Affymetrix GeneChip® Chromosome Copy Number Analysis Tool (CNAT) – Version 4.0

User Guide

March 2006

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

Trademarks

Limited License Notice

Patents

Copyright

CONTENTS

Chapter 1

WELCOME

Welcome to the Affymetrix GeneChip® Chromosome Copy Number Analysis Tool (CNAT) v.4.0 User Guide.

This chapter decribes the following:

CNAT 4.0 Application

CNAT is an application tool in the Affymetrix GeneChip® Genotyping Analysis (GTYPE) software that allows you to perform copy number (CN) and loss of heterozygosity (LOH) analysis on data from Affymetrix Mapping arrays (Mapping 500K, 100K, and 10K). CNAT 4.0 software implements new copy number and LOH algorithms, which provides the following features:

- Four analysis workflows:
  - CN for paired tumor/normal samples
  - CN for un-paired samples
  - LOH for paired tumor/normal samples
  - LOH for un-paired samples
- Perfect Match (PM) only copy number estimation
- Restricted analysis of SNPs based on PCR fragment size filter
- Probe-level quantile normalization or median scaling
- Probe summarization per SNP and per allele
- PCR fragment size and GC content normalization
- Dynamic generation of global reference based on the experiment for un-paired CN workflow

- Allele-specific CN estimation on paired tumor/normal samples
- Virtual array – combining of NSP and STY (or Xba and Hind) data from the same sample
- Gaussian smoothing
- Hidden Markov Model (HMM) for CN and LOH
- Outlier removal
- Leverages BRLMM.CHP files for LOH detection (BRLMM = Bayesian Robust Linear Model with Mahalanobis distance classifier algorithm)

Introduction

Identification and detection of DNA copy number changes is a major focus in current disease research. High resolution, whole genome, SNP arrays are changing the paradigm of detecting chromosomal imbalances by enabling researchers to analyze copy number alterations, LOH, and genotypes in a single experiment. In comparison to existing technologies, GeneChip® Mapping Arrays provide higher resolution, increased throughput, and higher reproducibility.

The Affymetrix GeneChip® Chromosome Copy Number Analysis Tool 4.0 (CNAT 4.0) implements novel algorithms to identify genome-wide chromosomal gains and losses using high density, oligonucleotide, array-based SNP genotyping methods and the Whole Genome Sampling Assay (WGSA). The CNAT 4.0 application is integrated into the Affymetrix GeneChip® Genotyping Analysis Software (GTYPE) to allow genotype calling and copy number calculations to be performed within one software application.

CNAT 4.0 contains several enhancements focused on improving the copy number analysis workflow including dynamic filtering and improved data visualization. The CNAT 4.0 application also enables users to visualize data in genomic context by automatically launching CNAT graphics into the Affymetrix® Integrated Genomic Browser (IGB) and saving graphs readable by the UCSC Genome Browser.

Intended User

This software and user guide is intended for users who have a working knowledge of the Affymetrix GCOS and GTYPE software packages. Please contact your local FAS for support and training on GCOS and GTYPE prior to using CNAT 4.0.

Overview of CNAT 4.0 Batch Analysis Workflow

Step 1 — Selection of Sample Type

Paired or Un-paired Sample Analysis

- Paired sample analysis – The sample and reference DNA are obtained from the same individual.
- Un-paired sample analysis – The sample is compared to a set of references.

Step 2 — Selection of Analysis Type

Copy Number (CN), Loss of Heterozygosity (LOH), or Both

- CN analysis – Quantile normalization or median scaling is performed at the PM probe level followed by summarization of the signal intensity for each allele of each SNP. This is the starting point for estimating copy number.
  Additionally, in the case of the CN paired workflow, genotype calls are used for enabling estimation of allele-specific CN.
- LOH analysis – Genotype calls is the starting point for determining LOH.

Step 3 — Selection of Files for Analysis

The user selects the appropriate CHP files by either dragging and dropping from the data tree into the appropriate columns (sample or reference) or uploads a file containing the CHP file names.

Step 4 — Advanced Analysis Options (Optional)

Advanced options allow you to set advanced CN and LOH algorithm parameters.

Step 5 — Output File Naming (Optional)

The output files are named according to the CHP file name and the analysis type selected in Step 2.

Step 6 — Analyze

Analysis is initiated.

About this Manual

This manual presents information about the CNAT 4.0 application tool in the following chapters and appendices:

- Chapter 2 *Installation* – Describes how to install the CNAT application.
- Chapter 3 *CNAT Batch Analysis Tool* – Describes how to access and use the CNAT Batch Analysis Tool for generating copy number and LOH estimates.
- Chapter 4 *CNAT Viewer* – Describes how to use CNAT Viewer to view the copy number and LOH data files.
- Appendix A *CNAT 4.0 Algorithm* – Describes the algorithm details.
- Appendix B *CNAT Version History* – Describes the version history for the CNAT application.
- Appendix C *How To Create Virtual Sets* – Describes how to set up and use Virtual Sets.

Conventions Used in This Guide

This manual provides a detailed outline for all tasks associated with the Affymetrix GeneChip® Chromosome Copy Number Analysis Tool. Various conventions are used throughout the manual to help illustrate the procedures described. Explanations of these conventions are provided below.

Steps

Instructions for procedures are written in a numbered 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.

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

Font Styles

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

1. Click the **Find** toolbar button  ; or

   Select **Edit** → **Find** from the menu bar.

   The Find dialog box opens.

2. Enter *Hind_50K* 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.

NOTE: The screen captures depicted in this manual may not exactly match the windows displayed on your screen.

Additional Comments

Throughout the manual, text and procedures are occasionally accompanied by special notes. These additional comments and their meanings are described below.

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

NOTE: The Note format presents supplemental 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 a 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

Tel: 1-888-362-2447 (1-888-DNA-CHIP)
Fax: 1-408-731-5441

sales@affymetrix.com
support@affymetrix.com

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
Fax: +44 (0) 1628 552585

saleseurope@affymetrix.com
supporteurope@affymetrix.com

Affymetrix Japan K.K.
Mita NN Bldg. 16F
4-1-23 Shiba Minato-ku,
Tokyo 108-0014 Japan

Tel. 03-5730-8200
Fax: 03-5730-8201

salesjapan@affymetrix.com
supportjapan@affymetrix.com

www.affymetrix.com

Chapter 2

INSTALLATION

This chapter describes how to install the Affymetrix GeneChip® Chromosome Copy Number Analysis Tool (CNAT) v.4.0 software and details the recommended and mandatory hardware requirements.

Installing CNAT 4.0

To install the Affymetrix GeneChip® Chromosome Copy Number Analysis Tool (CNAT), you must first install Affymetrix GeneChip® Operating Software (GCOS) and Affymetrix GeneChip® Genotyping Analysis Software (GTYPE) in that order. GTYPE 4.1 contains a new genotyping algorithm called BRLMM. It is highly recommended that you upgrade to GTYPE 4.1 in order to leverage the new BRLMM.CHP files for LOH analysis. See *Recommended Hardware Requirements* before installing the CNAT 4.0 software.

NOTE: The GCOS install must be version 1.4 or higher if you are working locally, or version 1.3 or higher, if you are working from a server.

1. Follow GCOS and GTYPE download instructions on the Affymetrix web site. Go to www.affymetrix.com, navigate to **Products** → **Software**, and select the appropriate software for download.

   NOTE: You must have administrative rights on the computer workstation running Microsoft Windows to install any software.

2. Follow CNAT download instructions on the Affymetrix web site. Go to www.affymetrix.com, navigate to **Products** → **Software** → **Chromosome Copy Analysis Tool (CNAT)**, and download the software.

IMPORTANT: **CNAT 4.0 will not overwrite CNAT 3.0. If you want to remove 3.0, use the Microsoft Add/Remove Program feature. It is acceptable to have both 4.0 and 3.0 on the same system. However, if CNAT version 1.0 or 2.0 are installed on the workstation, they MUST be uninstalled BEFORE installing CNAT version 4.0.**

NOTE: **Installation is required only on the local machine, whether the instrument is run in local or server mode.**

3. Double-click the install file.
4. Follow the install instructions on the screen.

Recommended Hardware Requirements

The following recommended hardware requirements for CNAT 4.0 are the same hardware requirements needed for GTYPE:

- 2 GHz Pentium Processor
- 2 GByte RAM (**required**)
- 100 GByte Hard Drive

Chapter 3

CNAT Batch Analysis Tool

This chapter decribes the following:

CNAT 4.0 Batch Analysis

The CNAT 4.0 Batch Analysis Tool allows the user to batch process .chp files from single array types (10K, 50K, or 250K) and multiple array types (100K for 50K Hind and 50K Xba arrays, or 500K for 250K Nsp and 250K Sty arrays) using four workflows:

- Copy number (CN) for paired tumor/normal samples
- Copy number (CN) for un-paired samples
- LOH for paired tumor/normal samples
- LOH for un-paired samples

The results are stored in copy number and LOH data files (*.CN.cnt or *.LOH.cnt).

Opening the CNAT 4.0 Batch Analysis Tool

1. Click the Windows **Start** button. 

2. Select **Programs** → **Affymetrix** → **GTYPE**.
   The Affymetrix GeneChip® Genotyping Analysis Software (GTYPE) opens.

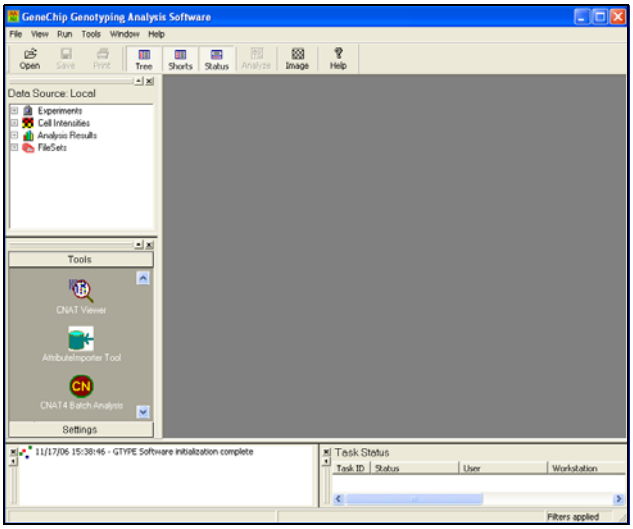

Figure 3.1 Affymetrix GeneChip® Genotyping Analysis Software main page

NOTE: Affymetrix recommends that you use GTYPE 4.1 to leverage BRLMM CHP files for your LOH analyses.

3. In the Tools window below the data tree, scroll until you see the following two icons:
   - CNAT 4 Batch Analysis
   - CNAT Viewer
4. Click the **CNAT 4 Batch Analysis** icon in the Shortcut bar, or select **Run → CNAT 4 Batch Analysis** from the menu bar.

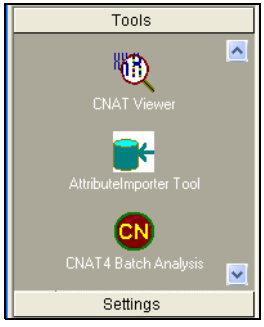

Figure 3.2 CNAT 4 Batch Analysis icon

The **CNAT 4 Batch Analysis** window opens.

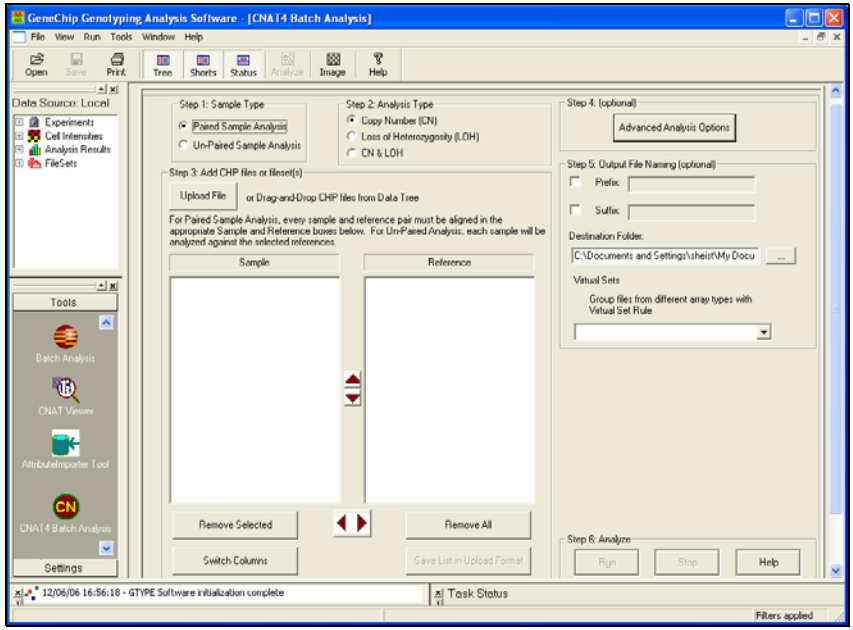

Figure 3.3 CNAT 4.0 Batch Analysis Window

Generating CN and LOH Data

Copy number and LOH data is generated using paired or un-paired analysis. Prior to running the analysis, use the filters page to select information about the experiment.

Filtering

Before you generate CN and LOH data, use the Filters page to select projects, sample type, assay type, probe type, experiment, and more to generate data.

To filter the data for CN and LOH analysis:

1. Select **Tools** → **Filters** or right-click a .chp file and select **Filters**.
   The Filters window opens.

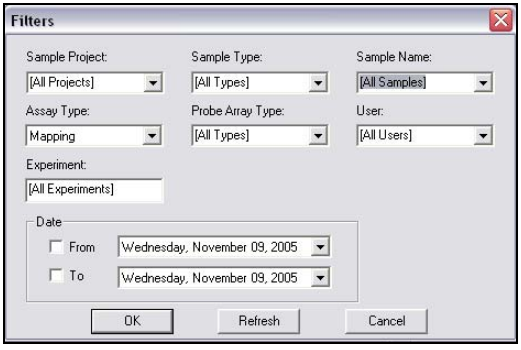

Figure 3.4 Filters window

2. Select **Mapping** from the **Assay Type** drop-down list.
3. Make selections from the other drop-down lists as needed.
4. Click **OK**.

Paired CN and LOH Analysis

Paired copy number analysis refers to tumor samples with a paired normal from the same individual. If you do not have paired normal controls, skip to the *Un-Paired CN and LOH Analysis* section. In the CNAT 4.0 Batch Analysis window (Figure 3.3), follow the steps outlined below to complete a paired sample analysis:

Step 1: Sample Type

- In the **Step 1: Sample Type** group box, select the **Paired Sample Analysis** option.

  The Paired Sample Analysis option refers to samples with a paired normal from the same individual.

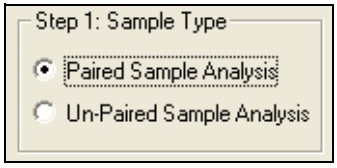

Figure 3.5 Sample Type – Paired Sample Analysis

Step 2: Analysis Type

- In the **Step 2: Analysis Type** group box, select Copy Number (CN), Loss of Heterozygosity (LOH), or both (Figure 3.6).

  If only the CN radio button is selected or only the LOH radio button is selected, the software performs only copy number analysis or only LOH analysis on the selected samples. If you select the CN and LOH radio button, both CN and LOH analyses are performed on the selected samples.

> **NOTE: Allele-specific results are obtained only when a paired analysis is run and the Generate Allele-Specific Copy Number check box in the Advanced Analysis Options window is selected.**

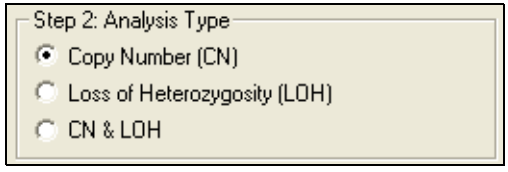

Figure 3.6 Analysis Type group box

Step 3: Add CHP files or fileset(s)

- In the **Step 3: Add CHP files or fileset(s)** group box (Figure 3.7):
  - Drag and drop sample or tumor CHP files from the data tree to the **Sample** list box.
  - Drag and drop reference or normal CHP files from the data tree to the **Reference** list box.

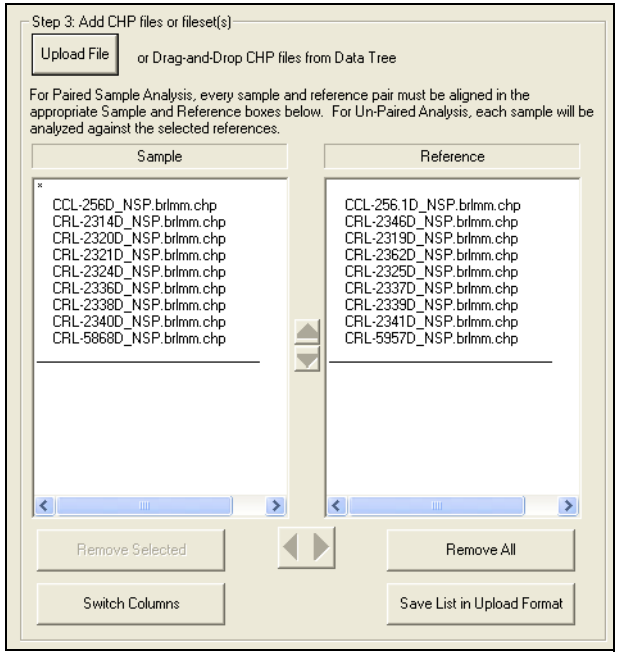

Figure 3.7 Add CHP files or fileset(s) group box

  - Alternatively, you can upload a file list containing the CHP file names of the sample and reference files. The file consists of two columns where the first column is the sample/tumor CHP file name and the second column is the paired reference/normal CHP file name (Figure 3.8).

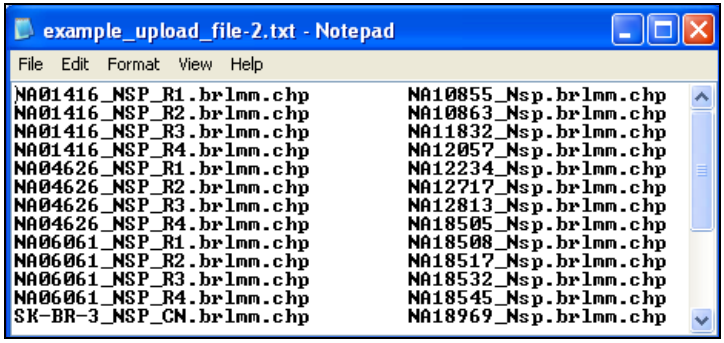

Figure 3.8 Example Upload File format

NOTE: Each sample and the corresponding paired normal reference file must be aligned horizontally in the list box or in the upload file. This alignment indicates which files to pair in the analysis.

Step 4: (optional) Advanced Analysis Options

- Click the **Advanced Analysis Options** button in the **Step 4** box (Figure 3.9).

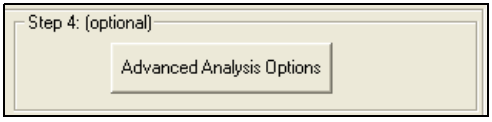

Figure 3.9 Advanced Analysis Options

The **Advanced Options** page opens (Figure 3.10).

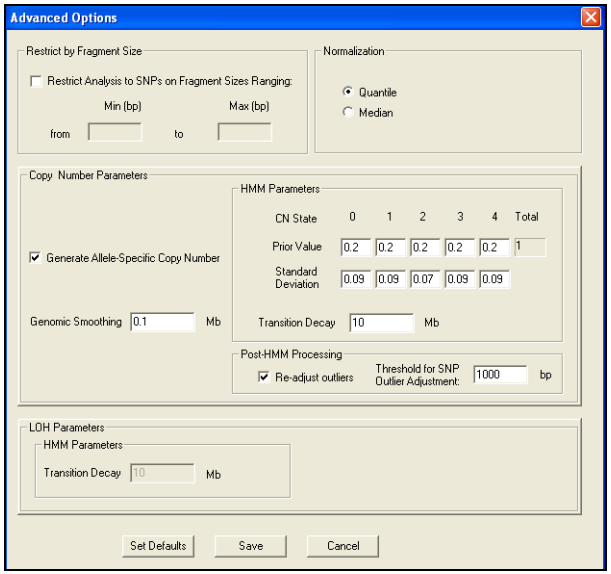

Figure 3.10 Advanced Options page

The Advanced Options page is divided into four sections:

- *Restrict by Fragment Size*
- *Normalization*
- *Copy Number Parameters*
- *LOH Parameters*

NOTE: If only CN or only LOH was selected in Step 2, then only the corresponding parameters are editable.

Restrict by Fragment Size

This option allows the analysis to be performed on only a subset of SNPs based on the fragment size where the SNPs reside. The default is unchecked and all SNPs are included in the analysis.

To enable this option:

1. Check the box next to **Restrict Analysis to SNPs on Fragment Sizes Ranging** (Figure 3.11).
2. Enter the size of fragments that you want to be included in the analysis.

   For example, if a sample has a known amount of degradation (no fragments larger than 600 bp), you can enter the following:

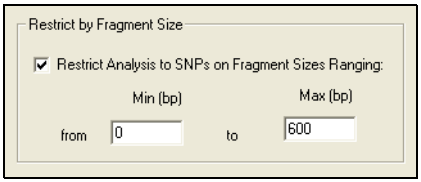

Figure 3.11 Restrict by Fragment Size group box

NOTE: In the example above, only SNPs that reside on fragment sizes less than or equal to 600 bp will be included in the analysis.

Normalization

This option allows specification of the probe-level normalization. Select one of the following two options in the Normalization group box (Figure 3.12).

Quantile

Quantile normalization performs a sketch normalization, based on 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 MM probes are included to compute the median intensity of a CEL file.

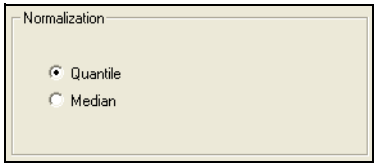

Figure 3.12 Normalization group box

Copy Number Parameters

Allele-Specific Copy Number

For paired analyses, 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 box next to Generate Allele-Specific Copy Number (Figure 3.13). See the Copy Number table (Table 3.1) for recommended CN parameter settings.

Genomic Smoothing

The user can 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) which results in a window size of 400 Kb. This default is optimized for 500K analyses. For 100K analyses, use 0.5 Mb. Genomic smoothing can be disabled by applying a smoothing bandwidth of 0 bp.

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 megabase smoothing bandwidth.

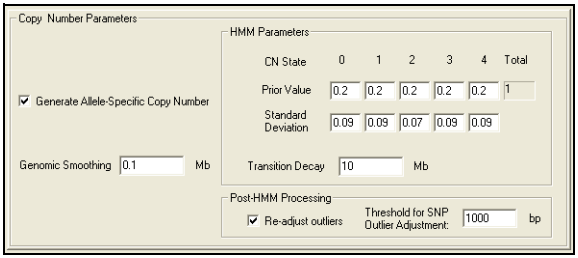

Figure 3.13 Copy Number Parameters - Advanced Options page

HMM Parameters – Priors

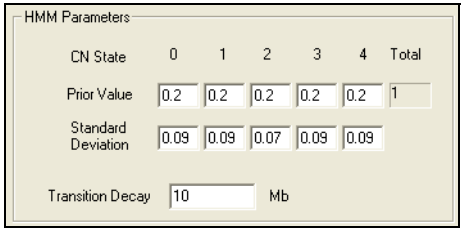

Figure 3.14 HMM Parameters

CN State – Prior Value:

A 5-State Hidden Markov Model (HMM) is applied for smoothing and segmenting the CN data. The Priors and Transition Decay Length are the two user-tunable parameters.

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

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.

CN State – 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 (see *Copy Number Parameter Settings* for suggested changes to this parameter). The default is 0.07 for state 2 and 0.09 for all other states (0, 1, 3, 4).

HMM Parameters – Transition Decay:

This parameter (Figure 3.14) 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

Re-adjust outliers:

This parameter allows for adjusting 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 readjust outliers parameter. It is the distance that is applied for determining 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) (Figure 3.13).

> **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.

Copy Number Parameter Settings

Analysis can be optimized to the specific copy number experiment by changing the algorithm parameters. The table below (Table 3.1) describes a set of recommended parameter settings for some common experimental conditions.

Table 3.1 Affymetrix recommended parameter settings for copy number

Copy Number	Footprint of Change	Restrict by Fragment Size	Reference Set	Probe-level nomaliza-tion	Gaussian Smoothing (kb)	HMM Priors	HMM Transition Decay (Mb)	HMM Standard Deviation	Adjust Outliers
Microdele-tions	< 4Mb		Un-paired ≥25	Median Scaling	low	Equal	≤1000	refer to BW versus SD table (algorithm in manual)	off
Chr X changes	Size of chr X		Un-paired ≥25	Quantile	100	Equal	1000	0.09 for states 0, 1, 3, 4 & 0.07 for state 2	on
Trisomy/ Disomy	Variable		Un-paired ≥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		Un-pairedun-paired ≥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)		Un-paired ≥25	Quantile	500	Equal	10	0.06 for states 0, 1, 3, 4 & 0.03 for state 2	on
Karyotype	1–5 Mb		Un-paired ≥25	Quantile	50	Equal	1	0.11 for states 0,1,3,4 & 0.08 for state 2	on
FISH (BAC clones)	200 Kb		Un-paired ≥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)	Un-paired ≥30	Quantile	100	Equal	1–100	0.09 for states 0,1,3,4 & 0.07 for state 2	on

LOH Parameters

HMM Parameters – Transition Decay

This parameter determines how correlated adjacent SNPs are to each other. The LOH state of any given SNP is partially dependent on that of its neighboring SNPs and are weighted based on the distance between them. By adjusting this parameter (Figure 3.15), neighboring SNPs can either have more or less of a dependence on each other. See the table below for recommended LOH parameter settings (Figure 3.16).

- The default value is 10 Mb.
- To reduce the influence of neighboring SNPs, decrease this value (transition faster). For example, if we 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).

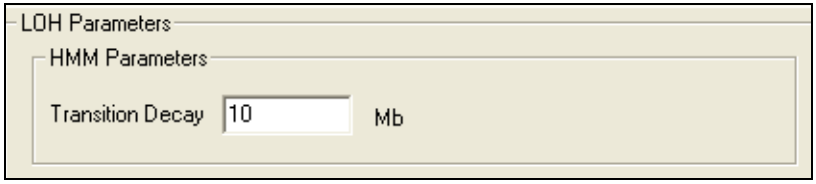

Figure 3.15 LOH Parameters - Advanced Options page

LOH Parameter Settings

Analysis can be optimized to the specific LOH experiment by changing the algorithm parameters. The table below (Figure 3.16) describes a set of recommended parameter settings for some common experimental conditions.

LOH	Reference Set	HMM Transition Decay (Mb)
Tumor - Normal Pairs	1	10
Unpaired	> 30 from mixed population	10
Unpaired	~ 30 from same population	10

Figure 3.16 Affymetrix recommended parameter settings for LOH

Step 5: Output File Naming (optional)

This step allows you to edit the output file name and select the virtual set rule, if applicable. Both of these steps are optional (Figure 3.17).

- By default, the output files are named as follows:
  - *CHPfilename.CN.cnt* – for copy number output: CCL-256D_NSP.brlmm.CN.cnt
  - *CHPfilename.LOH.cnt* – for LOH output: CCL-256D_NSP.brlmm.LOH.cnt
- To modify these names by adding a prefix or suffix to the default names, check the appropriate box and enter the text in the adjacent box.

NOTE: All *.cnt files are text files.

- In addition, the location of these output *.cnt files can also be chosen by selecting the appropriate Destination Folder (Figure 3.17).
- Virtual Sets allow you to group data from the 100K and 500K Human Mapping array sets into a single output *.cnt file. The CNAT 4.0 algorithm adjusts the baseline noise in both arrays individually, such that the median of the SNPs, which has a normal copy number of 2, are centered near zero. This corrects for different variations between the two arrays. Appendix C *How To Create Virtual Sets* contains additional information about setting up Virtual Sets.

  By default, when a Virtual Set is applied, the output files will be named as follows:
  - *VirtualSetRule.CN.cnt* – for copy number output *SampleID.brlmm.CN.cnt* or *CCL-256D.brlmm.CN.cnt*
  - *VirtualSetRule.LOH.cnt* – for LOH output *SampleID.brlmm.LOH.cnt* or *CCL-256D.brlmm.LOH.cnt*

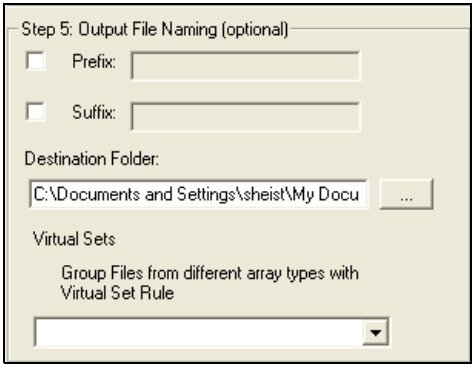

Figure 3.17 Output File Naming (optional)

NOTE: If a virtual set is not applied, each array will be analyzed individually. If a virtual set is not used, you can only run the analysis on one single array type at a time.

Step 6: Analyze

This step facilitates the following:

- Initiates the analysis by clicking the **Run** button (Figure 3.18).
- The **Stop** button stops the analysis.
- The **Help** button opens a copy of this document in a Help file format.
- When the run is complete (the status window indicates that the CN .exe was run successfully), continue with the instructions in the *CNAT Viewer* chapter to view the results.

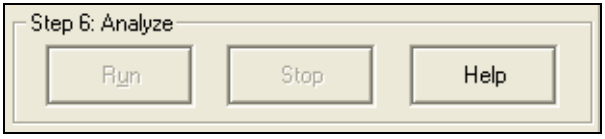

Figure 3.18 Run Analysis

Un-Paired CN and LOH Analysis

Samples without a paired normal from the same individual are compared to a set of references.

Availability of Reference Datasets From Affymetrix/HapMap Websites

There are 48 Normal HapMap samples available on the Affymetrix web site. Go to www.affymetrix.com and navigate to *Support/By Support Type/Data Resource Center/ Genotyping Mapping Data Sets/Mapping 500K Sample Data Set*. Additional HapMap samples are available on the HapMap web site and the GEO website. Go to www.hapmap.org and navigate to *Project Data/Bulk Data Downloads/Raw Data/ affy500K*. Or, go to http://www.ncbi.nlm.nih.gov/geo/ and enter accession number GSE5173.

Selection of the Reference Set

The selection of the reference set depends on several factors:

- The recommended number of samples in the reference set is:
  - 25 for CN analysis –
    For CN, references less than the stipulated number can be used, but depending on

the variability across references, data interpretation can be impacted.

- 30 for LOH analysis –
  In LOH, use of a small reference set eventually impacts the estimation of the normal rate of heterozygosity in samples. Greater than 30 references should be used if the reference set is constituted from a mixed-population pool.

- Gender of the samples in the reference set:
  The gender of samples used in the reference set impacts the results for Chromosome X. To avoid difficulty in data interpretation, a mixed gender reference set should not be used. If normal male samples are used, it should be understood that in this case the copy number on Chromosome X for the test sample(s) is compared against a haploid reference state for Chromosome X, unlike the autosomes.
  - If all male reference files are used, the denominator for the log ratios on Chromosome X will be 1, unlike the autosomes that will be 2, and results should be interpreted accordingly.
  - If reference files are used that contain a mix of male and female files, the results on Chromosome X may be difficult to interpret.
- Using the same reference set for CN and LOH. If you are going to perform both CN and LOH, it is recommended that the same set of reference samples be used. Therefore, it is advantageous to run the analysis for both CN and LOH simultaneously.

Un-Paired Analysis

To begin un-paired analysis:

1. Click the CNAT 4 Batch Analysis icon in the Shortcut Bar, or select **Run → CNAT 4 Batch Analysis** from the Menu bar (Figure 3.19).

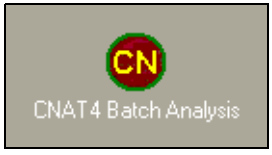

Figure 3.19 CNAT 4 Batch Analysis icon in the Shortcut Bar

The **CNAT 4 Batch Analysis** window opens (Figure 3.20).

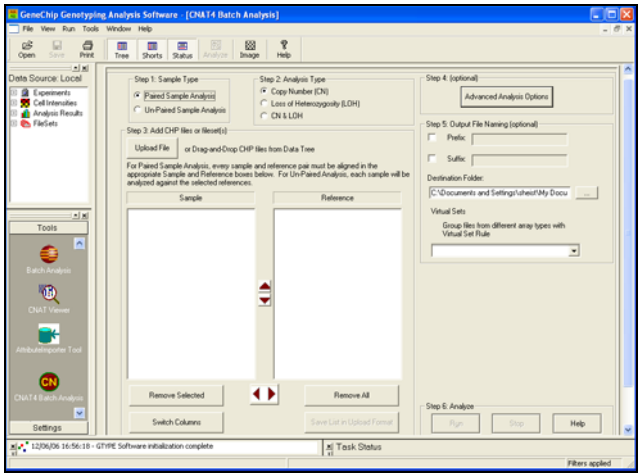

Figure 3.20 CNAT 4.0 Batch Analysis Window

2. Follow the steps outlined below to complete an **un-paired** sample analysis.

Step 1: Sample Type

- Select the **Un-Paired Analysis** option (Figure 3.21).

  In an un-paired analysis, the sample/tumor is compared to a set of normal references. See *Selection of the Reference Set* for more information on generating a reference set.

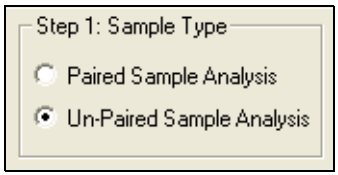

Figure 3.21 Sample Type group box – Un-Paired Analysis

NOTE: Affymetrix recommends that you use a reference set of at least 25 normal files for CN and 30 for LOH.

Step 2: Analysis Type

- In the **Step 2: Analysis Type** group box, select CN, LOH, or both (Figure 3.22).

  If only the CN radio button is selected or only the LOH radio button is selected, the software performs only copy number analysis or only LOH analysis on the selected samples. If you select the CN and LOH radio button, both CN and LOH analyses are performed on the selected samples.

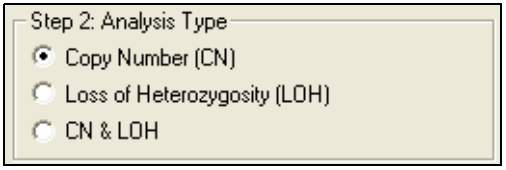

Figure 3.22 Analysis Type group box

- If you are going to perform both CN and LOH, Affymetrix recommends that the same set of reference files be used. Therefore, it is advantageous to run the analysis for both CN and LOH simultaneously.

Step 3: Add CHP files or fileset(s)

In the **Step 3: Add CHP files or fileset(s)** group box (Figure 3.23):

- Drag and drop sample or tumor CHP files from the data tree to the **Sample** list box.
- Drag and drop reference or normal CHP files from the data tree to the **Reference** list box.

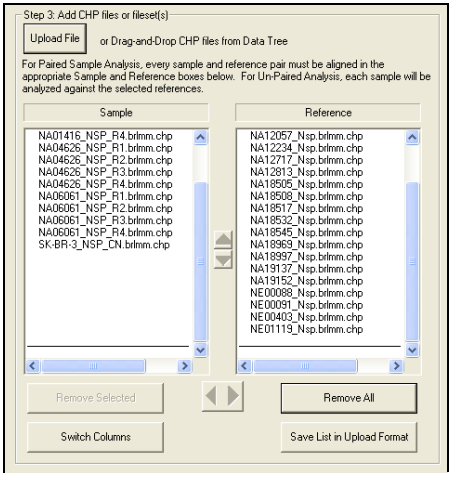

Figure 3.23 Add CHP files or fileset(s) group box – Un-paired Analysis

- Alternatively, you can upload a file list containing the CHP file names of the sample and reference files. The file consists of two columns where the first column is the sample/tumor CHP file name and the second column is the paired reference/normal CHP file name (Figure 3.24).

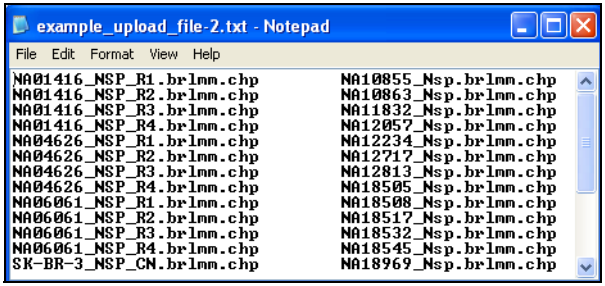

Figure 3.24 Upload File Format – Un-paired

NOTE: In an un-paired analysis, each sample is compared to the set of reference samples selected. The upload file format is the same (i.e., samples in the first column, references in the second column). However, the horizontal order has no effect on the analysis. See *Selection of the Reference Set* for more information.

Step 4: (optional) Advanced Analysis Options

- Click the **Advanced Analysis Options** button in the **Step 4** box (Figure 3.25).

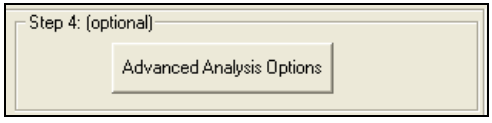

Figure 3.25 Advanced Analysis Options box

The **Advanced Options** page opens (Figure 3.26).

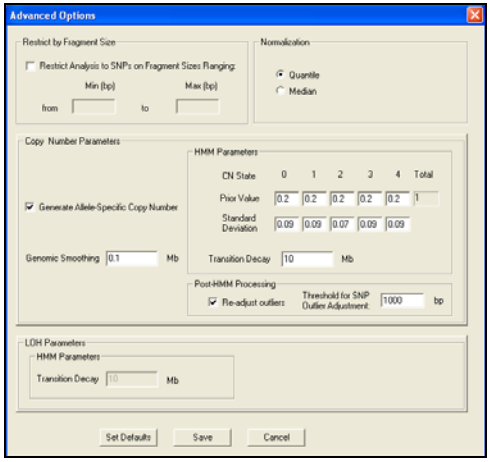

Figure 3.26 Advanced Options page

The Advanced Options page is divided into four sections:

- *Restrict by Fragment Size*
- *Normalization*
- *Copy Number Parameters*
- *LOH Parameters*

NOTE: If only CN or only LOH was selected in Step 2, then only the corresponding parameters are editable.

Restrict by Fragment Size

This option allows the analysis to be performed on only a subset of SNPs based on the fragment size that the SNPs reside on.

To enable this option:

1. Check the box next to **Restrict Analysis to SNPs on Fragment Sizes Ranging** (Figure 3.27).
2. Enter the size of fragments that you want to be included in the analysis.

   For example, if a sample has a known amount of degradation (no fragments larger than 600 bp), you can enter the following:

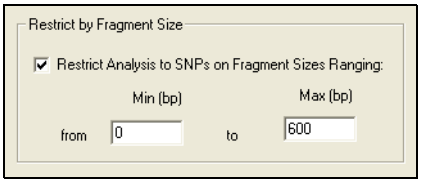

Figure 3.27 Restrict by Fragment Size - Advance Options page

In the example above, only SNPs that reside on fragment sizes less than or equal to 600 bp will be included in the analysis.

NOTE: The default is checked, which means that all SNPs will be included in the analysis.

Normalization

This option allows specification of the probe-level normalization. Select one of the following two options in the Normalization group box (Figure 3.28):

Quantile

Quantile normalization performs a sketch normalization, based on 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 MM probes are included to compute the median intensity of a CEL file.

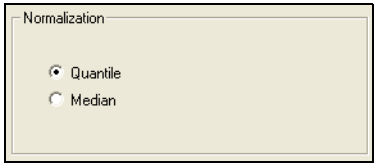

Figure 3.28 Normalization group box

Copy Number Parameters

Genomic Smoothing

The genomic smoothing option allows the user to specify the genomic smoothing length (in megabases) to be used (Figure 3.29). The genomic smoothing that is applied is a gaussian smoothing. The default bandwidth value is 100 Kb (0.1 Mb) which results in a

window size of 400 Kb. This default is optimized for 500K analyses. For 100K analyses, use 0.5 Mb. Genomic smoothing can be disabled by applying a smoothing bandwidth of 0 bp. See the table below for recommended CN parameter settings (Table 3.2).

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.

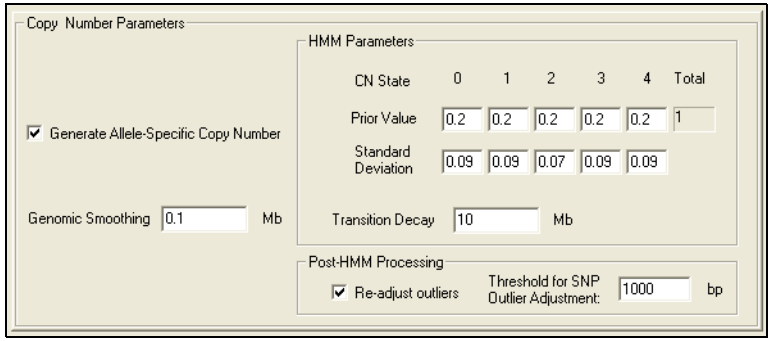

Figure 3.29 Copy Number Parameters – Advanced Options page

HMM Parameters – Priors

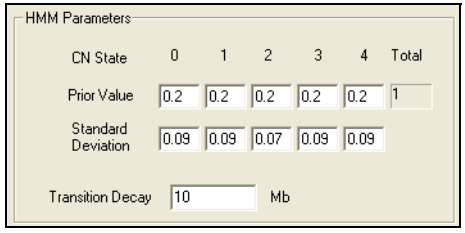

Figure 3.30 HMM Parameters

CN State – Prior Value:

A 5-state Hidden Markov Model (HMM) is applied for smoothing and segmenting the CN data. The priors and transition decay length are the two user tunable parameters.

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

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 (see *Copy Number Parameter Settings* for suggested changes to this parameter). The default is 0.07 for state 2 and 0.09 for all other states (0, 1, 3, 4).

HMM Parameters – Transition Decay:

This parameter (Figure 3.30) 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 are 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

Re-adjust outliers:

This parameter allows for adjusting 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 readjust outliers parameter. It is the distance that is applied, for determining if the flanking SNPs should impact the readjustment of the singleton SNP (Figure 3.29).

- 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.

Copy Number Parameter Settings

Analysis can be optimized to the specific copy number experiment by changing the algorithm parameters. The table below (Table 3.2) describes a set of recommended parameter settings for some common experimental conditions.

Table 3.2 Affymetrix recommended parameter settings for copy number

Copy Number	Footprint of Change	Restrict by Fragment Size	Reference Set	Probe-level nomaliza-tion	Gaussian Smoothin g (kb)	HMM Priors	HMM Transition Decay (Mb)	HMM Standard Deviation	Adjust Outliers
Microdele-tions	< 4Mb		Un-paired ≥25	Median Scaling	low	Equal	≤1000	refer to BW versus SD table (algorithm in manual)	off
Chr X changes	Size of chr X		Un-paired ≥25	Quantile	100	Equal	1000	0.09 for states 0, 1, 3, 4 & 0.07 for state 2	on
Trisomy/ Disomy	Variable		Un-paired ≥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		Un-paired ≥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)		Un-paired ≥25	Quantile	500	Equal	10	0.06 for states 0, 1, 3, 4 & 0.03 for state 2	on
Karyotype	1–5 Mb		Un-paired ≥25	Quantile	50	Equal	1	0.11 for states 0,1,3,4 & 0.08 for state 2	on
FISH (BAC clones)	200 Kb		Un-paired ≥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)	Un-paired ≥30	Quantile	100	Equal	1–100	0.09 for states 0,1,3,4 & 0.07 for state 2	on

LOH Parameters

HMM Parameters – Transition Decay

This parameter determines how correlated adjacent SNPs are to each other. The LOH state of any given SNP is partially dependent on that of its neighboring SNPs and are 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 (Figure 3.31). See the table below for recommended LOH parameter settings (Figure 3.32).

- The default value is 10 Mb.
- To reduce the influence of neighboring SNPs, decrease this value (transition faster). For example, if we 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).

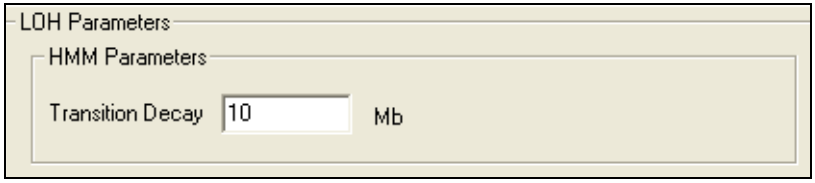

Figure 3.31 LOH Parameters - Advanced Options page

LOH Parameter Settings

Analysis can be optimized to the specific LOH experiment by changing the algorithm parameters. The table below (Figure 3.32) describes a set of recommended parameter settings for some common experimental conditions

LOH	Reference Set	HMM Transition Decay (Mb)
Tumor - Normal Pairs	1	10
Unpaired	> 30 from mixed population	10
Unpaired	~ 30 from same population	10

Figure 3.32 Affymetrix recommended parameter settings for LOH

Step 5: Output File Naming (optional)

This step allows you to edit the output file name and select the virtual set rule, if applicable. Both of these steps are optional.

- By default, the output files are named as follows:

- *CHPfilename.CN.cnt* – for copy number output:
  CCL-256D_NSP.brlmm.CN.cnt
- *CHPfilename.LOH.cnt* – for LOH output:
  CCL-256D_NSP.brlmm.LOH.cnt

- To modify these names by adding a prefix or suffix to the default names, check the appropriate box and enter the text in the adjacent box.

NOTE: All *.cnt files are text files.

- In addition, the location of these output *.cnt files can also be chosen by selecting the appropriate Destination Folder (Figure 3.33).
- Virtual Sets allow for the grouping of data on the 100K and 500K Mapping array sets into a single output *.cnt file. The CNAT 4.0 algorithm adjusts the baseline noise in both arrays individually such that the median of the SNPs, which has a normal copy number of 2, are centered about zero. This corrects for different variations between the two arrays.
  By default, when a Virtual Set is applied, the output files will be named as follows:
  - *VirtualSetRule.CN.cnt* – for copy number output
    SampleID.brlmm.CN.cnt or *CCL-256D.brlmm.CN.cnt*
  - *VirtualSetRule.LOH.cnt* – for LOH output
    SampleID.brlmm.LOH.cnt or *CCL-256D.brlmm.LOH.cnt*

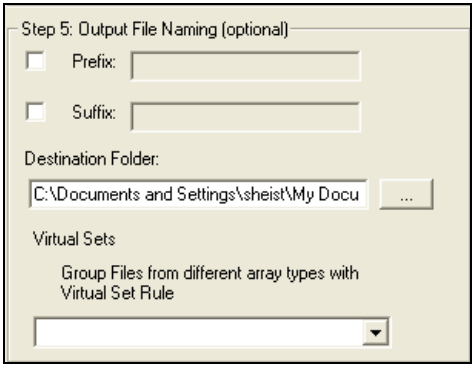

Figure 3.33 Output File Naming (optional)

Step 6: Analyze

This step facilitates the following:

- Initiates the analysis by clicking the **Run** button (Figure 3.34).
- The **Stop** button stops the analysis.
- The **Help** button opens a copy of this document in a Help file format.

- When the run is complete (the status window indicates that the CN .exe was run successfully), continue with *CNAT Viewer* in Chapter 4 to view the results.

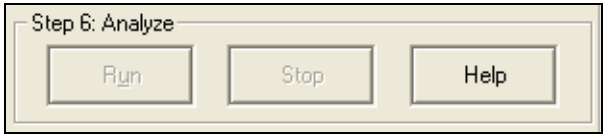

Figure 3.34 Run Analysis

CN Report File

Once the copy number analysis is complete, a report file will automatically open and display the interquartile range (IQR) of the un-smoothed log2ratio smoothed total CN for each sample. The report file will contain an IQR value for each chromosome as well as for the whole sample. In addition, in a paired analysis, the IQR values will be 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). 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 metric measure of dispersion. In general the sample-level IQR should be comparable to the chromosomal IQR for the given sample. An observed discordance in a chromosomal observation is potentially indicative of a biological change.

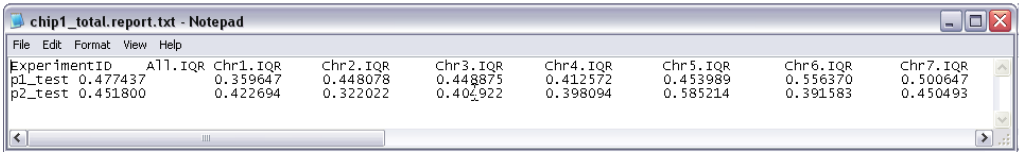

Figure 3.35 Copy Number Report File

NOTE: When doing a paired analysis, 3 report files will be generated, one for each allele and one for the total un-smoothed log2ratio. Only the total un-smoothed log2ratio report will be automatically displayed.

NOTE: The report file(s) are written to the output directory selected in Step 5 of the procedure for paired analysis (on page 23) and for un-paired analysis (on page 35).

CNT File Format

Header Section

The resulting *.cnt data files contain the following information in the header:

- Information about the array (# SNPs, probe array type, 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 *.cn.cnt Files

The resulting *.cnt data files contain the following data:

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.

Column Heading	Definition
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)

Column Heading	Definition
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 *.loh.cnt Files

The resulting *.cnt data files contain the following data:

Column Heading	Definition
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)

Chapter 4

CNAT VIEWER

The CNAT Viewer enables the visualization and manipulation of data generated by the CNAT 4.0 and 3.0 Batch Analysis Tools.

This chapter decribes the following:

Overview of CNAT Viewer

Key features include:

- Whole genome and chromosome specific views
- Multi-sample view to facilitate identification of trends in copy number or LOH data
- Dynamic filtering to enable thresholding of data
- Data exports to genomic browsers (IGB and UCSC)

Using CNAT Viewer

To view CN or LOH data files in the CNAT Viewer:

1. Click the **CNAT Viewer** button in the Shortcut bar, or select **Run → CNAT Viewer** in the menu bar (Figure 4.1).

Figure 4.1 CNAT Viewer icon in the Tools menu

The CNAT Viewer opens and the user is prompted to select a *.cnt file (Figure 4.2).

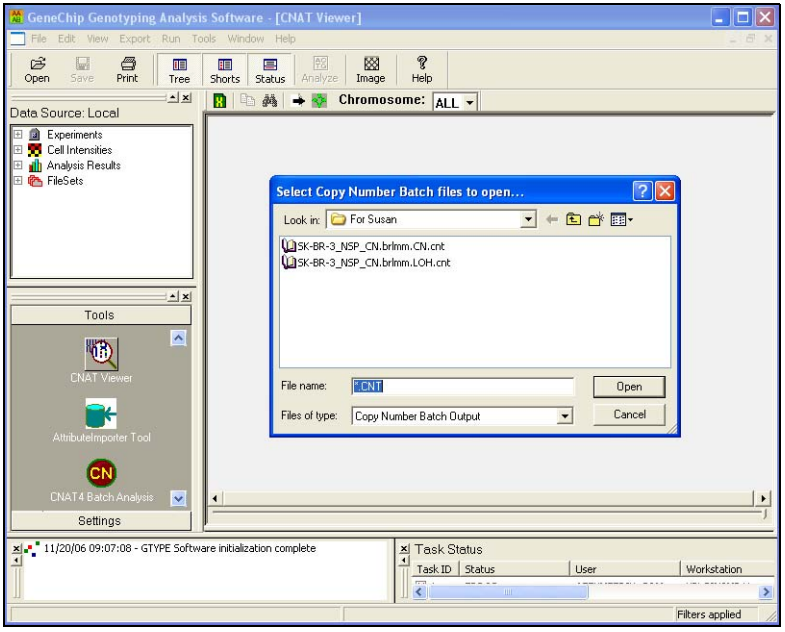

Figure 4.2 CNAT Viewer – Select Copy Number Batch Files window

2. Select the CN or LOH data files (select *.CN.cnt or *.LOH.cnt) you want to view and click **Open**.

   The selected *.cnt files open in the viewer and the whole genome is displayed (Figure 4.3).

Figure 4.3 CNAT Viewer – Whole Genome View

NOTE: When opening data types, the array type must be the same for all the selected samples. If you open two different array types, a warning box appears indicating that the two array types are not compatible and cannot be visualized (Figure 4.4).

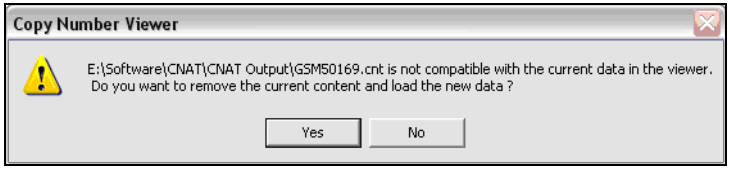

Figure 4.4 Warning message when switching between array types

3. To add additional *.cnt files, select **Edit** → **Add**, or click the **Add** icon .

CNAT Viewer Graphs

By default, the following graphs are displayed for paired and un-paired analysis:

Paired Analysis

CN

Pane 1: Log2Ratio

Pane 2: CNState

Pane 3: Log2RatioMin and Log2RatioMax

Pane 4: HmmMedianLog2Ratio

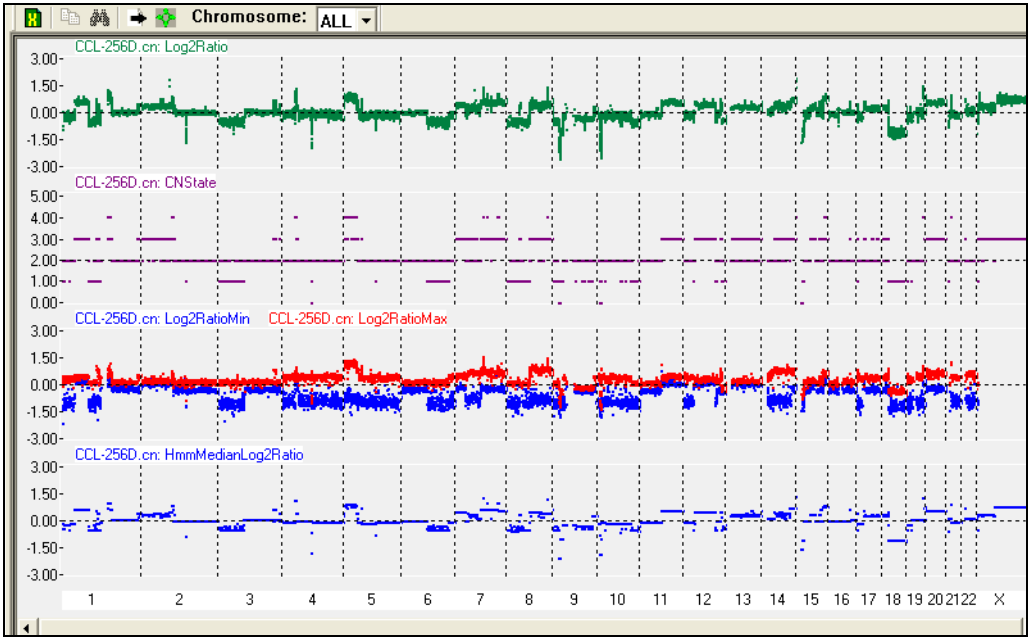

Figure 4.5 Default CN Graph – Paired

LOH

Pane 1: LOHState

Pane 2: LOHProb and RetProb

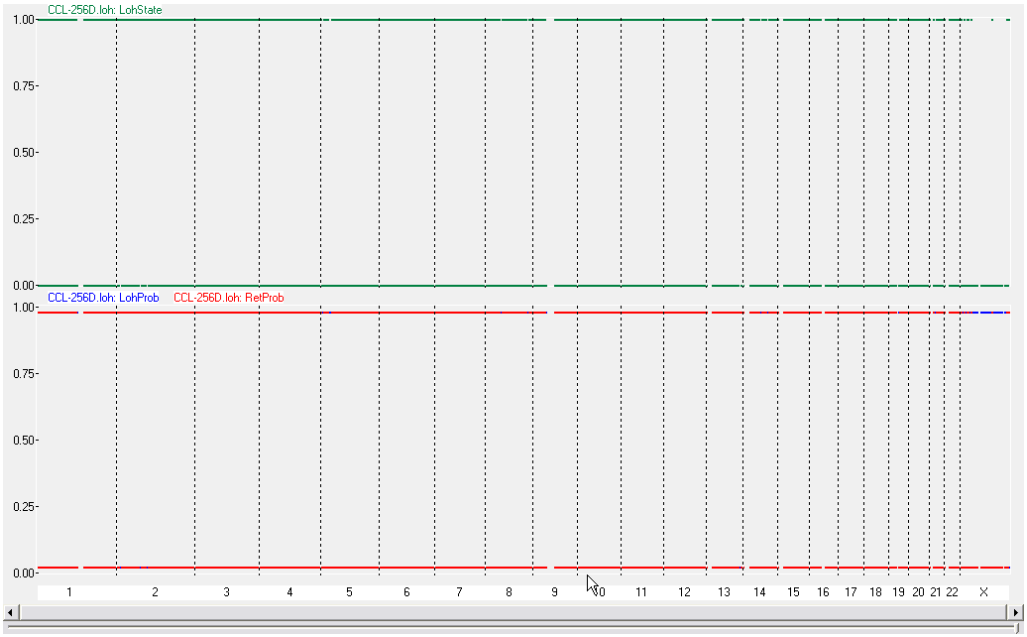

Figure 4.6 Default LOH Graph – Paired

Un-paired Analysis

CN

Pane 1: Log2Ratio

Pane 2: CNState

Pane 3: HmmMedianLog2Ratio

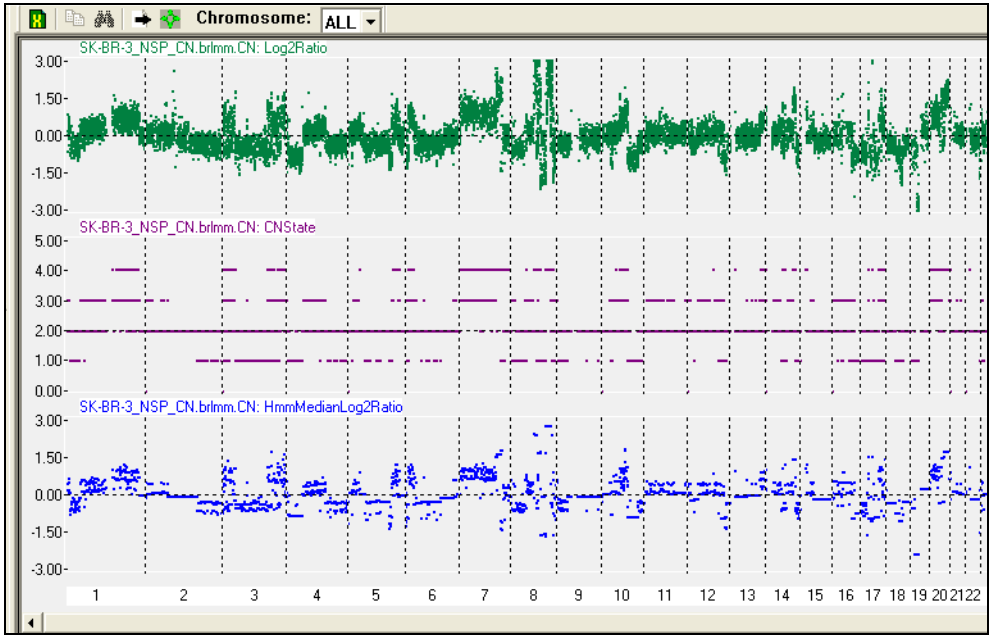

Figure 4.7 Default CN Graph – Un-paired

LOH

Pane 1: LOHState

Pane 2: LOHProb and RetProb

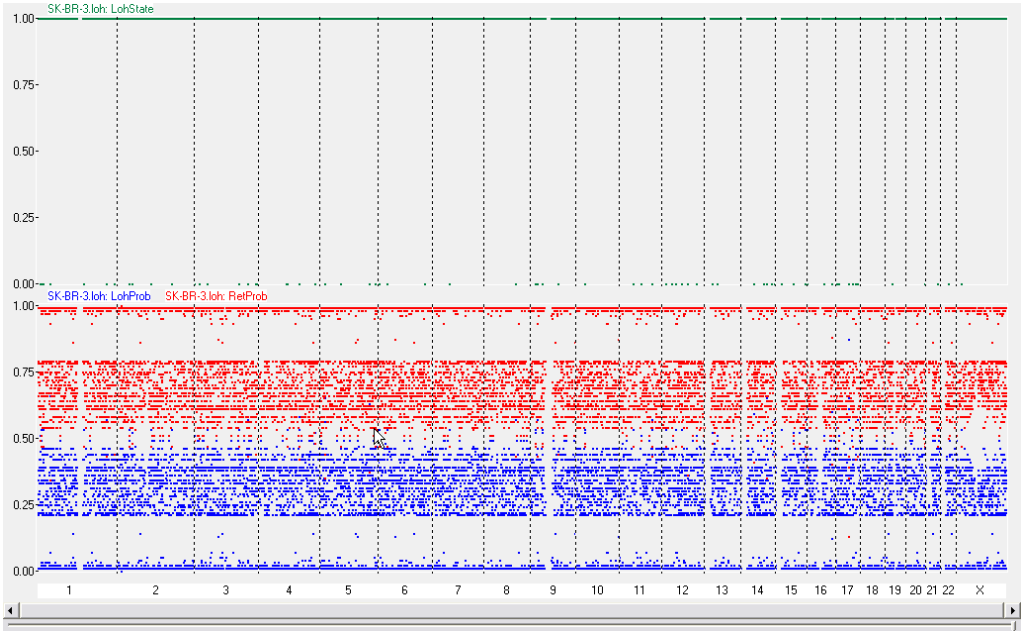

Figure 4.8 Default LOH Graph – Un-paired

Features

There are two main features of the graphical view:

- In all graphical views, the X-axis corresponds to a physical position along the chromosome and the Y-axis is the value of the metric displayed.
- The graph name (displayed above each graph) indicates the *.cnt filename followed by the metric graphed. For example, *SK-BR-3_NSP_CN.brlmm.CN:Log2Ratio* indicates that the Log2Ratio metric of the sample file *SK-BR-3_NSP_CN.brlmm.CN.CNT* is displayed in the first chart (Figure 4.7).

Column Headings

The *.CN.cnt and *.LOH.cnt data files contain specific information in the column headings. See *Data Section – For *.cn.cnt Files* and *Data Section – For *.loh.cnt Files* in Chapter 3.

Data Display Options

To switch between whole genome view and single chromosome view:

1. Go to the drop-down menu located above the graph and select a chromosome to be displayed in the graphical view.

   The graph displays the selected chromosome (Figure 4.9).

> **NOTE:** The Table view still displays all data from all chromosomes.

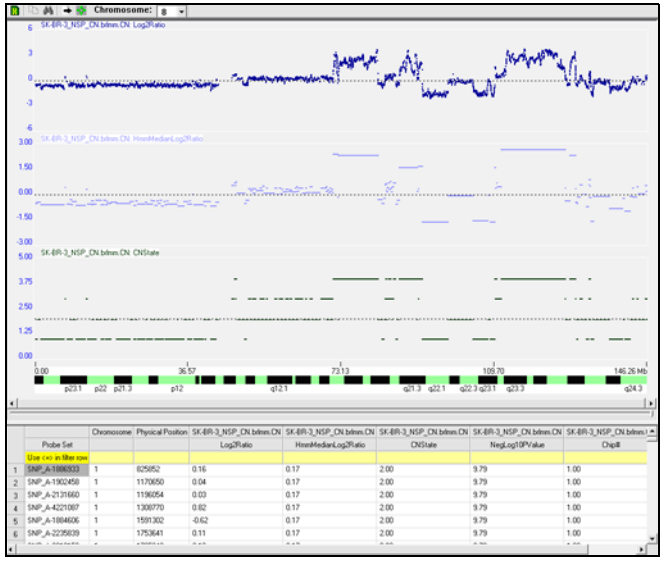

Figure 4.9 SKBR3 Breast Cancer Sample – Chromosome 8

2. To view the whole genome, select the **ALL** option under the drop-down menu.

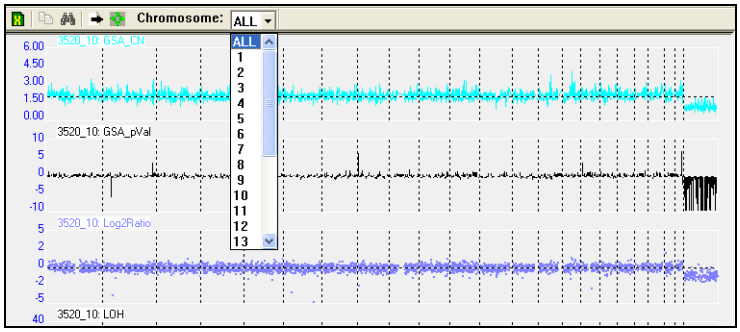

Figure 4.10 Select **ALL** to view whole genome

To change the graph display settings:

1. Select **View** → **Default Settings**, or the  icon to open the Default Settings dialog box (Figure 4.11).

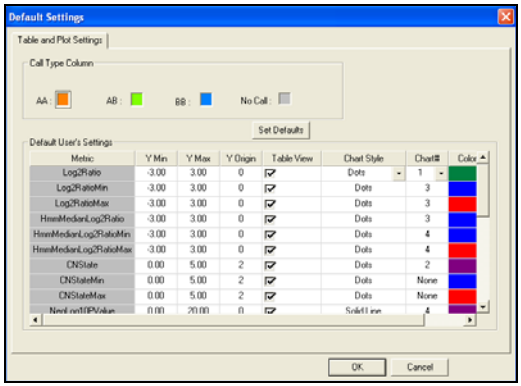

Figure 4.11 Default Settings dialog box

2. In the Default Settings dialog box, specify the following:
   - The columns you want to display in the table (check the appropriate box)
   - The X- and Y-axis maximum and minimum values (Y Min, Y Max)
   - The Y-origin (where the horizontal X-axis crosses the Y-axis)
   - The default Chart Style

   For each graph type, you can select the default chart style (**Dots**, **Dotted Line**, **Solid Line**, **Vertical Line**) by clicking in each row under the **Chart Style** column, clicking the drop-down menu, and selecting the style (Figure 4.12).

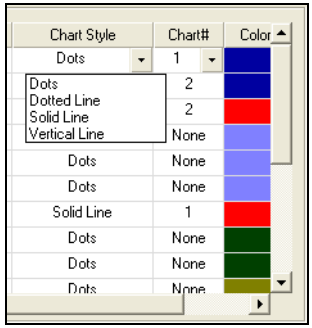

Figure 4.12 Chart Style default menu

- The default Color of the graphs

> **NOTE:** **The graph types can also be adjusted when new graphs are added. See *Adding Graphs* for more information.**

- The default Chart # corresponds to the location of the graph in the graphical pane. A metric with a chart # of 1 is plotted at the top of the graphical pane. For example, the Log2Ratio graph is blue in the figure above (Figure 4.12).

> **NOTE:** **Multiple charts can be graphed in a single pane. See *Adding Graphs* for more information.**

- The settings in the Default Settings dialog box are remembered from session to session. If you change one of the default settings, the software displays this change the next time you launch the application.

> **IMPORTANT:** **When you change any of the defaults, the change is only applied to new graphs (Figure 4.13). The changes will be applied to all graphs in the next session.**

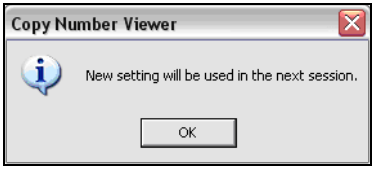

Figure 4.13 New Settings Warning

Adding Graphs

To add graphs:

1. Select **View** → **Plot Column,** or right-click in the graph space and select **Add Graph** (Figure 4.14).

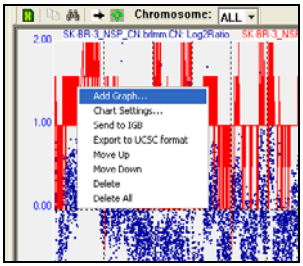

Figure 4.14 Right-click in the graph to select Add Graph

The **Plot Column** dialog box opens. This dialog box allows you to add graphs to either an existing chart or a new chart (Figure 4.15).

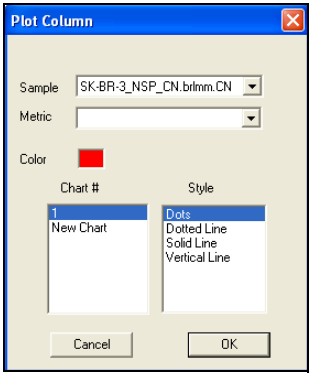

Figure 4.15 Plot Column dialog box

2. Select the sample from the drop-down menu (samples displayed correspond to open .cnt files).
3. Select the metric you want to plot in the second drop-down menu.
4. Select the color of the graph by clicking on the color swatch and selecting the new color from the color palette.
5. Select the chart # where you want the new graph to be displayed.
   - If you select **New Chart**, an additional chart is displayed on the bottom of the graphing pane.

- If you select an existing chart #, the graph is displayed in the selected chart #.

6. Select the style of the new graph (**Dots**, **Dotted Line**, **Solid Line**, or **Vertical Line**).
7. Select **OK** to plot the graph according to the selected settings.

Chart Settings

To modify chart settings:

1. Select **View** → **Chart Settings**, or right-click in the graph space and select **Chart Settings**.

   The **Chart Settings** dialog box allows you to modify the chart settings (Figure 4.16).

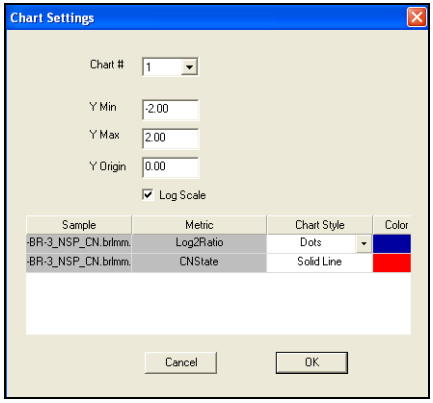

Figure 4.16 Chart Settings dialog box

2. Select the chart you want to modify by using the drop-down menu.
3. Modify the Y-axis minimum and/or maximum values as well as the Y-axis origin value by entering new values in the appropriate boxes.
   - To change the graph from linear to log scale, check the **Log Scale** box.
   - To change the chart style or color, use the drop-down menu or select the color swatch.

Moving and Deleting Graphs

- To move a graph up or down in the graphical pane, right-click on the graph and select either **Move Up** or **Move Down** (Figure 4.17).
- To delete a graph from the graphical pane, right-click on the graph and select either **Delete** or **Delete All**.

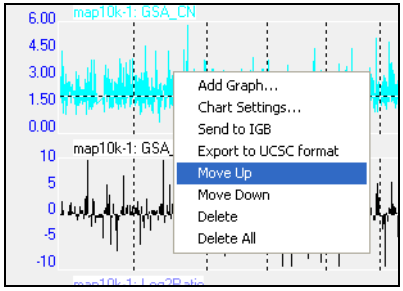

Figure 4.17 Right-click on graph to see Move and Delete options

Table Options

Several options are available that allow you to manage the tabular form of the data. Right-clicking on any of the column headers results in the following options (Figure 4.18):

- Show All Columns
- Plot Column
- Hide Column
- Sort Ascending
- Sort Descending
- Sort

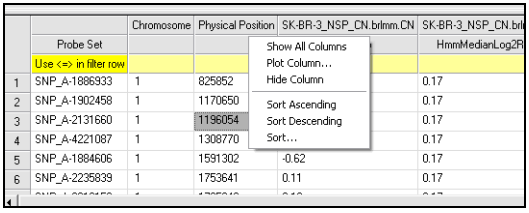

Figure 4.18 Right-click Options from the Table Header

To manipulate the column data:

- Select one of the following options:

- **Show All Columns** displays all data in the table.
- **Plot Column** allows you to select any column of data and add it to the graphical pane above. When selected, the Plot Column dialog box is displayed (Figure 4.15).
- **Hide Column** allows you to hide a column by right-clicking on a specific column and selecting Hide Column (Figure 4.18).

To sort the table data:

- Select one of the following options:
  - **Sort Ascending** sorts the selected column of data in an ascending order.
  - **Sort Descending** sorts the selected column of data in a descending order.
  - **Sort** brings up a new dialog box that allows you to sort on multiple columns, or right-click anywhere in the table, other than the header, to view the **Sort** dialog box (Figure 4.19).

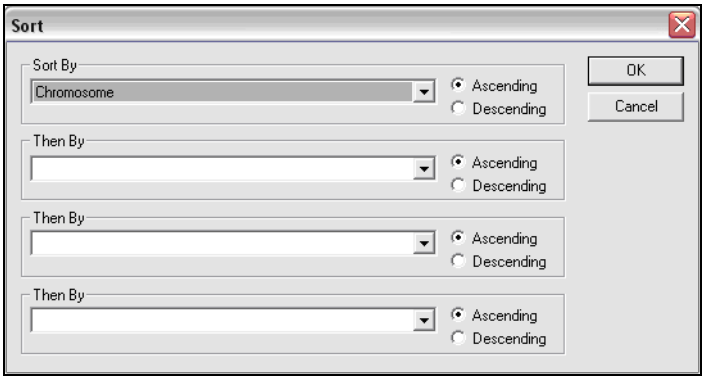

Figure 4.19 Sort dialog box

- Right-clicking anywhere in the table, other than the header, results in a different set of options (Figure 4.20):
  - Use **Edit** → **Find** or the **Find** icon  to search the table for a particular SNP or any other value in the table.
  - Clicking and selecting data from the table and selecting the **Copy Cells** option, the table data can be copied to a text editor.
  - The sort option brings up a new dialog box (Figure 4.19) that allows you to sort on multiple columns.

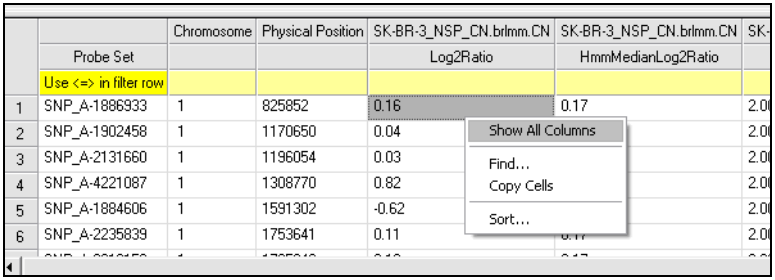

Figure 4.20 Right-click anywhere EXCEPT table header to view these options

Dynamic Filtering

- To filter by any metric or combination of metrics, double-click the yellow color row in the Table view and enter the filter condition.
- The filter conditions entered for the column are automatically applied to both the table and graphs. The table and graphs will only display the SNPs satisfying the filter condition (Figure 4.21).

Figure 4.21 Filtered graph and table by Log2Ratio greater than 1.5

The following filter conditions are valid:

>	Filter all values greater than the entered number
<	Filter all values less than the entered number
=	Filter all values equal to the entered number
>=	Filter all values greater than or equal to the entered number
<=	Filter all values less than or equal to the entered number

NOTE: **Exporting filtered data results in only SNPs meeting filtered criteria written to the exported file.**

Export

Tab-Delimited Report

To create a report:

1. Select **Export** → **Table** or click on the  icon. The **Export As** dialog box opens (Figure 4.22).
2. Select the location for the exported file and the file name.
3. Select **Export All** for the entire table or **Export Selected** to export a selected region of the table.
4. Click **Save**.

NOTE: **Exporting from the CNAT Viewer allows you to combine copy number results from different samples into a single file (if more than one .cnt file is open in the table).**

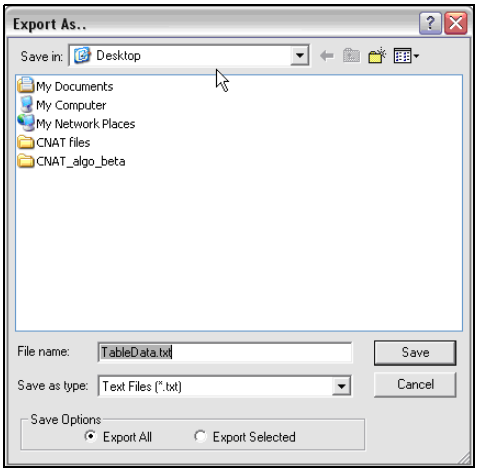

Figure 4.22 Tab-delimited Export As dialog box

IGB Format

To export to IGB format:

1. Select **Export** → **IGB Format** **(Figure 4.23)**.

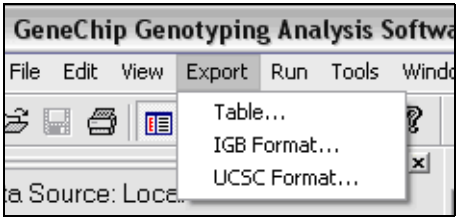

Figure 4.23 Export options

The **Export to IGB** dialog box opens (Figure 4.24).

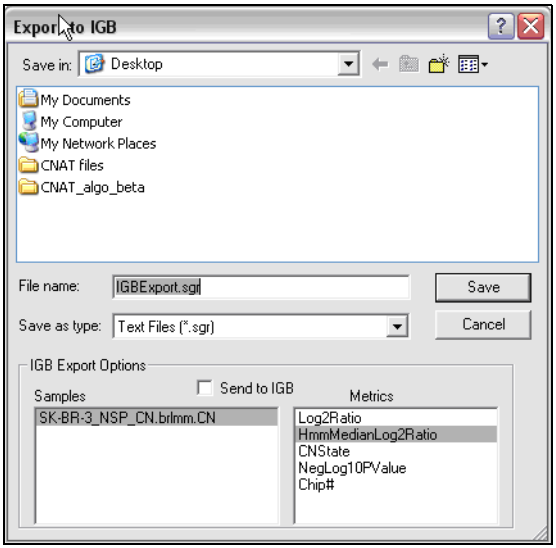

Figure 4.24 Export to IGB dialog box

2. In the **IGB Export Options** field, select one **Sample** and one **Metric** to export to the IGB format (*.sgr file).
3. Select the **File name** and the location where you want to save the file.
4. Click **Save**.

> **NOTE: If you checked the Send to IGB checkbox, the corresponding graph automatically opens in IGB (if you have IGB software installed).**

5. Right-click on a graph in the viewer and select **Send to IGB** (Figure 4.25).

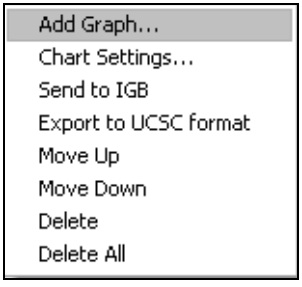

Figure 4.25 Right-click on graph in viewer to view this menu.

The selected graph automatically opens in IGB to the correct genome build and window size (e.g. whole chromosome view or zoomed in view) (Figure 4.26).

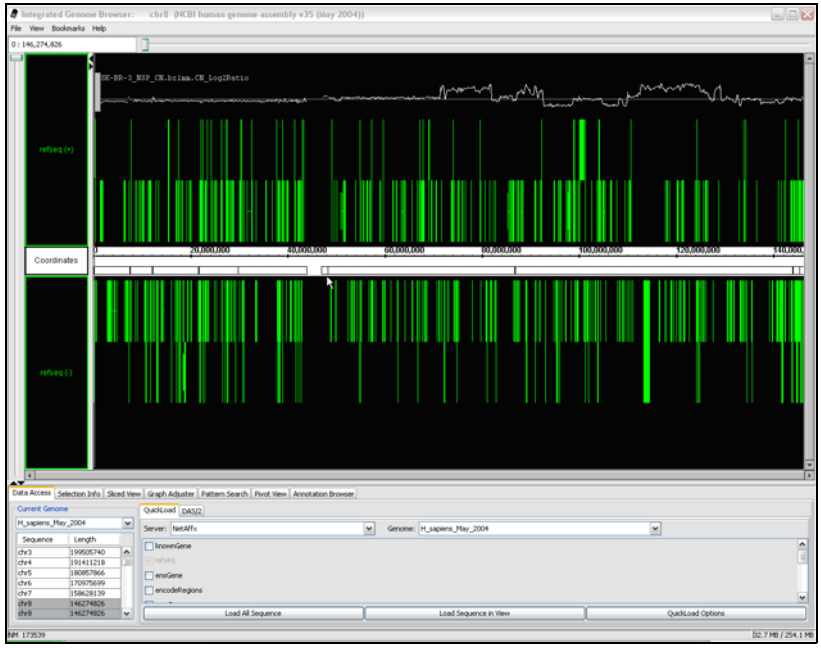

Figure 4.26 Log2Ratio graph exported to IGB

UCSC Browser

1. Select **Export** → **UCSC Format**.
2. In the **UCSC Export Options**, select one **Sample** and one **Metric** to be exported into the UCSC Wiggle format (*.txt file).

> **NOTE:** The wiggle (WIG) format allows display of continuous-valued data in a track format. For more information, see **www.genome.ucsc.edu**.

3. Select the **File name** and location where you want the file to be saved.
4. Save the file.
5. To open the file in the UCSC browser, go to www.genome.ucsc.edu and click **Genome Browser** (Figure 4.27).

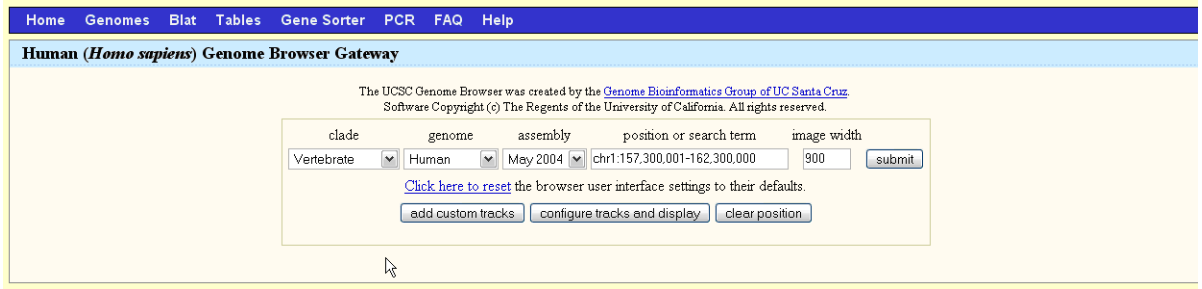

Figure 4.27 UCSC Genome Browser

6. Click the **add custom tracks** button (Figure 4.27).
   The **Add Custom Tracks** page opens (Figure 4.28).

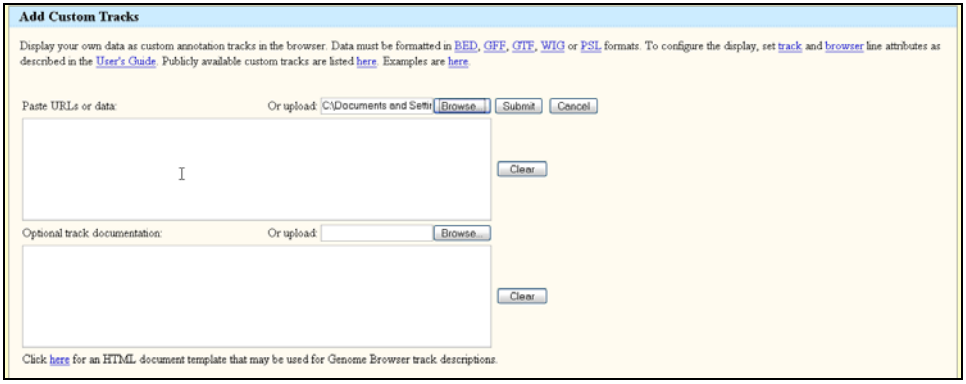

Figure 4.28 UCSC Genome Browser – Add Custom Tracks

7. Upload the .txt file generated by CNAT, when selecting **Export** → **UCSC** format.
   Data is displayed in the UCSC Genome Browser (Figure 4.29).

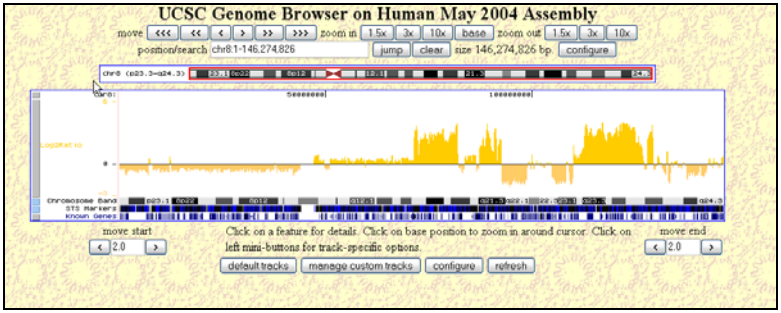

Figure 4.29 UCSC Genome Browser – Log 2 Ratio Data

Printing/Capturing Graphs

Images of graphs can be copied to the system clipboard and pasted into a graphic file.

To take a screen shot:

1. Press **ALT + Print Screen** on the keyboard. The image of the window is copied to the system clipboard.
2. Open a graphics program, such as Microsoft Paint, Corel® Paint Shop Pro, or Adobe® Photoshop®, and paste the image into a graphics file. The graphics file can then be printed or pasted into a document.

Appendix A

CNAT 4.0 ALGORITHM

This appendix describes the algorithm used in the Affymetrix GeneChip® Chromosome Copy Number Analysis Tool (CNAT) Software v.4.0:

Overview

The Affymetrix GeneChip® Chromosome Copy Number Analysis Tool (CNAT) Software v.4.0 implements a Hidden Markov Model (HMM) [1] based algorithm to identify chromosomal gains and losses from intensity data generated from whole genome SNP Mapping arrays. The CNAT algorithm compares sample values to SNP-specific distributions derived from either a paired normal sample from the same individual or a pooled set of unrelated references to quantitate copy number.

The dynamics of genomic copy number changes fundamentally follow a binary mechanism, in that it is either representative of a diploid state or not. The nuances are introduced by the latter where the deletions are either homozygous or hemizygous in nature, and amplifications can be representative of a gain in a single chromosomal copy or multiple copies. From a data-model point of view, this is a perfect scenario for the detection of change-points in the chromosomal copy number and/or state of heterozygosity. Motivated by the above rationale, Affymetrix has applied an HMM-based framework to segment the Total Copy Number (TCN) changes, Allele-Specific Copy Number (ASCN) changes, and Loss of Heterozygosity (LOH) into the various states.

This appendix describes the algorithmic steps used in the CNAT 4.0 software to analyze data from the Affymetrix GeneChip® Human Mapping 500K array set, Human Mapping 100K array set, and the Human Mapping 10K array. Four analysis pipelines are incorporated into CNAT 4.0, and the algorithmic workflow for both Copy Number (CN)

and LOH are categorized under two experimental scenarios, paired and un-paired analyses.

- Un-paired samples: In this case, samples potentially unrelated and spanning a diverse population are analyzed together.

  Output:

  - Total Copy Number (TCN) estimates
  - Loss of Heterozygosity (LOH) estimates
- Paired samples: In this case, the test sample and reference are derived from tissues of the same individual.

  Output:

  - TCN and allele-specific copy number (ASCN) estimates
  - LOH estimates

Algorithm Workflows

CEL files with Perfect Match (PM) and Mismatched (MM) intensities and genotype calls (AA, AB, BB) that are generated with the Dynamic Model (DM) mapping algorithm or the Bayesian Robust Linear Model with Mahalanobis (BRLMM) distance classifier algorithm constitute the inputs to the CN and LOH analysis pipelines, respectively.

Copy Number Algorithm

The copy number algorithm performs the following steps as depicted in the analysis workflow diagram (Figure A.1).

1. Probe-level and SNP level filtering.
2. Probe-level normalization of signal intensity.
3. Allele-specific summarization for each SNP.
4. Global reference generation for un-paired experiments.
5. Raw copy number estimation.
6. Linear regression on the raw CN estimate to correct for artifacts introduced by the PCR fragmentation process.
7. PCR normalized CN data – gaussian smoothed for enhancement of the signal-to-noise ratio (SNR).
8. HMM-based segmentation to obtain the different CN state partitions. The HMM assumes 5 states and is described in detail below.

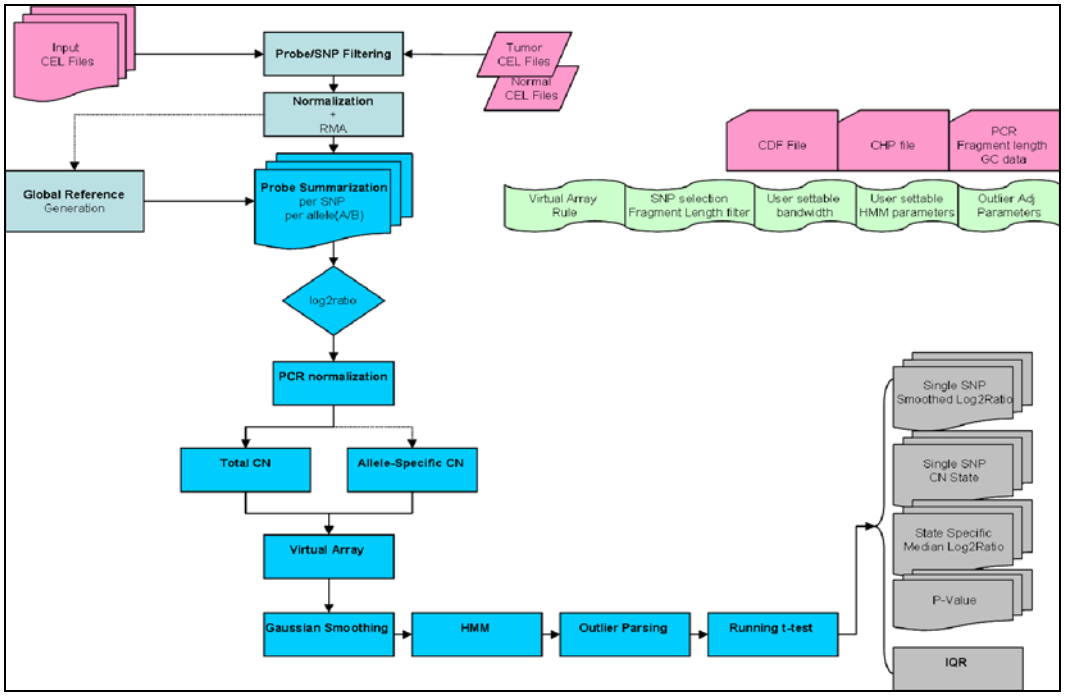

Figure A.1 Copy Number algorithm workflow

LOH Algorithm

The LOH algorithm performs the following steps as depicted in the analysis workflow diagram (Figure A.2):

1. A SNP-level filtering step.
2. Segmentation of the data into two states.

   LOH and normal retention of heterozygosity CEL files (which have Perfect Match (PM) and Mismatched (MM) intensities and genotype calls (AA, AB, BB) generated with the Dynamic Model (DM) mapping algorithm or the Bayesian Robust Linear Model with Mahalanobis (BRLMM) distance classifier algorithm) constitute the inputs to the CN and LOH analysis pipelines, respectively.

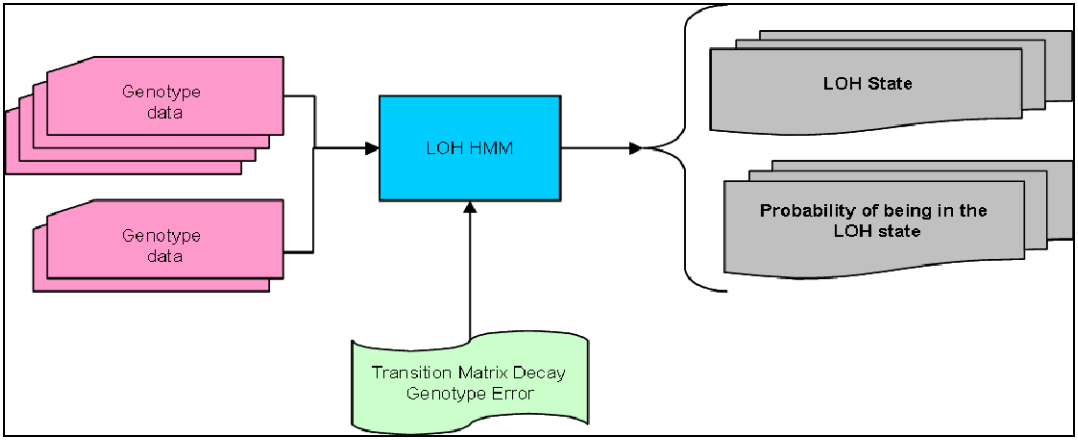

Figure A.2 LOH algorithm workflow

Algorithm Processes

Details of the algorithmic processes are described below.

Probe/SNP Filtering

This step determines the input to the Copy Number algorithm. Analysis is performed exclusively on PM probes present on the current array sets (Affymetrix GeneChip® Human Mapping 500K array set, Human Mapping 100K array set, and the Human Mapping 10K array). Additionally, users can exclude SNPs based on the length of the PCR fragment they are derived from. This length-based filtration is particularly effective for estimation of copy number in degraded DNA samples, for example fresh frozen paraffin embedded (FFPE) samples. Generally for FFPE samples, SNPs on smaller fragment lengths exhibit reduced degradation; the exclusion of SNPs on larger PCR fragment sizes have been shown to improve analytical accuracy.

Probe-Level Normalization of Signal Intensity

The normalization and summarization operations are performed across multichips with the goal to reduce experimental noise due to chip-to-chip variation, background, and relative variation in the performance of probes interrogating a given SNP, among others. Two modes of probe-level normalization have been implemented, median scaling and quantile normalization.

Median Scaling

This is a linear scaling operation. Here the scale factor is based on the median of all chip medians. A single chip median is computed based on the intensity of all probe pairs on the array. For paired sample analysis, Affymetrix recommends that median scaling be used.

Quantile Normalization

This method can be effective in normalizing non-linearities in the data [2]. A probe-level quantile normalization is effective in the elimination of biases introduced by experimental covariates such as, scanners, operators, and replicates among others. The sketch distribution against which all samples are quantile normalized is determined from a set of user selected references and by sampling across 50K PM probes in each of the references.

The choice and number of reference samples used is critical and is discussed in greater detail in the *Global Reference Generation For Un-Paired Experiments* section. It should be noted that tumor and normal samples are not quantile normalized separately within each of their respective groups. Rather, a full quantile normalization is performed across the test (tumor) and normal samples. If there is concern about any dilution effect in the data, median scaling should be used instead.

Results of a Receiver Operating Characteristic (ROC) curve analysis performed on the *Log2Ratio* is summarized for ten samples hybridized to both NSP and STY arrays and analyzed in the un-paired mode (Table A.1). The ROC analysis integrates the concept of sensitivity and specificity of a diagnostic test. The metric presented here is the Area under the Curve (AUC). For perfect performance of a test, the AUC is *one*, indicative of 100% specificity and 100% sensitivity. However, a generalized interpretation is when several methods are being compared, and the method with the highest AUC is the method with the highest performance, and hence the preferred one. Based on the AUC, quantile normalization seems to provide better performance (Table A.1) (*values in shown with asterisk*) in the case of NSP array and median scaling in the case of STY array. The margin of difference is slightly higher in the case of NSP, potentially favoring the quantile normalization overall. Analysis performed on X chromosomes of normal male samples, where the expected chromosomal copy number is *one*, showed variable response with the two normalization methods. In the case of NSP, for four out of ten samples, the median scaling erroneously estimated a dominantly diploid state; this aberration was not observed in the case of STY.

In general, for:

- Un-paired analysis – quantile normalization is recommended.
- Paired analysis – linear operation of median scaling is recommended, since both the test and reference samples originate from the same individual.

Table A.1 AUC Comparison for Quantile Normalized Versus Median Scaled Data. (Asterisks indicate higher AUC values.)

Experiment ID	NSP Quantile	NSP Median Norm	STY Quantile	STY Median Norm
K1	*0.942	0.922	0.904	*0.912
K2	*0.856	0.808	0.911	*0.915
K3	*0.837	0.792	*0.865	0.864
K4	*0.837	0.831	0.902	*0.912
K5	0.900	*0.904	0.884	*0.862
K6	*0.843	0.829	*0.849	0.837
K7	*0.812	0.795	0.882	*0.885
K8	*0.843	0.818	0.878	*0.879
K9	*0.836	0.817	0.884	*0.889
K10	*0.784	0.753	0.901	*0.909

Allele-Specific Summarization For Each SNP

Initially the intensity values are summarized for the A and B alleles of each SNP, where the allele designation is based on the alphabetical order of the base represented at the mutation. For example, a SNP with a sequence AT {G/A} TTT has an A and B allele designated by AT{A}TTT and AT{G}TTT, respectively. Subsequent to the probe-level normalizaton, a median polish operation is performed to generate a SNP level, allele-specific summarization across all relevant probes. (*In the command-line mode, the additional option of plier in RMA mode is also available.*) This operation takes into account both (a) sample-level effects arising from chip-to-chip (sample-to-sample) variation and (b) feature-level effects arising from systemic differences in feature intensities due to the hybridization specificity of different probe sequences. For N experiments performed, each SNP therefore has a 2 x N summary matrix, corresponding to the two alleles.

In general, the summarized intensity of the A versus the B allele is governed by the quantity of the particular allele in the target genome. Hence, the ratio of the summarized A allele to B allele intensity in normal populations might exhibit an ethnicity bias. It should be noted, that the A versus B allele distinction is different from the minimum versus maximum allele intensity distinction where the latter is analogous to allelic frequency. The output of the allele-specific analysis comprises the CN estimates for the min and max alleles.

Global Reference Generation For Un-Paired Experiments

In the case of an un-paired analysis, a global reference is generated from a pool of unrelated reference samples. Pooling mitigates impact from outliers, random variations across the samples among others. There are three factors governing the global reference generation.

- Number of samples
- Gender of samples
- Ethnicity of samples

Number of normal samples required to generate the optimal reference distribution:

In the un-paired mode, a number of normal controls are required to generate a stable sketch distribution against which all samples can be quantile normalized. Since the quantile is a smooth and slow varying function, the quantile distribution for each chip is interpolated from approximately 50K PM probes. There is negligible difference, in terms of CN estimation, between the sketch being derived from a sampling of 50K versus all available PM probes. Stabilization of the sketch is estimated by starting with a nominal number of normal samples, two in this case, and increasing the number of samples in the normal population until the resultant quantile distributions converge to within 1e-4. Based on this titration, convergence was achieved for 25 normal samples. This optimization was performed without consideration of ethnicity.

Gender of the references:

It is recommended that only normal female samples are used for reference generation, otherwise it might lead to interpretational difficulty in the estimation of changes observed in chromosome X. If CN estimation in chromosome X is not a concern, a pooled set of male and female samples can be utilized.

Impact of ethnicity on the generation of the reference:

The references can either be ethnically matched or not to the test sample. Provided sample-size considerations are addressed, the outcome does not demonstrate a significant ethnicity bias. In many cases, the ethnicity of the test sample is unknown; however, if there is concern about dilution of the estimated copy number from sampling across a diverse population set, the references can certainly be restricted to the known ethnicity type. It was established that for a given ethnicity, more than 25 samples, independent of quality, did not enhance the quantile sketch. This method obviates the need for the selection of optimal references [3], and, in fact, a random sampling of 25 references provides a sufficient mechanism for removal of outliers. If reference samples are unavailable in the lab, the HapMap references (www.affymetrix.com) can be used, and equal proportions can be pooled across ethnic diversity.

Raw Copy Number Estimation

The raw copy number estimation involves generating the log2ratio between the test or tumor sample and a reference sample. In the case of the paired and un-paired analyses, the paired normal and global references are used as the respective references. CN is generated for every allele of every SNP. $\Lambda_i^{1,2}$ is the raw CN for the i^{th} SNP between test sample, S^1, and reference, S^2.

$$\Lambda_i^{1,2} = \log_2\left(\frac{S_i^1}{S_i^2}\right) \text{ for } i^{th} \text{ SNP}$$

Two variants on Total Copy Number (TCN) estimate have been implemented. The methods are referred to as *sumLog* and *logSum* as detailed in the equations below. Although the TCN estimate is co-optimized for bias and variance; the logSum and sumLog approaches further optimize the output for variance and bias, respectively. In the GUI mode only the variance optimized method is available (*in the command-line mode, the type of optimization can be selected*). The choice of the optimization method has an impact on the degree of gaussian smoothing required and consequently, the choice of the HMM parameters as discussed in the following sections.

$$\text{logSum} \rightarrow TCN = \log_2\left(\frac{S_A}{R_A + R_B} + \frac{S_B}{R_A + R_B}\right)$$

$$\text{sumLog} \rightarrow TCN \sim \log_2\left(\frac{S_A}{R_A}\right) + \log_2\left(\frac{S_B}{R_B}\right)$$

R_A, R_B: copy number in the allele {A,B} of the reference

S_A, S_B: copy number in the allele {A,B} of the sample being tested

Therefore, for each SNP, three raw CN estimates are generated:

- TCN
- CN for allele A
- CN for allele B

PCR Normalization

Subsequent to the estimation of the raw copy number, an allelic and SNP-level normalization is performed to correct for potential artifacts introduced by the polymerase chain reaction (PCR) process. The normalization methodology was developed by researchers at University of Tokyo [3, 4]. The corrected CN is indicated by ${}^cA_i^{1,2}$. In the quadratic PCR correction implemented as a linear regression, GC content and fragment length of the PCR fragment that a given SNP resides on are the covariates.

In general, the PCR correction has less of an impact in 500K compared to 100K data. Furthermore, in the case of the 500K array, length (rather than GC content) seems to be a stronger contributor to the PCR issue. This is evident in the plots of the length and GC content versus CN pre- and post-PCR correction (Figure A.3). Slight straightening of the curve is observed in the case of the length; no change is observed for GC content.

$$\Lambda_i^{1,2} = \log_2\left(\frac{S_i^1}{S_i^2}\right) \text{ for } i^{th} \text{ SNP}$$

$$(\Lambda_i^{1,2} = {}^{c}A_i^{1,2}) + q(x, y)$$

$$q(x, y) = (A + bx + b'y + cx^2 + c'y^2)$$

where x = GC and y = *fragment length*

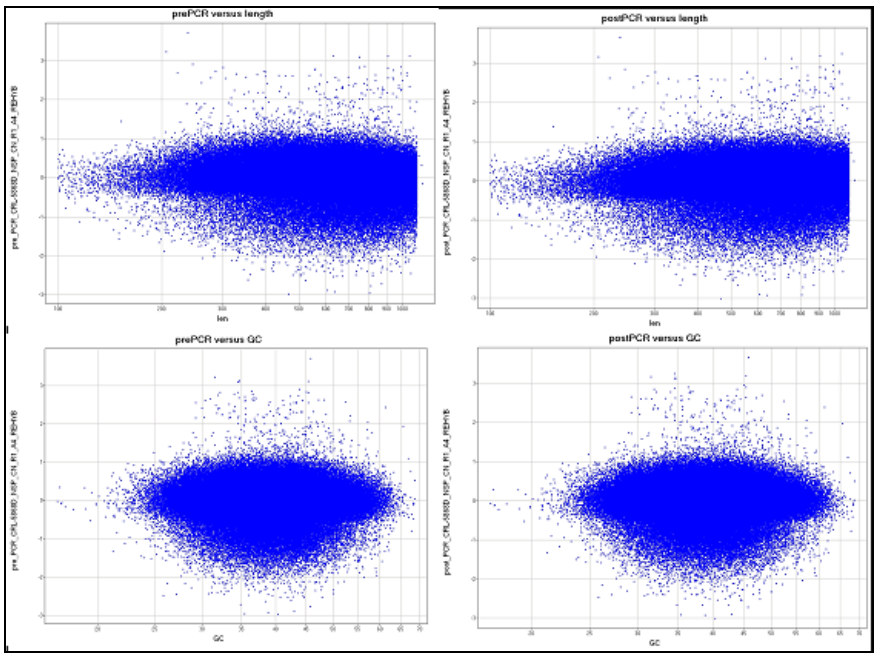

Figure A.3 Pre and post PCR versus length and GC

Generation of a Virtual Set

Generation of the virtual set constitutes combining 2 arrays to create a single virtual array. The arrays combined must be of the same type; for example, two 50K arrays are combined to create a 100K virtual array. This method includes an inter-array normalization to adjust for the noise baseline and variance across the arrays. The normalization parameters are estimated based on a subset of SNPs belonging to the copy-neutral region (CNR). The exact bounds (*adjustable in the command line application*) of the CNR envelope are not adjustable in the GUI and are fixed to ±0.2. The underlying transformation is a shift-and-scale one, where the corrected log2ratio is shifted by the mean of the SNPs in the CNR and is scaled by the variances of the arrays. After the transformation, the SNPs are sorted based on chromosomal and physical location and merged across both arrays. Without the inter-array normalization, there is a high probability of certain regions of the genome being erroneously labeled as CN aberrations.

Gaussian Smoothing of Data

Dependant on the noise in the data, it is often advisable to perform some smoothing prior to estimation of CN. Prior to the HMM segmentation of the CN data into discrete states representative of copy-neutral, amplification and deletion, an optional gaussian smoothing operation might be performed. Smoothing basically increases the signal-to-noise ratio in the data and enhances the demarcation between regions of CN change. The degree of smoothing is often governed by the experimental question at hand, especially the genomic footprint of the aberrations being detected. The table in *Copy Number Parameter Settings* displays various experimental scenarios and suggested smoothing parameters. In the case of applying a gaussian smoothing, it is possible for the user to set a smoothing bandwidth (σ) or use the default σ = 100K bases.

Smoothing is performed on the PCR smoothed (and inter-array normalized for virtual array) CN data. It is performed for every index SNP(i) by considering the CN contributions from all flanking SNPs encompassed in a window (W), where $W \approx 4\sigma$. Specifically, all flanking SNPs within 2σ to the left and right are considered. The choice of bandwidth affects the level of variance-reduction obtained in the data, as shown in the figure below. While smoothing preserves the overall trend in the data, the level of smoothing might mask micro-aberrations and a $\sigma \sim 10 - 15$KB (instead of 100KB) might be more appropriate in these cases. The modulation of σ is interconnected with the modulation of the HMM parameters, specifically the standard deviation discussed in the Hidden Markov Model section.

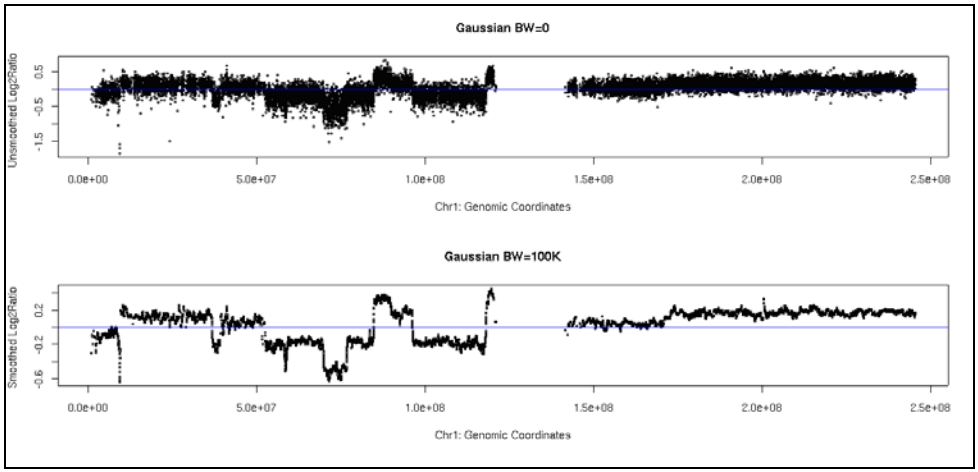

Figure A.4 Comparison of log2ratio data using σ =0 (no smoothing) versus σ =100kb. Smoothing enhances the boundaries of CN changes.

Hidden Markov Model

The underlying assumption in the HMM, is that the experimental observation, CN and/or LOH value of a SNP, is generated by a hidden or unknown process. However, the emitted (observed) CN values, which are more of a continuum, hint of the true hidden copy number state.

The underlying states are discrete in nature. In this HMM model, there are five hidden states representing the following biological phenomenon:

- Homozygous deletion or State = 0
- Hemizygous deletion (haploid) or State = 1
- Copy neutral (diploid) or State = 2
- Single copy gain or State = 3
- Amplification (multiple copy gain) or State ≥ 4

According to the HMM, a SNP_i can exist in any one of the five hidden states (S). Based on the observed CN (O) of SNP_i and its neighboring SNPs, as well as the model λ discussed below, we need to determine exactly which hidden state SNP_i exists in.

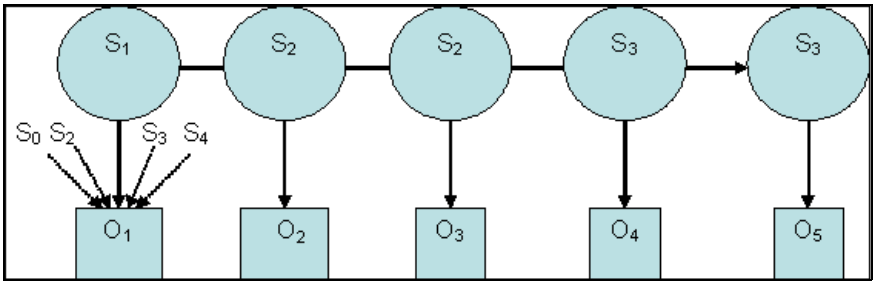

Figure A.5 Probability of an observation sequence

In the absence of knowing the truth, we take a probabilistic approach, where a given SNP has a non-zero probability that it can exist in each of the five states (Figure A.5). The goal here, based on a certain model λ (with parameters discussed below) is to determine which of the five states is the most likely one. Therefore, we construct the path of maximum likelihood while traversing all the SNPs. Stated succinctly, we are trying to evaluate the probability of an observation sequence $\{0_1, 0_2, 0_3 ... 0_n\}$ given a model λ.

There are three parameters governing λ:

- Prior probability
- Transition probability
- Emission probability

The HMM parameters that are user-tunable via the GUI relate to the above described probabilities. It should be noted that all HMM related parameters entered in the GUI are initial guesses and are optimized by the underlying HMM algorithm. So while these are not expected to be precise by any means, a certain combination of values, as described below, can lead to difficulties with interpretation of data.

Prior Probability (referred to as the *prior,* π)

Prior probability refers to the *a priori* knowledge of a SNP to exist in a preferred state. Fundamentally, under normal diploid conditions, the prior probability of a SNP to be in State 2 is almost 100%. However, since there are five states in this model (and in a probabilistic approach we should assume that a given SNP has a non-zero probability of being in any of the five states), we shall restate the prior probability distribution for a diploid sample as such: where the prior probability of a SNP to be in state 2 is 96% (maximal) and to be in any one of the other states is 1% (minimal):

States $\{0, 1, 2, 3, 4\} \rightarrow \pi \{0.01, 0.01, 0.96, 0.01, 0.01\}$

The property governing the prior requires their sum be equal to 1 (see the equation below). It should also be noted that priors are not assigned on a SNP by SNP basis, but to the population as a whole.

$$\sum_{s=1}^{S} \pi_s = 1$$

where S = states

In a disease condition where aneuploidy is a hallmark, the prior probability is significantly different from the above example and is not known *a priori.* In this case, it suffices to let the priors be equal, with $\pi = 0.2$ for all states, that is 0.2 (0.2*5=1), and let the algorithm adjust to determine the optimal prior per state. In general for all experiments, it suffices to assume that no *a priori* information is available about the states of the SNPs, and the prior for each state is initialized to 0.2.

Transition Probability

This is the probability of transitioning from one hidden state, S_i, to another, S_j. Often in annotating CN aberrations, it is evident that a contiguous set of SNPs exist in a deletion state; in LOH, clusters of contiguous SNPs are shown to exhibit loss or retention, and it is very seldom a singleton SNP effect. This indicates that an underlying correlation exists across the hidden states of neighboring SNPs. In general terms, the hidden state of any SNP is impacted by the status of the prior SNP(s) in genomic space – this introduces a concept of the probability of sustaining/transitioning from a hidden state to another. Here we support the ergodic model described as follows:

- If SNP_i exists in hidden state 0, then the allowed transition for SNP_{i+1} can be any of the states from 0–4.

If SNP_{i+1} happens to be proximal, there is a high probability that it will continue to exist in state 0. Conversely, if SNP_{i+1} were distal, then it might have equal or increased probability of transitioning to an amplification or copy-neutral state.

The UI tunable parameter determines the rate of decay of the transition correlation. A smaller value implies a faster transition from the current underlying state to a different underlying state and vice-versa. The units of this parameter are in base pairs, and a minimum value of 100 bp is advised.

Emission Probability

Emission probability reflects the probability with which the underlying state (S) is emitted to produce the observable (O). There are two parameters that govern this probability, the mean and the standard deviation (SD) of each underlying state. The SD which reflects the dispersion in each hidden CN state is the only user tunable parameter. It sets an initial guess on the expected dispersion in the hidden state. If the CN data is smoothed, then the expected SD should be tight (low); if the data is noisy, the SD should be high. As a rule, the lower the bandwidth (implying that the underlying CN data is unsmoothed or noisy), the higher the SD parameter. Refer to the table below for suggested bandwidth versus SD estimates. Generally, the SD of state 2 is tighter than the

SD in the other states.

Table A.2 Bandwidth versus SD estimates for States 0 thru 4

Bandwidth	Standard Deviation	
	States 0, 1, 3, 4	State 2
0	0.25	0.19
15K	0.15	0.13
30K	0.12	0.09
50K	0.11	0.08
100K	0.09	0.07
500K	0.06	0.03

P-Value Estimation

A running t-test is performed on the PCR-normalized CN estimate on a per chromosome/per sample basis. The null hypothesis is: the mean CN in a given window (W) is equal to the baseline CN. The baseline is computed per sample and the W corresponds to the gaussian window size. In this regard, it is a windowed t-test, and therefore, p-values cannot be computed for unsmoothed data.

Allele-Specific Copy Number

Allele-specific copy number is computed exclusively for paired experiments. In this case, only the subset of SNPs that have heterogenous, genotype {AB} calls in the reference sample are considered. For this subset, the PCR-normalized A and B allele data (discussed in the *Raw Copy Number Estimation* section) is transformed into min and max allele, based on their intensity. The ASCN analysis pipeline is analogous to the TCN pipeline where the min and max allele data are processed independently.

LOH Algorithm Processes

The HMM for the LOH uses a two-state model:

- LOH or State = 1
- Retention/Normal or State = 0

Affymetrix recommends that BRLMM, as opposed to DM, be used for generation of the genotype data, since the former recovers a significantly higher percentage of heterozygous calls.

The underlying HMM is analogous to the CN analysis described above, but there are a few differences as highlighted below:

- Only the transition decay parameter is user tunable (*in the command line the*

genotyping error is user tunable), and all other parameters are computed based on the input data.

- For paired LOH, the analysis is performed exclusively on SNPs with a heterozygous genotype {AB} call in the reference.
- For un-paired LOH, based on ROC analysis, Affymetrix recommends that at least 40 references be used. In this case, the probability of being in LOH versus retention is modeled based on the SNP heterozygosity rate (in the reference pool) and the genotyping error rate. Therefore, data is generated for all SNPs.
- The LOH probability is estimated based on the marginal probability of the HMM; the Retention probability is estimated as 1-LOH probability. The marginal probability of a state is the possibility that we will observe that state without the knowledge of any other events.

References

1. Rabiner, L., *A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition*. Proceedings of the IEEE, Vol. 77(2), pp. 257–86. (1989)

2. Bolstad, B.M., Irizarry, R.A., et.al., *A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Bias and Variance*. Bioinformatics, Vol. 19(2), pp. 185–93. (2003)

3. Nannya Y., Sanada M., Nakazaki K., et.al., *A Robust Algorithm for Copy Number Detection Using High-Density Oligonucleotide Single Nucleotide Polymorphism Genotyping Arrays*. Cancer Res. Vol. 65(14), pp. 6071–79. (2005)

4. Ishikawa S., Komura D., Tsuji S., et.al., *Allelic Dosage Analysis With Genotyping Microarrays*. Biochem. Biophys. Res. Commun. Vol. 333(4), pp. 1309–14. (2005)

Appendix B

CNAT VERSION HISTORY

This chapter describes the following:

CNAT 1.0 (Updated June 2004) – Detailed History

CNAT 1.0.01 (original version) supported the following probe array type:

Mapping10K_Xba131

Analysis outputs are:

- Copy number
- P-value
- Meta p-value (calculated using Contiguous Point Analysis (CPA))
- LOH

Compatible with files from GCOS 1.0; GDAS 2.0

CNAT 1.1.01 (updated June 2004) supported the following probe array types:

- *Mapping10K_Xba131*: Algorithm and analysis outputs are the same as for version 1.0.01.
- *Mapping10K_Xba142*: Algorithm and outputs are the same as for *Mapping10K_Xba131*.
- *Mapping50K_Xba240* and *Mapping50K_Hind240*

Analysis outputs are:

- Copy Number
- P-value
- Meta p-value (for 10K arrays)
- Kernel smoothed p-value (for 50K arrays)
- LOH

CNAT 2.0 (Updated February 2005) – Detailed History

CNAT version 2.0 supported the following probe array types:

- *Mapping10K_Xba131*
- *Mapping10K_Xba142*
- *Mapping50K_Xba240* and *Mapping50K_Hind240*
- *Mapping 100K* (combination of Xba240 and Hind 240)

Analysis outputs are:

- Genotype Call
- SPA_CN (Single Point Analysis Copy Number)
- SPA_pVal (Single Point Analysis p-Value)
- CPA_pVal (Contiguous Point Analysis p-Value)
- GSA_CN (Genome Smoothed Analysis Copy Number)
- GSA_pVal (Genome Smoothed Analysis p-Value)
- LOH (Loss of Heterozygosity)

Compatible with files from GCOS 1.2; GDAS 3.0

CNAT 3.0 (Updated November 2005) – Detailed History

CNAT version 3.0 supported the following probe array types:

- *Mapping10K_Xba142*
- *Mapping50K_Xba240* and *Mapping50K_Hind240*
- *Mapping 100K* (combination of Xba240 and Hind 240)

Analysis outputs are:

- Genotype Call

- SPA_CN (Single Point Analysis Copy Number)
- SPA_pVal (Single Point Analysis p-Value)
- CPA_pVal (Contiguous Point Analysis p-Value)
- GSA_CN (Genome Smoothed Analysis Copy Number)
- GSA_pVal (Genome Smoothed Analysis p-Value)
- LOH (Loss of Heterozygosity)
- Log2 Ratio

Compatible with files from GCOS 1.3; GTYPE 4.0

CNAT 4.0 (Updated March 2007) – Detailed History

CNAT version 4.0 supports the following probe array types:

- *Mapping10K_Xba142*
- Mapping 100K (combination of Xba240 and Hind 240 below)
  - *Mapping50K_Xba240*
  - *Mapping50K_Hind240*
- Mapping 500K (combination of Nsp and Sty below)
  - *Mapping 250K_Nsp*
  - *Mapping250K_Sty*

Analysis outputs for *.cn.cnt files:

NOTE: Values in the chart below that are labeled ***matched analysis only*** require that the check box next to **Generate Allele-Specific Copy Number** (Advanced Options page) be selected.

Column Heading	Definition
Log2Ratio	Smoothed Log2Ratio value
HmmMedianLog2Ratio	Median Log2Ratio 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 Log2Ratio 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 (*matched 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 (*matched analysis only*)
CNStateMin	HMM copy number state of the allele with the lower signal intensity (matched 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 normal state (CN State 2) for that particular sample (*matched analysis only*)
Log2RatioMax	Smoothed Log2 ratio value for the allele with the higher signal intensity (*matched 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 (*matched analysis only*)
CNStateMax	HMM copy number state of the allele with the higher signal intensity (*matched 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 normal state (CN State 2) for that particular sample (*matched analysis only*)

Analysis outputs for *.loh.cnt files are:

Compatible with files from:

- GCOS 1.4 local
- GCOS 1.3 or a higher server
- GTYPE 4.0 or higher

Column Heading	Definition
Call	Genotype call for the tumor/test sample
RefCall	Genotype call for the matched reference sample (matched 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 an SNP is in LOH state (closer to 1 indicates a strong likelihood of LOH)
RetProb	Likelihood that an SNP is in Retention state (closer to 1 indicates a strong likelihood of Retention)

Genome Builds and SNP Positions

The physical positions in CNAT 4.0 are from NetAffx™. NetAffx annotations are updated quarterly and have the most recent information.

Array Type	Build Used in NetAffx – Available in GTYPE and CNAT 4.0
Mapping10K_Xba131	Build 35, May 2004
Mapping10K_Xba142	Build 35, May 2004
Mapping50K_Xba240	Build 35, May 2004
Mapping50K_Hind240	Build 35, May 2004
Mapping250K_Nsp	Build 35, May 2004
Mapping250K_Sty	Build 35, May 2004

Update the genome build and all other annotations by using the update NetAffx annotations in GTYPE. For more information about NetAffx, refer to the *Affymetrix GeneChip® Genotyping Analysis Software User's Guide* – P/N 702083.

Appendix C

How To Create Virtual Sets

This appendix contains information that illustrates the step-by-step process of creating a Virtual Set in GTYPE such that array files are grouped together to form Virtual arrays in the CNAT 4.0 batch analysis tool.

This appendix describes the following:

Using Attributes

Attributes in Virtual Sets

Before a virtual set can be created, all arrays that are included in the virtual set must contain a sample attribute that is common to all of them. If the arrays that you want to analyze by Batch Analysis do **not** have a common attribute, you will not be able to analyze them together. If you used a common Sample Name, this attribute can be used for setting up a virtual set.

Using Sample Name As a Common Attribute

In the figures below, the Mapping 250K_Nsp and Mapping 250K_Sty arrays both have a common Sample Name, *Test_Sample*. In these examples, Sample Name could be used as the Virtual Set Rule.

Adding Attribute Information to Samples

If the Sample Names are different for the Mapping 250K_Nsp and Mapping 250K_Sty arrays, there are two ways, Manual and Batch, of assigning attributes to your samples, which allows you to define and apply Virtual Set Rules.

For more information, see *Setting Up Sample Attributes* in the *Affymetrix GeneChip® Genotyping Analysis Software (GTYPE) User Guide* (PN 702083).

Adding Attribute Information Manually

In instances where arrays to be grouped do **not** have a common attribute such as Sample Name, a Sample Template must be created to facilitate a common attribute. Template attributes, such as SampleID, Person ID, or Individual ID, can be created manually. In the following figures, Sample ID, an attribute in the Pedigree template, is used (Figure C.1) (Figure C.2). For more information, see *Creating Sample Templates With Attributes* in the GTYPE User Guide.

Setting Up the Sample Template

To set up a sample template:

1. In GCOS, open the experiment file of the sample you want to add attribute information to.
2. From the template drop-down menu, select the template.

> **NOTE: Editing and assigning sample templates takes place within GCOS Manager. See *Creating Sample Templates With Attributes* in the GTYPE User Guide.**

3. Manually type the **Sample ID** name into the Sample ID row of the experiment file.

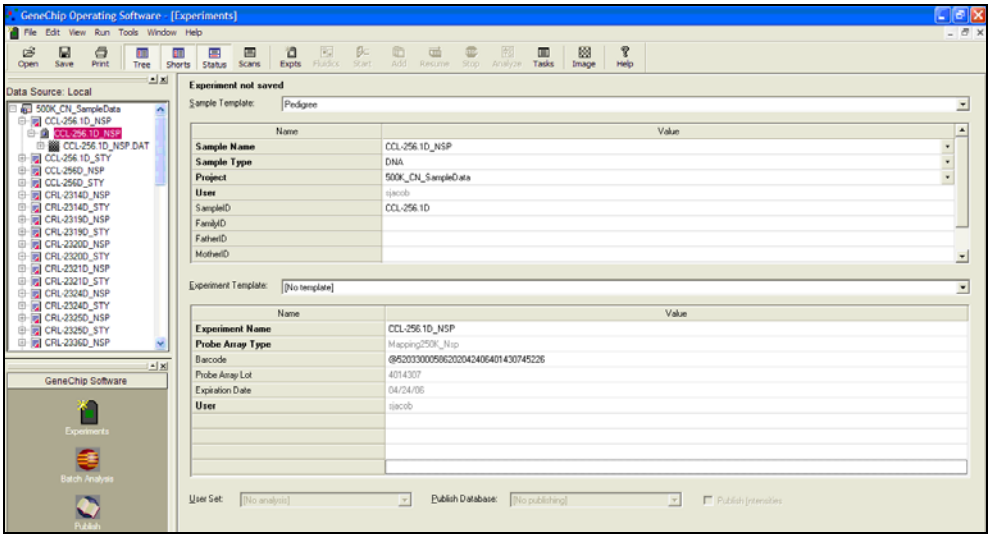

Figure C.1 Sample ID attribute for Mapping250K_Nsp

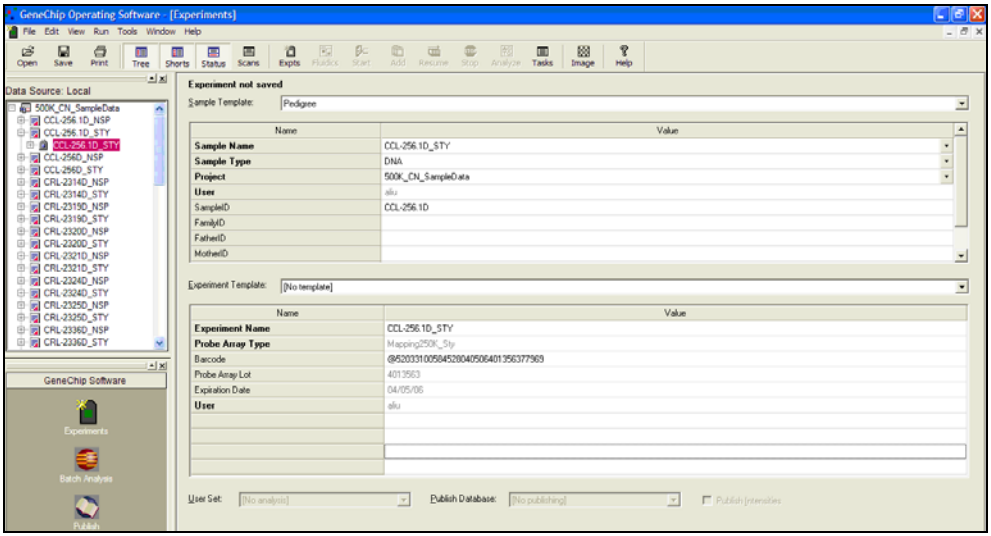

Figure C.2 Sample ID attribute for Mapping250K_Sty

NOTE: Assigning new sample attributes is done in the Experiment Setup Window in GCOS.

Adding Attribute Information with the Batch Feature

Using Attribute Importer Tool

Below is an abbreviated example (Figure C.4) using the Attribute Importer Tool to add attributes to samples in batch mode. See *Importing Pedigree and Attibute Information* in the GTYPE User Guide.

To add attribute information:

1. Before using the Attribute Importer Tool, you must:
   - **A.** Assign a sample template with attributes in GCOS. See *Creating Sample Templates With Attributes* in the GTYPE User Guide.
   - **B.** Create a data file in the correct format. (See the example table below (Table C.1).)The data file contains the following:
     - A tab-delimited text file (*Example_Importer.txt*) containing sample data to be imported.
     - Column headers in the first row, which are field names that identify the data type. Column headers do **not** need to match the attribute names.
     - Remaining rows contain information for the samples, one sample per row.

Table C.1 Data File Contents

Sample Name	Sample ID	Disease State
CRL-2325D_Nsp	CRL-2325D	Normal
CRL-2325D_Sty	CRL-2325D	Normal
CRL-5957D_Nsp	CRL-5957D	Normal
CRL-5957D_Sty	CRL-5957D	Normal
CCL-256.1D_Nsp	CCL-256.1D	Normal
CCL-256.1D_Sty	CCL-256.1D	Normal
CRL-2336D_Nsp	CRL-2336D	Tumor

Table C.1 Data File Contents

Sample Name	Sample ID	Disease State
CRL-2336D_Sty	CRL-2336D	Tumor
CRL-2338D_Nsp	CRL-2338D	Tumor
CRL-2338D_Sty	CRL-2338D	Tumor
CRL-2340D_Nsp	CRL-2340D	Tumor
CRL-2340D_Sty	CRL-2340D	Tumor

2. To batch-add attributes, open up GTYPE and select the **Attribute Importer Tool** from the Run menu, or click the Attribute Importer Tool icon in the tools window (Figure C.3).

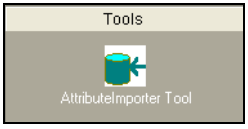

Figure C.3 Attribute Importer Tool icon in the Tools shortcut bar

3. Browse to the location of the tab-delimited text file containing your sample attribute information.
4. Edit the attribute name(s) in the Attribute column (Figure C.4) to assign data to the correct sample attribute, if necessary.
5. Select the attributes to be imported by using the check boxes.
6. Select options for import:
   A. Modify existing samples.
   B. Create new samples.
7. Select the template name from the drop-down box.
8. Select or enter the **Project Name** to assign to these samples.
9. Click the **Start Import** button.

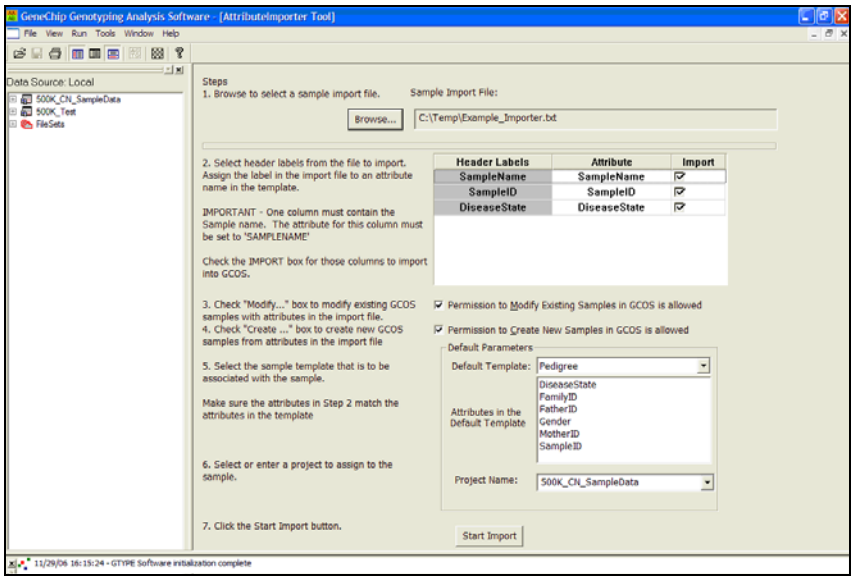

Figure C.4 Attribute Importer Tool window

Creating the Virtual Set

Below is an abbreviated example of creating a virtual set (Figure C.6). For more detailed information, see *Defining a Virtual Set Rule* in the GTYPE User Guide.

To create a virtual set:

1. Open the GTYPE software and select the **Virtual Set** from the Tools menu, or click the Virtual Set icon in the Tools window (Figure C.5).

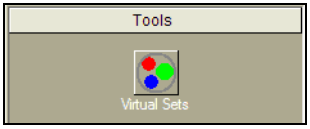

Figure C.5 Virtual Sets icon in the Tools shortcut bar

2. Create a unique name for this virtual set definition, for example, *Sample ID*.
3. Select the attribute(s) that are required to uniquely identify the sample:
   - **A.** Select the template used.
   - **B.** Use the Add and Remove buttons to select the attributes to be used in this virtual set rule. In this example, the attribute Sample ID from the Pedigree was used.

4. Select the arrays that you wish to combine into a virtual array. In the example below (Figure C.6), we use Mapping250K_Nsp and Mapping250K_Sty.
5. Select **Save** and close the dialog box.

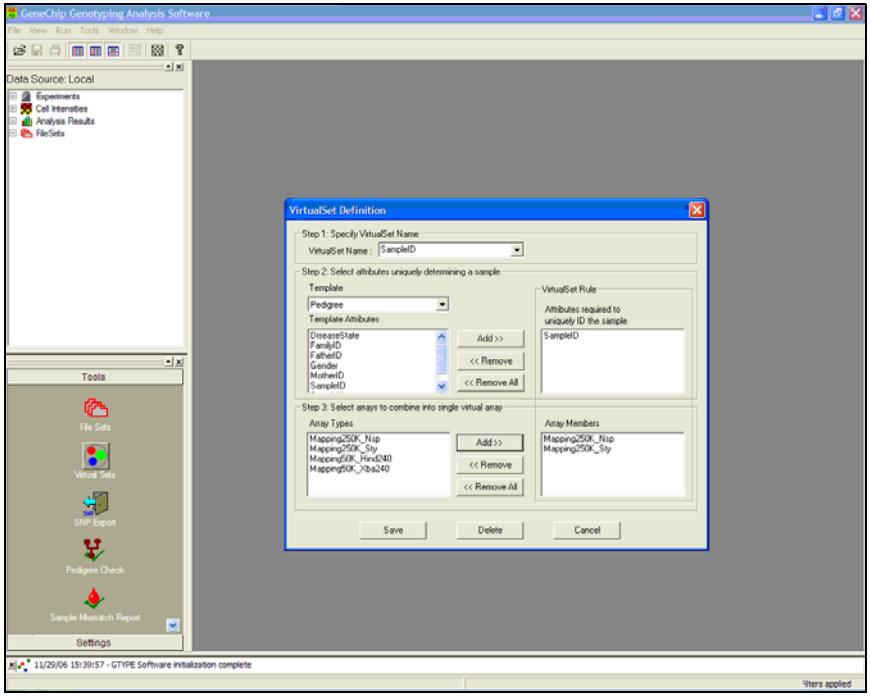

Figure C.6 Virtual Set Definition window

INDEX

L

M

N

O

P

Q

R

S

T

U

V

W

Y