Guide to Your CGH Data

NimbleGen Comparative Genomic Hybridization (CGH) Services

Outline

This document provides detailed information on how to view and interpret your CGH data. Roche NimbleGen calculates ratio data using test sample as the numerator and reference sample as the denominator, regardless of which Cy dye is specified for each sample in the online sample submission form (eForm).

Enclosed with this document you will find a data disk (refer to the "Data Disk Contents" section on page 6 for more information on the disk's directories and files)

Step 1. Installing the Necessary Software

❶ Download and install a free, fully functional demo version of SignalMap v1.9 Software, if necessary, at *www.nimblegen.com/products/software/.*

The demo version functions for 30 days after installation. To purchase the software for continued use after the 30-day trial period ends, contact your local sales affiliate at *www.nimblegen.com/arraysupport.*

SignalMap software is currently available only in Windows PC format.

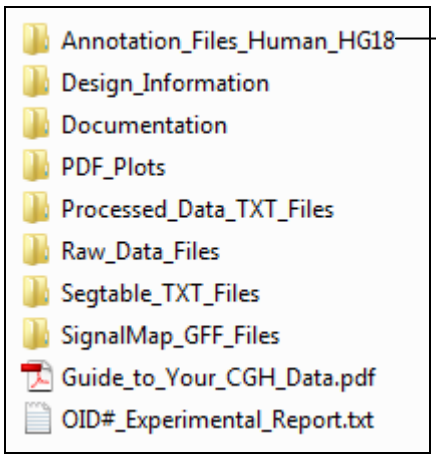
Follow the installation wizard instructions.

❷ Install other recommended software, if necessary:

- Adobe Acrobat Reader (*www.adobe.com*)
- Spreadsheet software (e.g. Microsoft Excel, *www.microsoft.com*)

Step 2. Reviewing Data Disk Contents and the Experimental Report

1. Insert the data disk into the appropriate drive. Figure 1 shows the files and directories on the data disk:



This directory is included only for array designs based on the human hg18 build. Annotation files for the human hg19 build are included for array designs based on the human hg19 build. These annotation files are also available for download at *www.nimblegen.com/products/cgh*.

Figure 1: Folders and Files Provided on a CGH Data Disk

2. Open and review the OID#_Experimental_Report.txt file for a detailed summary of your experiment.

Step 3. Reviewing Your Data

Both raw and processed data files are included on the data disk:

- The Raw_Data_Files directory contains the raw data (.pair files), which you can open using a text editor.
- The following directories contain processed data files. Note that the directory name identifies the format of the files within the directory: PDF, GFF, or TXT.

Directory	Contents
PDF_Plots	Whole-genome and individual-chromosome views of your data in PDF (portable document format).
SignalMap_GFF_Files	Whole-genome and individual-chromosome views of your data in GFF (general feature format) for interactive viewing using SignalMap software.
Segtable_TXT_Files	Summary of predicted segments in TXT (text) format.
Processed_Data_TXT_Files	Log_2 ratio (test divided by reference) data of all data points in TXT (text) format.

Reviewing PDF Plots

1. Open Acrobat Reader.
2. Select **File** -> **Open** and open .pdf plots in the PDF_Plots directory (hold down Ctrl and click to select multiple files). Be aware of the following when selecting .pdf plots:
    - Roche NimbleGen provides .pdf plots in two formats:

- Single panel rainbow plots show the $\log_2$ ratio of test divided by reference for all probes plotted versus genomic position for all chromosomes or regions in a single plot, with each chromosome or region differentiated by vertical dashed lines.
- Multi-panel plots show the data on a chromosome-by-chromosome (or region-by-region) basis.

- Roche NimbleGen provides .pdf plots of both unaveraged (1X) and 10X window-averaged data. The "Step 4. Learn More about Your Data" section on page 5 provides detailed information about window averaging.
- Roche NimbleGen recommends viewing the 10X window-averaged .pdf files for an initial snapshot overview of your data. Because of the high density of data points, the unaveraged .pdf files are often not as informative as the 10X window-averaged .pdf files.

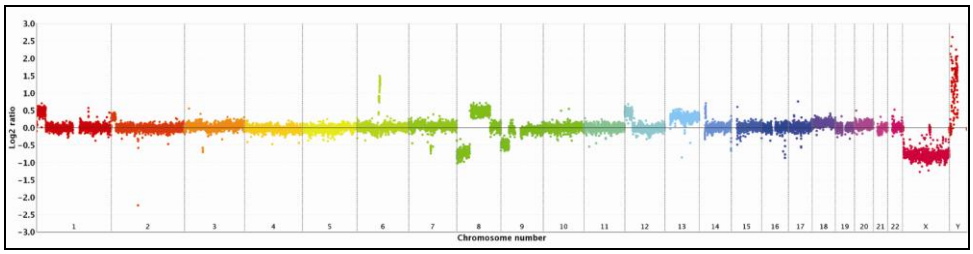


Figure 2: Human CGH Whole-Genome Array Data Displayed as a Single Panel Rainbow Plot

Reviewing GFF Files

1. Open SignalMap software.
2. Select **File** -> **New** and import .gff files from the SignalMap_GFF_Files directory (hold down Ctrl and click to select multiple files). Be aware of the following when selecting .gff files:
   - Roche NimbleGen provides .gff files containing normalized (no segmentation algorithm applied), unaveraged (1X), and 10X window-averaged data. The "Step 4. Learn More about Your Data" section on page 5 provides detailed information about normalization and window averaging.
   - Roche NimbleGen recommends initially viewing the unaveraged .gff files for a comprehensive and highest-resolution view of your data.
   - A gene annotation .gff file specific to your design is also included in the SignalMap_GFF_Files directory.

   You can drag and drop files into the SignalMap window.
3. (Optional) Select **File** -> **Import** and import the *design*_probe_locations.gff file from the Design_Information directory. This file displays the genomic location (x-axis) and uniqueness value (y-axis) of all probes on the array.
4. (Optional - for human hg18 or hg19 designs only) Select **File** -> **Import** and import files from the Annotation_Files_Human_HG18 directory or the Annotation_Files_Human_HG19 directory (hold down Ctrl and click to select multiple

files). HG18-19_Annotation_Files_Descriptions.pdf provides detailed information about the human hg18 and hg19 annotation files.

The complete suite of human hg18 and hg19 annotation files is also available for download at *www.nimblegen.com/products/cgh*.

5. Review your data using these SignalMap functions:

   - **Select chromosomes:** Use the pane selector field below the toolbar to change the view to either all chromosomes (All Tracks) or a selected chromosome (e.g. chr2).
   - **Set the $\log_2$ ratio scale:** To compare multiple .gff tracks, set the $\log_2$ ratio scale so it is the same for all selected tracks:
     a. Select **Edit** -> **Select All** (hold down the Ctrl key and click the y-axis to select individual tracks).
     b. Select **View** -> **Manual Scale**. Enter the desired minimum and maximum scale values (Roche NimbleGen recommends starting with a scale from -2 to 2) and click **OK**.
   - **Set track height:** Select the desired .gff tracks as described above, select **Track** -> **Set Height**, enter the desired track height (Roche NimbleGen recommends 120), and click **OK**.

Appendix A describes additional SignalMap functions you can use when reviewing .gff files.

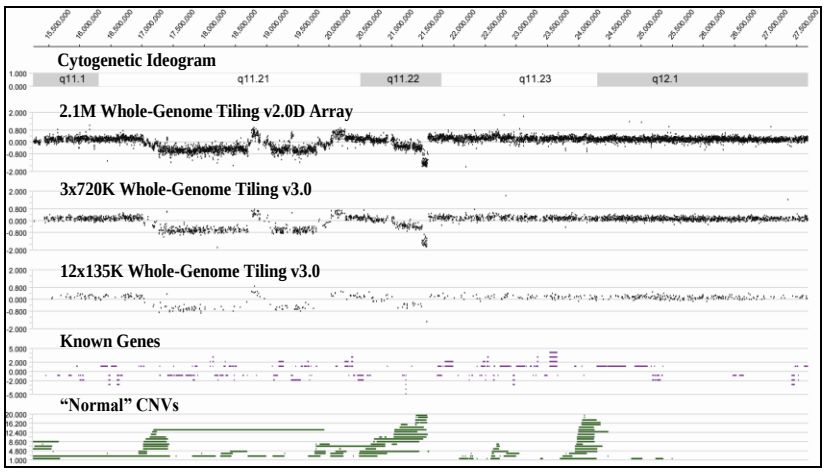


Figure 3: Cross-Platform Analysis of a Large (~4Mb) Deletion Region in Chromosome 22 in a VCFS Research Sample Referenced Against Normal Genomic DNA. A deletion region associated with Velocardiofacial Syndrome (VCFS) is detected using three different NimbleGen CGH Whole-Genome Tiling arrays, as indicated. Copy number analysis was performed using the segMNT algorithm, available in NimbleScan software. Data are displayed using a GFF file in SignalMap software alongside annotation tracks (provided with NimbleGen CGH arrays) showing a cytogenetic ideogram, known genes, and "normal" CNVs from the Database of Genomic Variants *(http://projects.tcag.ca/variation)*.

Reviewing Segtable and Processed Data Files

1. Open spreadsheet software (e.g. Excel) or a text editor (e.g. WordPad).

   Spreadsheet software allows you to sort segments by genomic position, size, or $\log_2$ ratio mean.

❷ Select **File** -> **Open** and open segtable .txt files from the Segtable_TXT_Files directory (hold down Ctrl and click to select multiple files). Be aware of the following when selecting segtable.txt files:

Roche NimbleGen provides segtable files that list the predicted segments and their corresponding genomic position (by chromosome and base pair coordinates), size, $\log_2$ ratio, and number of data points contained within the segment. Segtable files are provided for both unaveraged (1X) and 10X window-averaged data. The "Step 4. Learn More about Your Data" section on page 5 below provides detailed information about normalization and window averaging.

❸ *Optional:* Select **File** -> **Open** and open processed data .txt files from the Processed_Data_TXT_Files directory (hold down Ctrl and click to select multiple files). These files provide a comprehensive summary of your data in .txt format.

Step 4. Learning More about Your Data

Signal Intensity (Raw) Data

Signal intensity data are extracted from the scanned images of each array using Roche NimbleGen NimbleScan software. Signal intensities for each feature are provided in pair files (*arrayID*_532.pair and *arrayID*_635.pair), the raw data format for CGH experiments. Pair files can be viewed using a text editor and are stored in the Raw_Data_Files directory.

Spatial Correction

Spatial correction is applied to the raw signal intensities to reduce variation that occurs along the length, and to a lesser extent, across the width of the array. Specifically, locally weighted polynomial regression (loess) is used to adjust the signal intensity of each feature based on the X and Y coordinate position of the probe on the array.

Roche NimbleGen has found that spatial correction reduces some artifacts observed in CGH data and has minimal impact on overall noise and $\log_2$ ratio values in regions of copy number variation. The spatially corrected data are reported in two columns, *EXP_SPATIAL* and *REF_SPATIAL*, in the *arrayID*_segMNT.txt file in the Processed_Data_TXT_Files directory.

Normalization

Normalization compensates for inherent differences in signal between Cy3 and Cy5 dyes. After combining the signal intensity information with the genomic coordinate information, and applying spatial correction, the Cy3 and Cy5 signal intensities are normalized to one another using qspline normalization (Workman C, et al., 2002). The normalized data are reported in two columns, *EXP_NORM* and *REF_NORM*, in the *arrayID*_segMNT.txt file in the Processed_Data_TXT_Files directory. In addition, the normalized data can be viewed in SignalMap software using the *arrayID*_segMNT.gff file in the SignalMap_GFF_Files directory.

Window Averaging

Following normalization, a window-averaging step is applied. Window averaging involves performing a set of transformations on the signal data for all features within a given window and reducing the data to a single data point. For example, when a 10X averaging window is

applied to a whole-genome array in which probes are spaced every 1,000bp across the genome, each window-averaged data point represents the average of approximately 10 original data points contained within a 10,000bp window. Your data disk contains unaveraged (1X) and 10X window-averaged data reported in PDF, GFF, and TXT formats. Window averaging reduces the size of the data set and may reduce noise in the data, but at the expense of resolution.

CGH-segMNT Analysis

CGH-segMNT analysis identifies copy number changes using a dynamic programming process that minimizes the squared error relative to the segment means. Roche NimbleGen has found that the segMNT algorithm consistently shows increased accuracy and performance when compared with other available copy number calling algorithms, such as the DNACopy algorithm.

Data Disk Contents

Directory	File	Description
Annotation_Files_Human_HG18 or Annotation_Files_Human_HG19	Genes.gff[1] Genes_Exon-Intron.gff[1] Transcription_Start_Sites.gff[1] Structural_Variants.gff[1] 42M_CNV_Regions.gff[1] NimbleGen_CNV_Regions.gff[1] Segmental_Duplications.gff[1] Cytogenetic_Ideogram.gff[1] miRNA.gff[1] HG18-19_Annotation_Files_Descriptions.pdf[3]	HG18-19_Annotation_Files_Descriptions.pdf provides detailed information about these human hg18 or hg19 files.
Design_information	DesignNotes.txt[2, 4]	A short description of the design.
	design.ndf[2]	Complete information about the design, including feature locations and probe sequences.
	design.pos[2]	The genomic positions of the probes used in the experiment.
	*design*_probe_locations.gff[1]	The genomic location (x-axis) and uniqueness value (y-axis) of all probes on the array.
	design.ncd[2, 4]	The location of arrays on the slide.
Documentation	NimbleGen_Data_Formats_*version*.pdf[3]	Detailed descriptions of NimbleGen data files.
PDF_Plots	*arrayID*_unavg_single_panel_rainbow_segMNT.pdf[3]	A genome-wide view of your unaveraged data.
	*arrayID_XX*bp_single_panel_rainbow_segMNT.pdf[3]	A genome-wide view of your 10X window-averaged data. *XX*bp indicates the effective base pair (bp) spacing after window averaging.
	*arrayID*_unavg_multi_panel _segMNT.pdf[3]	A per chromosome (or per region) view of your unaveraged data.
	*arrayID_XX*bp_multi_panel _segMNT.pdf[3]	A per chromosome (or per region) view of your 10X window-averaged data. *XX*bp indicates the effective base pair (bp) spacing after window averaging.

Directory	File	Description
Processed_Data_TXT_Files	*arrayID*_segMNT.txt[2]	The log_2 intensity values for each feature are reported for the 532nm channel, the 635 nm channel, and the ratio of the test divided by reference channels. These values are reported for the raw data, spatially corrected data, and normalized data in table format.
	*arrayID*_unavg_segMNT.txt[2, 4]	The test divided by reference log_2 ratio values for your unaveraged data in table format.
	*arrayID*_ *XX*bp_avg_segMNT.txt[2, 4]	The test divided by reference log_2 ratio values for your 10X window-averaged data in table format. *XX*bp indicates the effective base pair (bp) spacing after window averaging.
Raw_Data_Files	*arrayID*_532.pair[2]	A raw data file that reports signal intensities for features in the 532nm channel.
	*array*ID_635.pair[2]	A raw data file that reports signal intensities for features in the 635nm channel.
Segtable_TXT_Files	*arrayID*_unavg_segtable_segMNT.txt[2, 4]	Summary of predicted segments in your unaveraged data.
	*arrayID_XX*bp_segtable_segMNT.txt[2, 4]	Summary of predicted segments in your 10X window-averaged data. *XXbp* indicates the effective base pair (bp) spacing after window averaging.
SignalMap_GFF_Files	*design*.gff[1]	Gene annotation specific to your design.
	*arrayID*_segMNT.gff[1]	The test divided by reference log_2 ratio values for your normalized data.
	*arrayID*_unavg_segMNT.gff[1]	The test divided by reference log_2 ratio values for your unaveraged data.
	*arrayID_XX*bp_avg_segMNT.gff[1]	The test divided by reference log_2 ratio values for your 10X window-averaged data. *XXbp* indicates the effective base pair (bp) spacing after window averaging.
root directory	OID#_Experimental_Report.txt[2, 4]	A detailed summary of the experiment.
	Guide_to_Your_CGH_Data_*version*.pdf[3]	(This document) Instructions on how to view and analyze your data.

1 Open using SignalMap software.
2 Open using a text editor, such as Microsoft WordPad.
3 Open using Adobe Acrobat Reader.
4 Open using spreadsheet software, such as Microsoft Excel.

Technical Support

If you have questions, contact your local Roche Microarray Technical Support. Go to *www.nimblegen.com/arraysupport* for contact information.

Appendix A. Additional Techniques for Reviewing GFF Files

The following SignalMap functions are also helpful when reviewing .gff files:

- **Zoom**: Select the magnifier button on the toolbar. Position the magnifier cursor to the region of interest then click and drag to draw a bounding box around the region to magnify. Alternatively, hold down the Ctrl key and press + to zoom in or - to zoom out.
- **Arrange tracks**: To move a data track above or below another track, click in the left margin of the track, move the cursor until a gray dashed line appears, and then click to place the track in the new position.
- **Search for genes**: To search for a particular gene by gene name or accession number, select **Edit** -> **Search**. Type the gene name or accession number in the search field and click **Find**. To jump to the data track where the gene is located, click **Go to Selected**.
- **Pointer information**: To gather information about a specific data segment or annotation feature, click the pointer button on the toolbar and position the cursor over the region of interest. For data segments, the $\log_2$ ratio and genomic position will be displayed in the top left corner of the SignalMap window. For annotation features, details including gene name, cytoband coordinates, and CNV reference information from the Database of Genomic Variants will be displayed.
- **Attach Cursor**: To move quickly from one data point or annotation feature to the next, click the y-axis of the track of interest and select **Cursor** -> **Attach Cursor**. A vertical line will appear at the left-most feature of the track. To jump to the next feature, use the left and right arrows. To remove the cursor from the track, select **Cursor** -> **Detach Cursor**. This function is particularly useful when there are large gaps between data points or annotation features.

05434432001 ⓔ 11/11

Published by
Roche NimbleGen, Inc.
504 S. Rosa Rd
Madison, WI 53719 USA

www.nimblegen.com/arraysupport