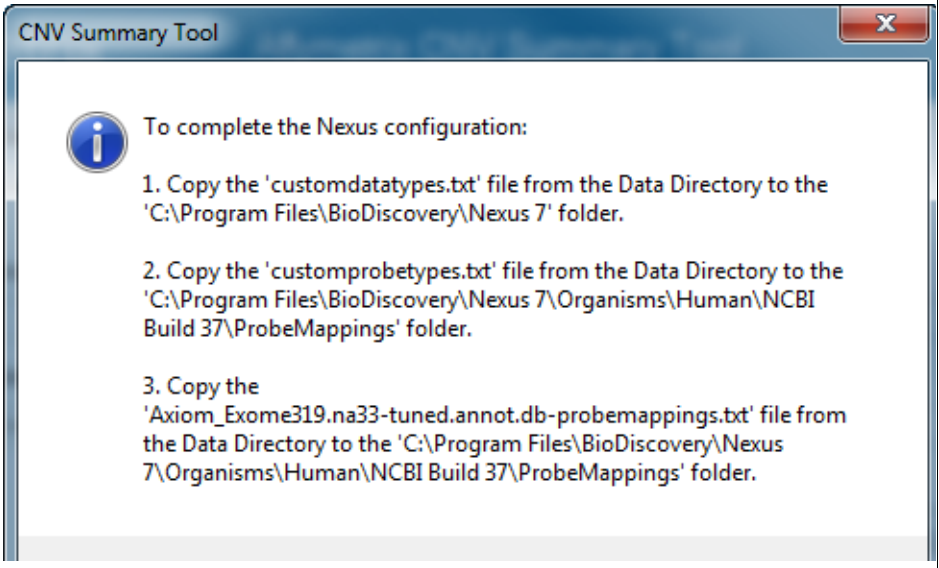
The message *Configuration Complete* appears.

5. Click **OK**.

IMPORTANT: You are responsible for knowing the location of Nexus's *ProbeMappings* folder. If you are unsure of its location, contact BioDiscovery. In most cases, the Nexus Probe Mapping folder resides here: *C:\Program Files\BioDiscovery\NexusX*

If you do not have access (Administrator Privileges) to some of your computer's folders, a message with file copying instructions appears. **(Figure K.10)** Follow the 3 steps shown to configure the Nexus software manually.

Figure K.10 To complete the installation message



6. Use the Nexus software as you normally would.
   If you want to perform a GC Correction in Nexus, see *Performing GC Correction in Nexus* on page 373

Performing GC Correction in Nexus

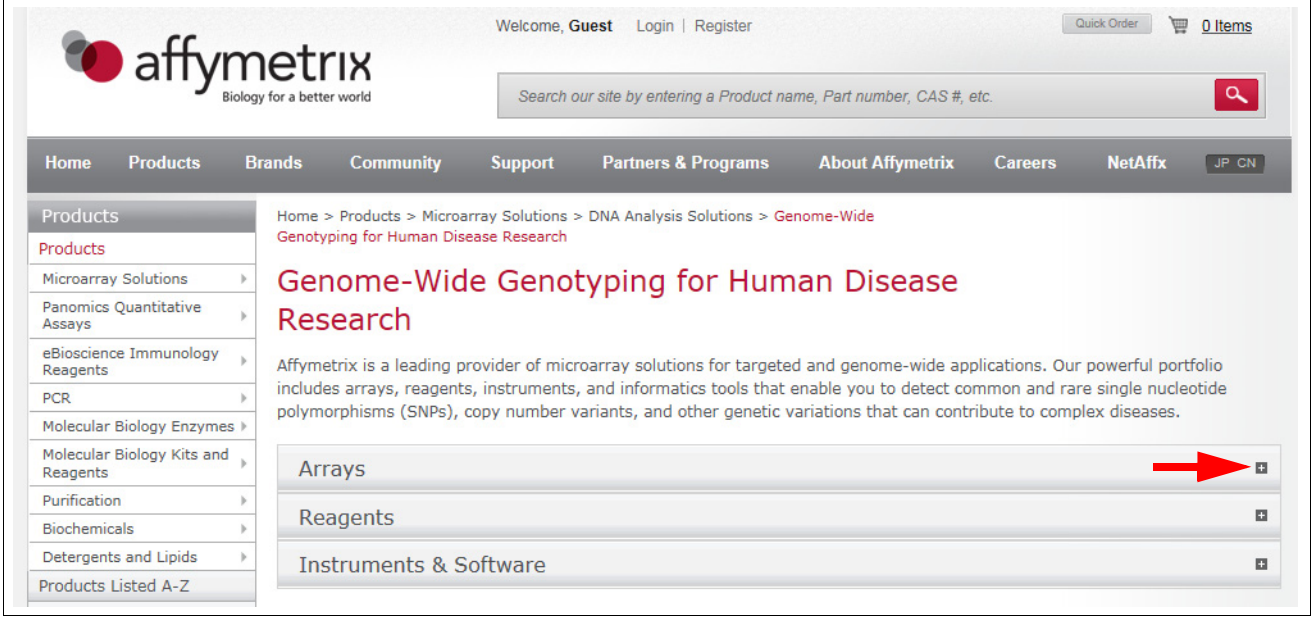
IMPORTANT: You must first download an appropriate BED file from affymetrix.com.

To download a BED file from affymetrix.com:

1. Go to *www.affymetrix.com*.
2. Login as you normally would or click **Register**, then follow the on-screen instructions.
3. Click **Products** -> **Products** (top left).
   The Products page appears.
4. Click **Microarray Solutions** (left pane).
   The Microarray Solutions pane appears.
5. Under the **DNA Analysis Solutions** header, click to choose the option you want. Example: *Genome-Wide Genotyping for Human Disease Research*.
   For the *Genome-Wide Genotyping for Human Disease Research* example, 3 options appear.

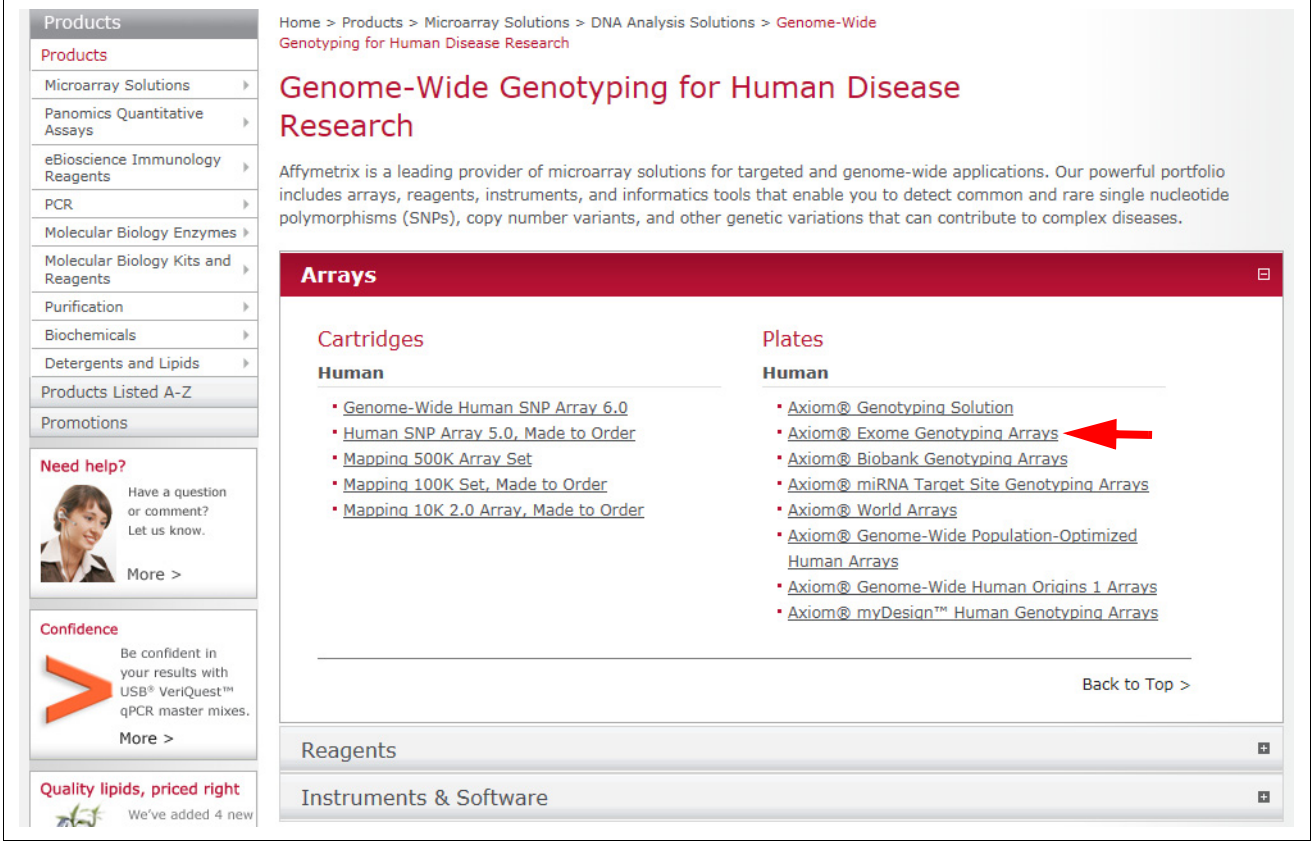
6. Click the *Arrays* adjacent **[+]** button. (Figure K.11)

Figure K.11 Arrays menu



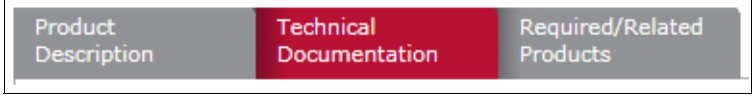
For the *Genome-Wide Genotyping for Human Disease Research* example, the following page appears. (Figure K.12)

Figure K.12 Arrays menu



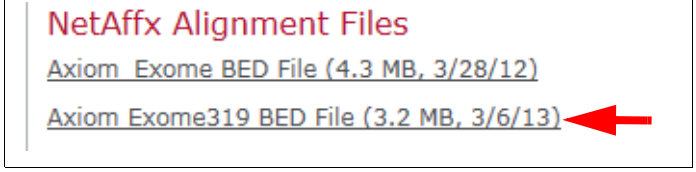
7. For this example, click the **Axiom® Exome Genotyping Arrays**. (Figure K.12)
   The *Axiom® Exome Genotyping Arrays* page appears.
8. Click on the **Technical Documentation** tab. (Figure K.13)\

Figure K.13 Technical Documentation tab.



9. Scroll down and locate *NetAffx Alignment Files*, then (for this example) click on **Axiom Exome319 BED File**. (Figure K.14)

Figure K.14 Technical Documentation tab.



A Windows Explorer window appears.

NOTE: BED files for *Axiom myDesign* arrays are provided directly to you from Affymetrix (for each custom array designed).

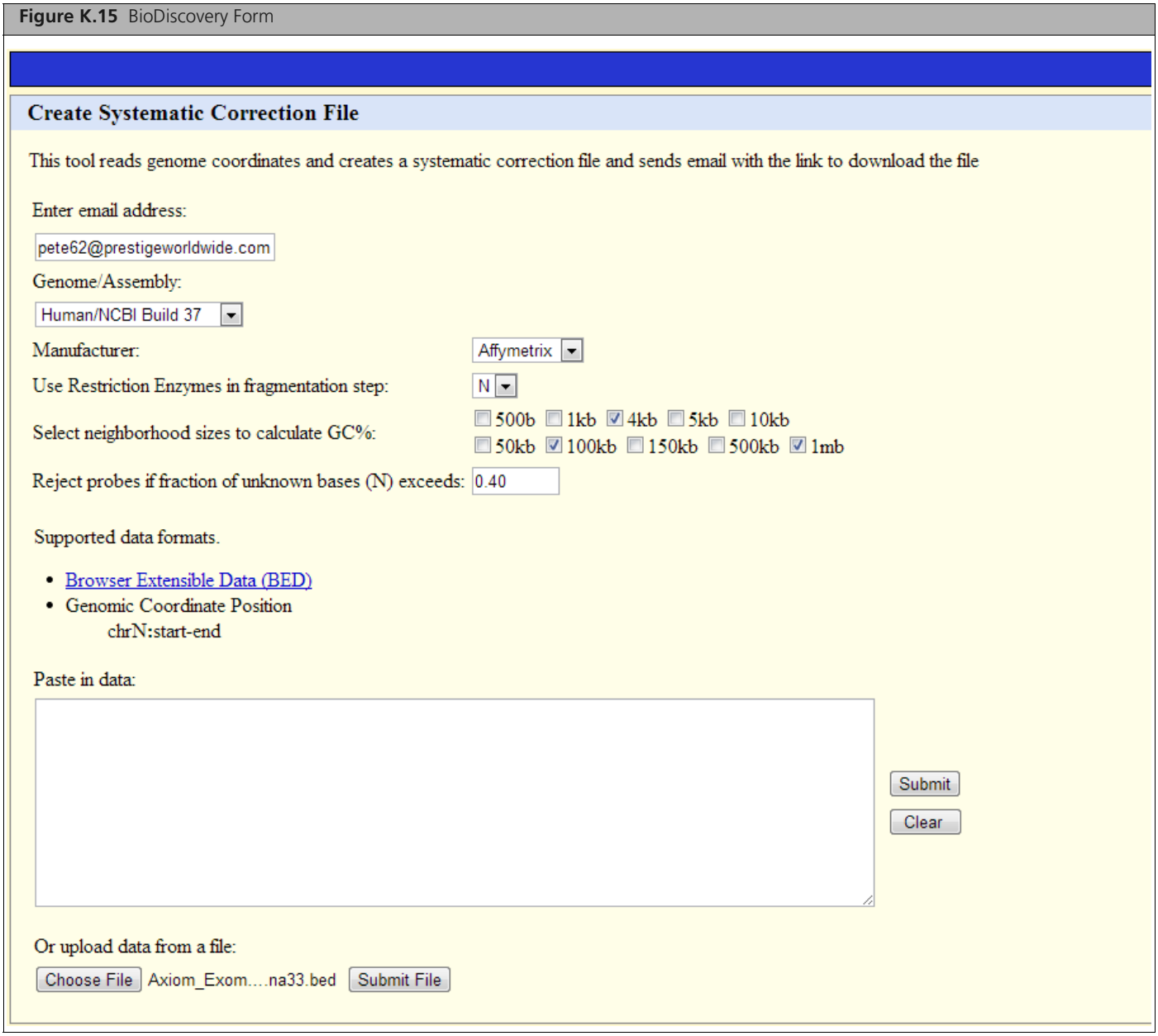
10. Save the zip file to a convenient location.

Do the following to submit your BED file to BioDiscovery for GC Correction:

1. Extract the downloaded BED.zip file, then contact BioDiscovery and tell them you need a GC Correction file created from a BED file.
   BioDiscovery will respond with an email containing a hyperlink.
2. Click on the hyperlink provided by BioDiscovery.
   The following form appears. (Figure K.7)

3. Complete the form, click **Browse** to upload your unzipped BED file, then click **Submit File**.

Figure K.15 BioDiscovery Form

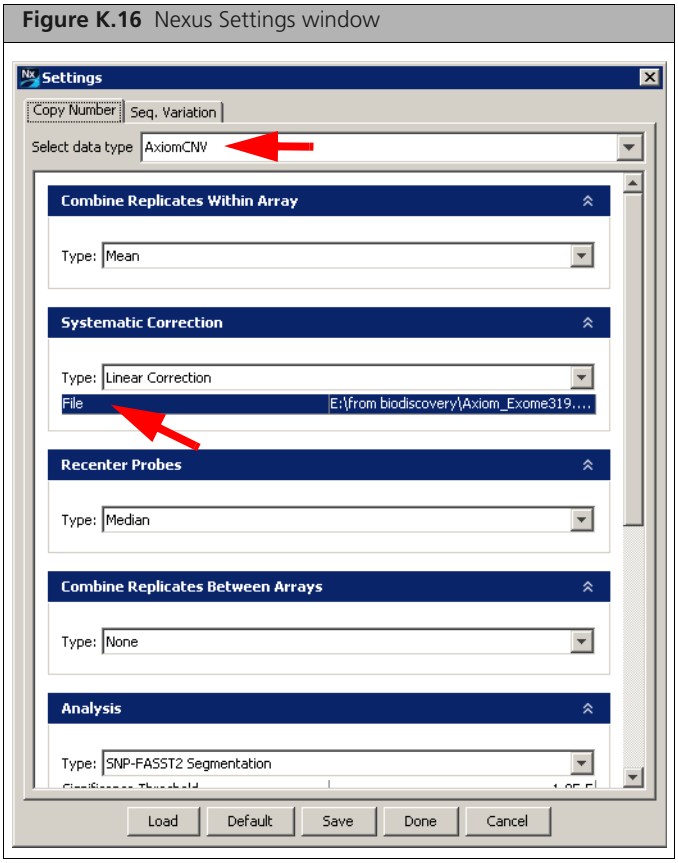


NOTE: BioDiscovery will email you a second hyperlink to download the GC Correction file for use with their Nexus software.

4. Click on the hyperlink to download/save the GC Corrected BED file. Make sure you save the file to a convenient location.
5. Open the *BioDiscovery Nexus* application as you normally would.
6. Click **Settings**.

The following window appears. (Figure K.16)

Figure K.16 Nexus Settings window



7. From the *Select data type* drop-down menu, click **AxiomCNV**. (Figure K.16)
8. From the *Systematic Correction* drop-down menu, select your GC Correction Type.
9. Click the File banner (Figure K.16), then select your GC Correction file.
10. Use the other applicable drop-down menu selections to complete the Settings form, then click **Done**.
11. Use the Nexus software as you normally would.

Troubleshooting

The following information is provided to help you troubleshoot GTC:

Troubleshooting Tips

Table L.1 Troubleshooting Tips

Issue	Resolution
Data file(s) (ARR, XML, CEL, GQC, or CHP) cannot be imported and/or causes the software to crash.	Confirm that the data files were generated by Affymetrix software or GeneChip compatible software partners can be imported into Genotyping Console and have not been tampered with or edited. Any data files which are edited outside of these software packages may cause import to fail or Genotyping Console software to crash.
I tried to import my CEL files and selected the auto-QC option. An error indicated that I was missing a library file and the QC step was aborted but no CEL files were added to the Workspace.	If an action is selected such as auto-QC and the required library files are missing, all current actions are aborted so no data files including the CEL files are added to the Workspace. To resolve this issue, download the required library files from the File menu and repeat the data import.
My analysis is taking a long time.	Confirm that the CEL files are located on the local machine and NOT on a network. Affymetrix recommends that you perform genotyping and QC analysis with all files stored locally. Close other applications to free up memory and CPU resources.
I copied data to the Clipboard but when I pasted it into a new document/file, not all of the text was copied.	The copy to Clipboard is a Windows operating system feature and can only hold a certain amount of data. If you copy a large amount of rows/columns of data, it may not all be able to handled by Windows. To resolve this issue, copy smaller sections of data or export to a text file.
I selected files to be added to the workspace but not all files were added.	Windows has a fixed buffer which limits how many files can be returned to the application. The control lets a user pick any number of files, but due to its buffer size it may return fewer files. The maximum number of files varies. As an example, when trying to add 800 ARR and CEL files to the Data Set at one time, although all files could be selected only a subset are actually added to the Workspace. The work-around is to either work with Windows folders containing smaller sets of data, or to perform the Add Data operation multiple times, each time selecting a different set of files in the Windows folder.
Sorting and/or scrolling the SNP Summary table is slow and unresponsive.	Since the SNP summary table holds all of the SNP results for all CHP files in the batch, it can become very large. Not all of the data is loaded into the memory. Sorting and scrolling this file may take time. If you select multiple actions, the software may become unstable. To resolve this issue, export the SNP summary table to text or use the Filter SNPs option to select a subset of the data for easier use.
Genotyping analysis failed.	View the log window; it may contain information relating to the issue. Confirm that the algorithm parameter values are valid. To resolve this issue, make sure you are using values within these bounds: Score Threshold: 0 – 1

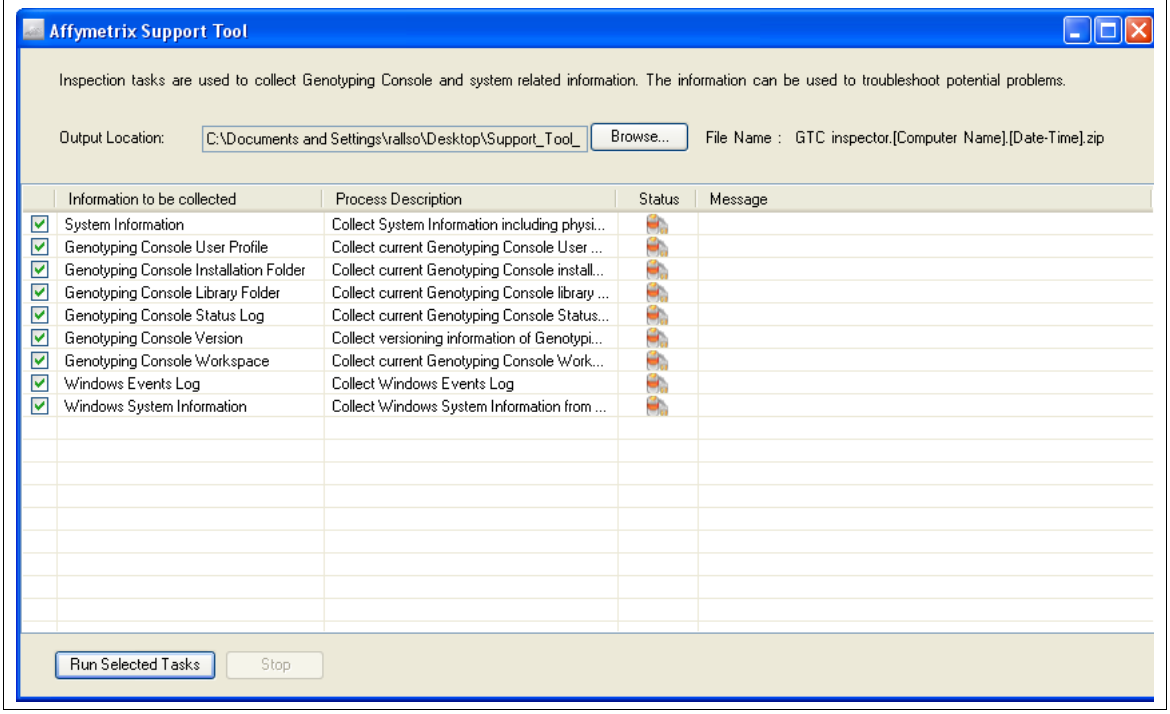
Table L.1 Troubleshooting Tips

Issue	Resolution
Genotyping Console could not perform the QC and/or the Genotyping analysis.	Confirm that the library files are present. Refer to the *Library and Annotation file section* of the manual for more information.
I got an error when I tried to add additional data to my Data Set.	Data Sets can consist of only one array type. Confirm that you are adding data which is the same probe array type (e.g. Genome-Wide SNP 5.0) to the existing Data Set.
I tried to perform QC and/or Genotyping analysis and Genotyping Console could not find the data files.	Confirm that the data files have not been moved or deleted by verifying the file locations. Go to Workspace/Verify File Locations.

Using the Troubleshooter Tool

The Affymetrix Support Tool (Figure L.1) can be used to collect information on the operation of GTC that may be useful to Affymetrix Support in troubleshooting problem.

Figure L.1 Affymetrix Support Tool

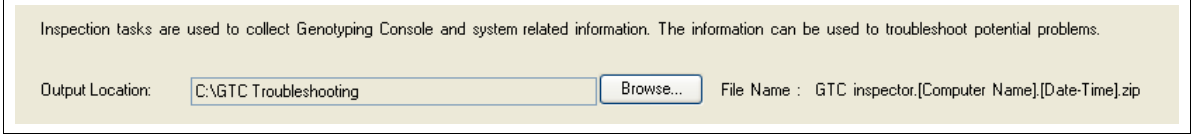


It creates a set of XML files in a zip package that can be sent to Affymetrix Support.

To collect troubleshooting information using the tool:

1. From the Tools menu, select **Troubleshooter**.
   The Affymetrix Support Tool dialog box opens.
2. Enter the path and name for the output location (Figure L.2); or

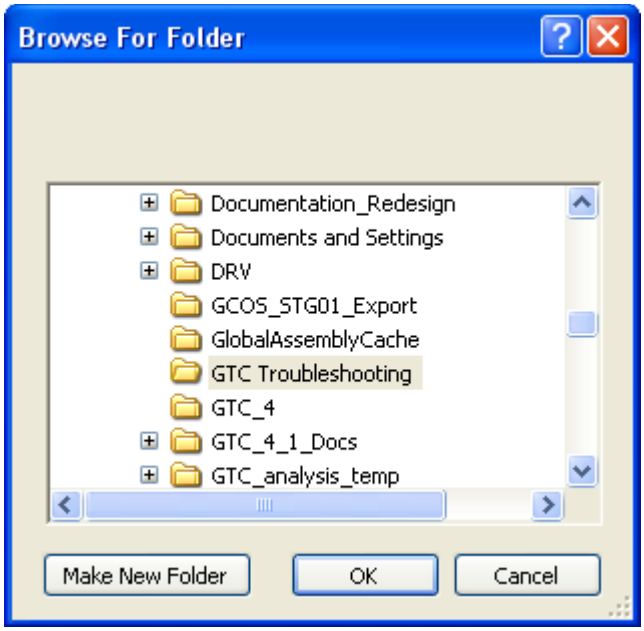
Figure L.2 Output Location



A. Click the **Browse** Button

The Browse for Folder dialog box opens (Figure L.3).

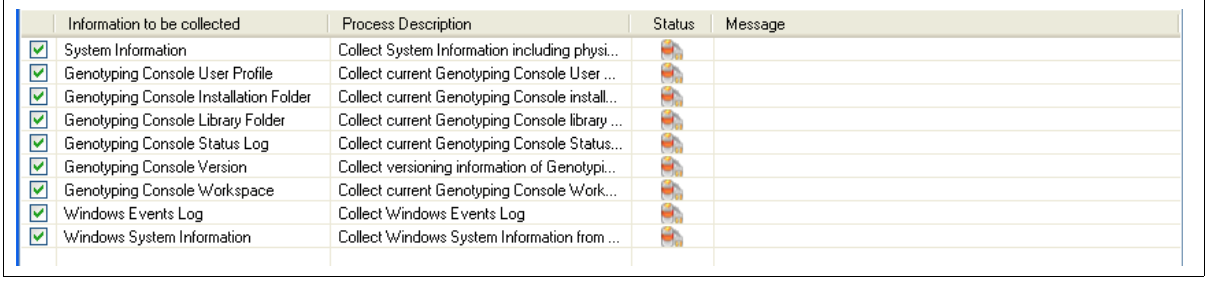
Figure L.3 Browse For Folder dialog box



B. Navigate to the folder location (making a new folder if necessary) and click OK in the Browse for Folder dialog box.

3. Select the Reports you wish to generate (Figure L.4).

Figure L.4 Options for collecting information



You can choose from the following options:

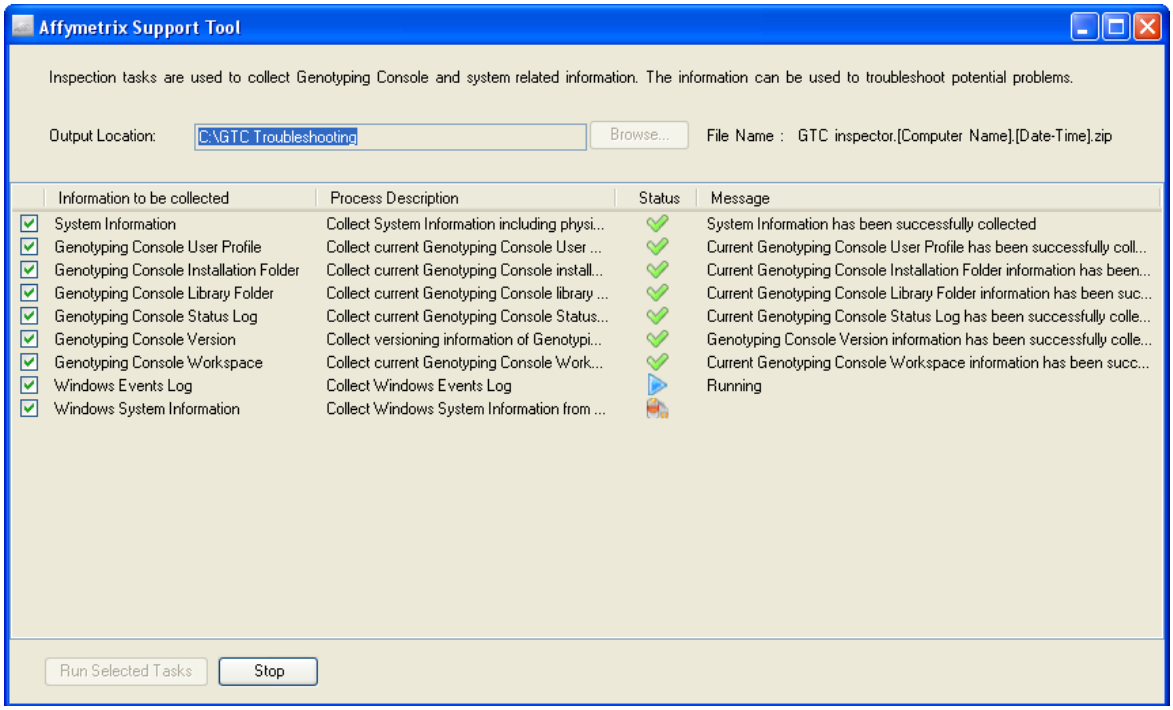
- Collect GTC System Information – The exported file contains information about Physical RAM, CPU, 32/64-bit OS, Total and Free Space of C Drive and Windows OS Version and Service Pack Version.
- Collect Library File Information – The exported file contains local library file path (including date and file size), lists of probe array types with complete library files and annotation files. Library and annotation file versions as specified in the file name.
- Collect Workspace Information – If there is no workspace opened, the exported file will contain a simple report saying that there is no workspace opened. If that's not the case, the exported file will be an xml file of the currently opened workspace.
- Collect Status Log Information – The exported file contains log information in the status log window in GTC.

- Collect GTC Version Information – The exported file contains GTC Version and Build number as reported in the about box.
- Collect Current User Profile – The exported file contains the user profile currently logged in.
- Collect Installation Information – The exported file contains information about file name, path, date and size of installation files under GTC folder. The extensions of the files can be set freely, for example, .exe and .zip.
- Collect Windows System Information – the exported file contains comprehensive windows system information generated via msinfo32.exe.
- Collect Windows Event Log – The exported file contains windows event log

4. Click **Run Selected Tasks**.

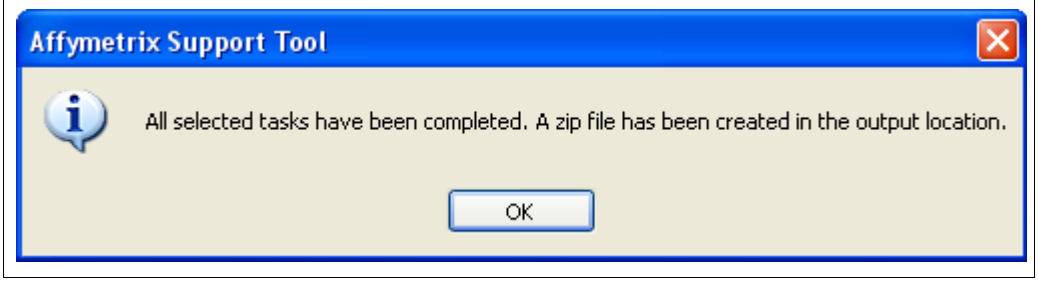
   The dialog box displays the progress of the various tasks (Figure L.5).

Figure L.5 Affymetrix Support Tool running



When the information has been collected, a notice appears (Figure L.6).

Figure L.6 Notice of completion



The selected information is collected in a zip file at the output location (Figure L.7).

Figure L.7 Zip file with information