NEXUS EXPRESS

FOR ONCOSCAN DATA ANALYSIS

2013

www.biodiscovery.com

COPYRIGHT NOTICE

The Nexus Express™ Users Manual was written at BioDiscovery, Inc., 5155 Rosecrans Ave, Suite 310, Hawthorne, CA 90250.

Printed in the United States of America. The software described in this book is furnished under a license agreement and may be used only in accordance with the terms of the agreement.

Every effort has been made to ensure the accuracy of this manual. However, BioDiscovery and Affymetrix make no warranties with respect to this documentation and disclaim any implied warranties of merchantability and fitness for a particular purpose. BioDiscovery shall not be liable for any errors or for any incidental or consequential damages in connection with the furnishing, performance, or use of this manual or the examples herein. The information within this manual is subject to change.

Document Revision: 1

TRADEMARKS

Nexus Copy Number ™, Nexus CGH™, GeneSight®, ImaGene®, GeneSight-Lite®, GenePie®, GeneDirector®, and CloneTracker® are registered trademarks of BioDiscovery, Inc.

OncoScan™ is a registered trademark of Affymetrix, Inc.

Windows, WordPad, and Excel are either registered trademarks or trademarks of Microsoft Corporation in the United States and/or other countries.

Other product names mentioned in this manual may be trademarks or registered trademarks of their respective companies and are the sole property of their respective manufacturers.

PATENTS

Several patents are pending.

SUPPORT FOR 3RD PARTY SOFTWARE

BioDiscovery makes every effort to support the latest technologies and provide solutions that the user needs, which often includes adding the ability to use third party software or utilizing output from third party software, but we cannot directly support third party (open source) software or another vendor's products. If the problem is within our software in utilizing third party software, our support team can assist with that but we cannot provide support for installation or execution of third party software. We use the information provided by the vendor/open source software at the time of development, but we cannot be responsible if something changes in the third party software which may then not work in Nexus. In these cases, we try to work with the vendor and come up with a solution as soon as possible, often issuing a minor release.

QUICK START GUIDE

See **Nexus Express Quick Start Guide.pdf** in the installation folder or access via **Help->Quick Start Guide** within the application.

NEXUS INTRODUCTION

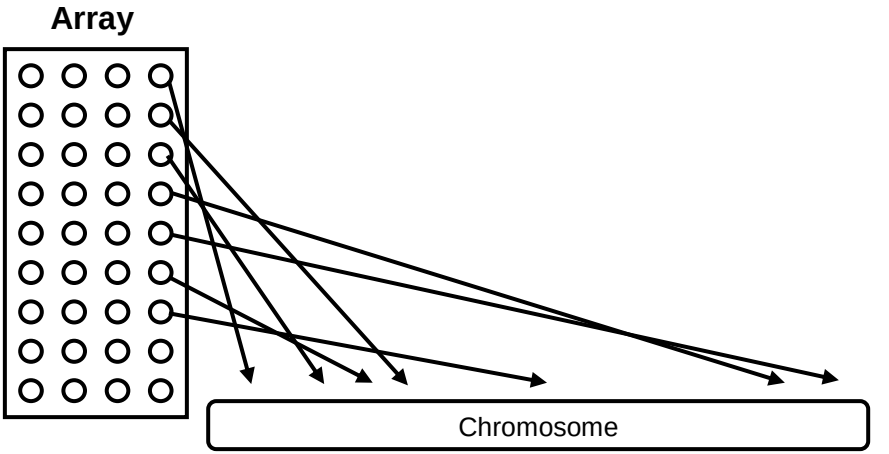
Nexus Express was designed to be a versatile and easy to use yet powerful application for analysis of SNP microarray data for copy number and somatic mutation analysis. The elegant user interface enables scientists to explore data and get to biological answers quickly. The flexible design enables viewing and analysis of large numbers of samples simultaneously.

Nexus Express reads in OncoScan FFPE Assay Kit data analyzed by the TuScan algorithm from OSCHP files. Nexus Express supports two levels of analysis. The first is on a per sample basis to identify regions of gain, loss, loss of heterozygosity (LOH), and the presence of somatic mutations in a single sample. The next level combines results from all samples to answer more elaborate questions at the population level about the copy number and LOH aberrations. For example, are there regions of copy number gains or losses associated with a particular group of samples compared to another *(e.g.*, tumor vs. normal).

SINGLE SAMPLE ANALYSIS- ONCOSCAN FFPE ASSAY KIT

Once Nexus Express has read in the data from the OSCHP files, it is ready to visualize copy number change on each sample. Each probe on an array maps to a location on a genome and this mapping information is available to the software along with the probe intensities of your experiment samples.

Probes map to locations along the genome:



Segmentation algorithms try to differentiate between adjacent clusters of probes to determine if any copy number change is apparent. There are many algorithms that perform segmentation. Nexus Express takes advantage of two different algorithms. This first is the TuScan (Tumor Scan) algorithm developed by Affymetrix, and is the default segmentation for the OSCHP files generated by the OncoScan FFPE Assay Kit assay. The second is BioDiscovery's SNP-FASST2 algorithm, which is a proprietary variation of a hidden Markov model (HMM), optimized for performance. The algorithm is discussed in more detail in a subsequent section on *SNP-FASST2 Segmentation Algorithm,* page 24. OncoScan version 2 data is always analyzed on SNPFASST2 in Nexus Express.

MULTIPLE SAMPLE ANALYSIS

Nexus' capabilities are more than segmentation. After the single sample analysis, you can explore your data on the population level by performing statistical class comparisons.

DEFINITIONS

Some important terms have special meaning in the context of this application and are defined here.

PROJECT

The concept of a project in Nexus is a collection of samples that the user would like to process and then collectively integrate their results. The user can add new samples to a Project and delete existing samples already contained in a Project.

AGGREGATE

It is a segmented region of gain or loss shared by a set of samples. The number of samples sharing the event is referred to as the frequency of the aberration in the region.

FACTOR

A characteristic of a sample used in grouping together all samples with a particular characteristic to compare different groups. A factor is often clinical data or phenotypic data. For example, samples in a Project can have factors **Age**, **Gender**, and **Tumor Grade**. Values for **Age** could be **Old** or **Young**. Values for **Gender** would be **Male** or **Female**. Values for **Tumor Grade** could be **1**, **2**, or **3**. Then an aggregate plot can be created for all samples that are **Male** and one for all samples that are **Female** and these two profiles can then be compared.

SAMPLE DESCRIPTOR

A text document in a specific format used to describe the samples in a **Project**, the location of the input data files, and any **Factors** associated with each sample.

DATA TYPE

A term used to describe the input data to Nexus. The Nexus Express Software can be used to analyze OncoScan version 2 and OncoScan FFPE Assay Kit data only.

EXTERNAL DATA

In addition to copy number and somatic mutations data that Nexus Express analyzes, external data such as expression and miRNA can be integrated with the copy number data and viewed alongside the copy number data.

INTERFACE OVERVIEW

Nexus Express visualization and analysis tools and results are organized into various tabs and sub-tabs within the main application window. When Nexus is launched, the **Project Workspace** tab is the only visible tab and allows you to open an existing project or create a new project. After a **Project** has been opened, the **Data Set** is visible and, the **Comparisons, External Data,** and **Nexus DB** tabs are present. Once the **View** button is pressed to process and view the results in the genome browser, the **Results** tab appears which is composed of a number of sub-tabs containing the results in graphical as well as table format. At the bottom of the **Project Workspace** screen is a clickable link to a **Quick Start guide,** the **User Manual**, and the **Customer Forum** where you can get the latest information on new developments with Nexus Express Copy Number as well as communicate with other users.

DATA SET TAB

The **Data Set** tab allows loading of data into a Project, contains information about each sample in the data set, and allows processing of selected samples.

RESULTS TAB

After samples have been processed, the results are displayed in various sub-tabs within the **Results** tab. These sub-tabs contain graphical as well as numerical output.

EXTERNAL DATA TAB

This tab allows loading of external data such as expression and miRNA into a project. Each type of external data has a sub tab within this tab.

COMPARISONS TAB

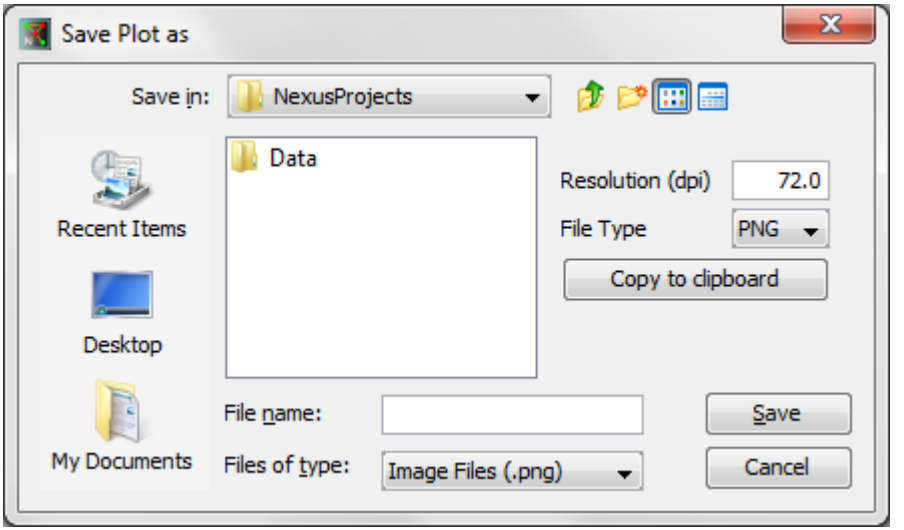
This tab allows creating comparisons between different sub-groups of samples in a project.

STATUS BAR

At the bottom of the main application window is the status bar which provides feedback to the user whenever Nexus is processing data. It provides a message indicating the type of processing being performed and usually states **Finished** when the processing is complete. All the way towards the right of this bar is a cancel button which can be clicked to cancel the current processing event. The mouse cursor turns into an hourglass icon when the application is processing data.

SAVING HIGH RESOLUTION IMAGES

In many places in the Nexus Express interface you are able to save an image either by using the save image icon  or via a **Save** button. When you do save an image, the **Save** dialog opens up asking you to specify a location for the file or you can choose to copy this to the clipboard by clicking the **copy to clipboard** button. Image saving options (resolution and file type) can be modified from the Save Dialog. The default dpi is set to 72 but can be increased via the **Image Saving Options**. When the **Save** button is clicked, Nexus adds the sample name and chromosome (if the image is a specific chromosome) to the bottom of the image.



DATA TABLE FUNCTIONS

EXPORT TXT BUTTON

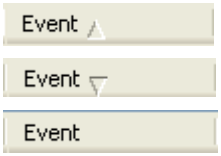
All tabs that contain numerical data in a table also have an **Export TXT** button to export the data into a text file. Upon clicking the button, a **Save** window opens up asking the user to specify the file name and location. The data is saved as a tab delimited text file. In the Data Set tab, this button is under the multi-level **Tools** button (**Tools->Export TXT**).

FORMATTING DATA TABLES

Whenever data is represented within a table in the Nexus tabs in the main window and within pop-up windows, the table can be formatted in several ways as described below.

SORTING COLUMNS

Columns can be sorted in ascending or descending order by clicking on the column headers. Clicking repeatedly on the header will go through a cycle of sorting in ascending order, descending order, and not ordered. An arrow next to the header name indicates that the data is sorted. If no arrow is present, data is not sorted.



Multi-column sorting can be performed as well by selecting the primary, secondary, tertiary, etc. sort columns. This is accomplished by holding down the CTRL key while clicking on the column headers of the columns you want to sort. The first column that is clicked on while holding the CTRL key becomes the primary sort column and the next column clicked becomes the secondary sort column and so on. The sequence of the multi-column sort can be determined quickly by glancing at the size of the arrows in the column headers. The primary sort column will display the largest arrow with the arrow size decreasing with each successive column sorted in a multi-column sort.

By looking at the size of the arrows in the header in the figure below, we can determine that the multi-column sort sequence is **Gender**, then **General Tumor Classification**, and finally **Specific Tumor Classification**.

Sample	Data Type	Status	Gender	General Tumor Classifica...	Specific Tumor Classification
AA3	Genepix	Processed	Male	Astrocytic tumors	Anaplastic astrocytoma
GBM1	Genepix	Processed	Male	Astrocytic tumors	Glioblastoma multiforme
GBM12	Genepix	Processed	Male	Astrocytic tumors	Glioblastoma multiforme
GBM16	Genepix	Processed	Male	Astrocytic tumors	Glioblastoma multiforme
GBM20	Genepix	Processed	Male	Astrocytic tumors	Glioblastoma multiforme
GBM24	Genepix	Processed	Male	Astrocytic tumors	Glioblastoma multiforme
GBM26	Genepix	Processed	Male	Astrocytic tumors	Glioblastoma multiforme
GBM3	Genepix	Processed	Male	Astrocytic tumors	Glioblastoma multiforme
GBM32	Genepix	Processed	Male	Astrocytic tumors	Glioblastoma multiforme
GS1	Genepix	Processed	Male	Astrocytic tumors	Glioblastoma multiforme
GNN1	Genepix	Processed	Male	Astrocytic tumors	Glioneuronal neoplasm
AOA3	Genepix	Processed	Male	Mixed gliomas	Anaplastic oligoastrocytoma
AOA5	Genepix	Processed	Male	Mixed gliomas	Anaplastic oligoastrocytoma
AOA7	Genepix	Processed	Male	Mixed gliomas	Anaplastic oligoastrocytoma
AO2	Genepix	Processed	Male	Oligodendroglial tumors	Anaplastic oligodendroglioma
AO3	Genepix	Processed	Male	Oligodendroglial tumors	Anaplastic oligodendroglioma

REARRANGING COLUMNS

The order of the columns can be changed by clicking on a column header and holding while dragging the mouse to where you want to move the column and then letting go.

Re-sizing columns

Column widths can be changed by clicking on the gray vertical bar marking the column boundary in the header row and holding and dragging the mouse left or right to decrease or increase the column width.

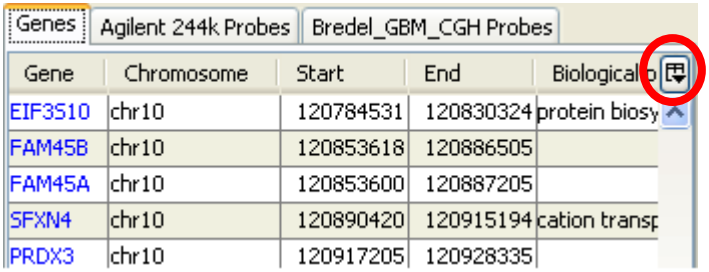
Highlighting rows

Clicking on any cell in the table will highlight the corresponding row. To highlight multiple rows, click on a table row and drag the mouse, or click and use shift, or click and use the **Ctrl** key to choose specific rows. The exception to this is the table in the **Data Set** tab where clicking on a cell in the checkbox column does not highlight the row, rather it either selects or de-selects the checkbox.

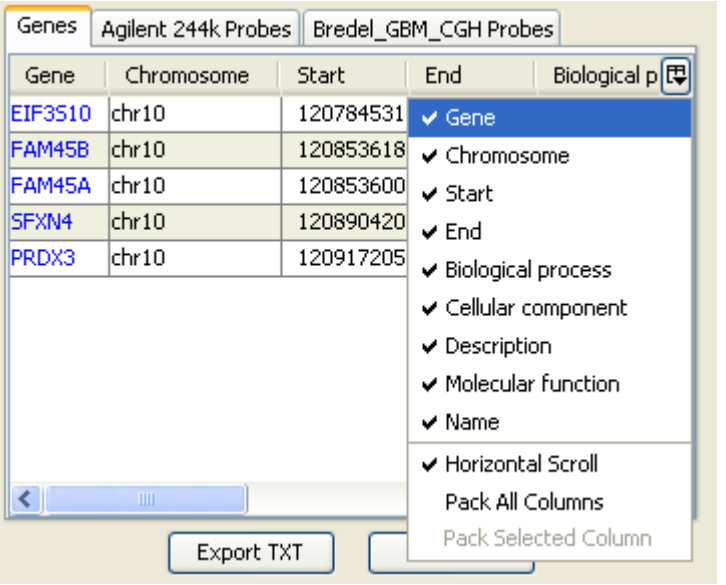
Column Control Button

Within some data tables, (e.g., annotation data for regions), a **Column Control** button can be found to the right of the last column header.

Column control button:



Clicking the **Column Control** button opens up a popup menu with options for configuring the table display.

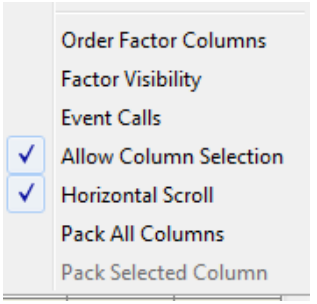


Each column header is listed in the menu. A check next to the header name indicates that the column will be displayed in the table. To hide any columns, just click on the header name in the menu to remove the checkmark and the column will now be hidden from view. By default, all columns are displayed. Additional options:

Horizontal scroll – When selected, this function resizes each column to its optimal width and allows columns to scroll off to the right. A scroll bar appears on the bottom. De-selecting this resizes the columns such that all columns fit in the visible area.

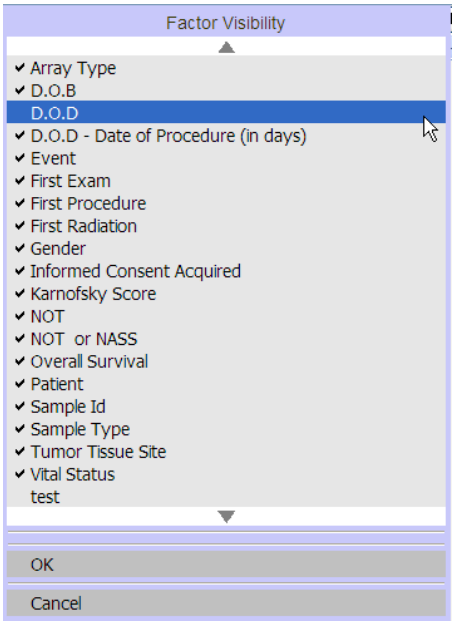
Pack all columns – Resizes all columns to their default width.

In the **Data Set** table, the column control button has additional options (**Order Factor Columns**, **Allow Factor Selection**, and **Factor Visibility**).



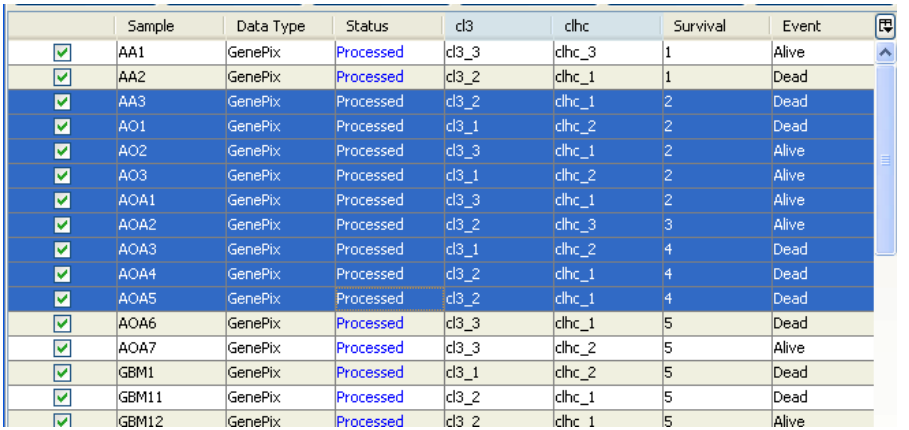
Order Factor Columns – Selecting this function puts all the Factor columns in alphabetical order.

Factor Visibility – Selecting this brings up a list of factors from which you can select the Factors you want visible in the table. Click on the Factor name to check or uncheck and then click **OK**.



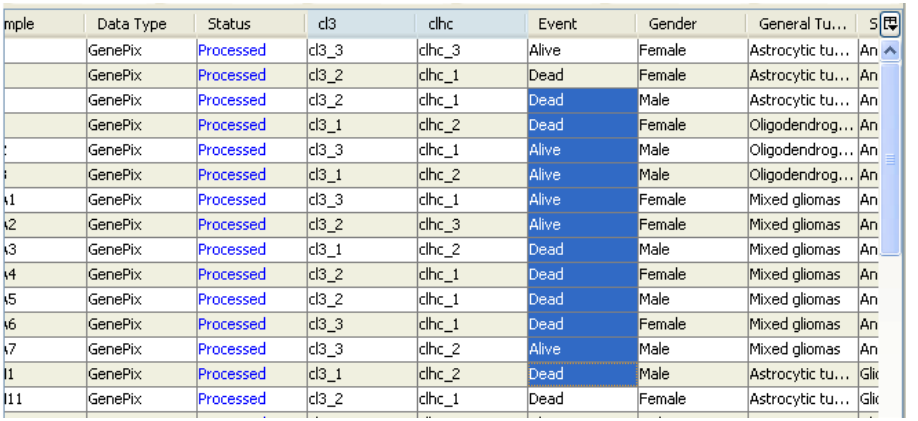
Event Calls – Selecting this bring up a list of event call columns (e.g. one copy gain, one copy loss, etc.) from which you can select those you want to be visible in the table. Click on the call type to check or uncheck and then click **OK**.

Allow Column Selection – Selecting this item will allow you to highlight the rows in individual columns. Normally when highlighting samples the rows in all columns are highlighted as seen in the figure below.



	Sample	Data Type	Status	cl3	clhc	Survival	Event
☑	AA1	GenePix	Processed	cl3_3	clhc_3	1	Alive
☑	AA2	GenePix	Processed	cl3_2	clhc_1	1	Dead
☑	AA3	GenePix	Processed	cl3_2	clhc_1	2	Dead
☑	AO1	GenePix	Processed	cl3_1	clhc_2	2	Dead
☑	AO2	GenePix	Processed	cl3_3	clhc_1	2	Alive
☑	AO3	GenePix	Processed	cl3_1	clhc_2	2	Alive
☑	AOA1	GenePix	Processed	cl3_3	clhc_1	2	Alive
☑	AOA2	GenePix	Processed	cl3_2	clhc_3	3	Alive
☑	AOA3	GenePix	Processed	cl3_1	clhc_2	4	Dead
☑	AOA4	GenePix	Processed	cl3_2	clhc_1	4	Dead
☑	AOA5	GenePix	Processed	cl3_2	clhc_1	4	Dead
☑	AOA6	GenePix	Processed	cl3_3	clhc_1	5	Dead
☑	AOA7	GenePix	Processed	cl3_3	clhc_2	5	Alive
☑	GBM1	GenePix	Processed	cl3_1	clhc_2	5	Dead
☑	GBM11	GenePix	Processed	cl3_2	clhc_1	5	Dead
☑	GBM12	GenePix	Processed	cl3_2	clhc_1	5	Alive

With **Allow Column Selection** turned on, rows in individual columns can be selected (see figure below). This is useful when copying and pasting a column of values from a spreadsheet into Nexus or vice versa.



mple	Data Type	Status	cl3	clhc	Event	Gender	General Tu...	S
	GenePix	Processed	cl3_3	clhc_3	Alive	Female	Astrocytic tu...	An
	GenePix	Processed	cl3_2	clhc_1	Dead	Female	Astrocytic tu...	An
	GenePix	Processed	cl3_2	clhc_1	Dead	Male	Astrocytic tu...	An
	GenePix	Processed	cl3_1	clhc_2	Dead	Female	Oligodendrog...	An
	GenePix	Processed	cl3_3	clhc_1	Alive	Male	Oligodendrog...	An
	GenePix	Processed	cl3_1	clhc_2	Alive	Male	Oligodendrog...	An
A1	GenePix	Processed	cl3_3	clhc_1	Alive	Female	Mixed gliomas	An
A2	GenePix	Processed	cl3_2	clhc_3	Alive	Female	Mixed gliomas	An
A3	GenePix	Processed	cl3_1	clhc_2	Dead	Male	Mixed gliomas	An
A4	GenePix	Processed	cl3_2	clhc_1	Dead	Female	Mixed gliomas	An
A5	GenePix	Processed	cl3_2	clhc_1	Dead	Male	Mixed gliomas	An
A6	GenePix	Processed	cl3_3	clhc_1	Dead	Female	Mixed gliomas	An
A7	GenePix	Processed	cl3_3	clhc_2	Alive	Male	Mixed gliomas	An
1	GenePix	Processed	cl3_1	clhc_2	Dead	Male	Astrocytic tu...	Gli
11	GenePix	Processed	cl3_2	clhc_1	Dead	Female	Astrocytic tu...	Gli

Scroll bars

Sometimes horizontal and vertical scroll bars will be present in the data tables and can be used to move the table horizontally or vertically to view the data.

TUSCAN ALGORITHM (SUPPORTED BY AFFYMETRIX)

The TuScan algorithm uses B-allele frequencies (BAFs) and log2 ratios to estimate the ploidy and percentage of aberrant cells in the sample (%AC) which in turn are used to calculate copy number calls (CN). The BAFs and log2 ratios contribute equally to CN determination. TuScan first uses the BAFs and log2 ratio data to identify segments of equal CN. Next TuScan uses the BAFs, log2ratios and segment data to find the combination of %AC and ploidy that best fits the data. When TuScan can successfully determine %AC, the algorithm assigns each aberrant segment an integer copy number representing the copy number in the tumor portion of the sample. This is possible because CN is well approximated by an integer when the tumor is nearly homogeneous. If the tumor is highly heterogeneous (*i.e.,* lacks a dominant clone), or contains a large amount of "normal" cells, the %AC cannot be determined. In other words, if the percentage of aberrant cells contributing to the various aberrations in the sample varies across all aberrations, %AC and ploidy cannot be determined. When %AC cannot be determined, the segmentation algorithm will still identify segments of equal CN, but the CN in just the aberrant cells cannot be determined. In this case, TuScan bins the copy numbers and returns fractional CN values in 1/3 increments (*e.g.,* 2, 2.33, 2.66, 3 etc.). This fractional copy number is derived from the normal contamination as well as the heterogeneous population of tumor cells; therefore, the fractional CN calls represent the average CN observed for that segment. Users should look at the value of %AC to determine whether the CN value represents the CN in the tumor (%AC= number) or the average CN in the sample (%AC=NA). Tumor heterogeneity also affects the interpretation of the CN number calls when %AC cannot be determined. For example, a TuScan call of 2.33 can result from 40% of the aberrant cells having 3 copies, 10% of aberrant cells having 5 copies, or a more complex heterogeneous mixture of copy numbers. Since nearly every tumor sample will have some amount of normal

contamination combined with tumor heterogeneity it is not possible to predict how often TuScan will be able to determine the %AC, it will vary depending on the sample.

Copy number calls are classified into “Gain”, “Loss”, “High Copy Gain”, and “Homozygous Copy Loss” based on their CN call from TuScan. In the settings section it is possible to adjust the threshold for High Copy Gain

SNP-FASST2 Segmentation Algorithm

Fast Adaptive States Segmentation Technique (FASST2) and SNP-FASST2 algorithms were developed to address the needs of increased density of array technology in the adaptation of new high-throughput sequencing technology. Although a number of algorithms have been proposed based on the well know Hidden Markov Model (HMM) approach which have linear time requirements, these methods often rely on rather restrictive assumptions that are not satisfied in common types of real world samples (e.g. cancer data which often contains significant mosaicism and normal cell contamination). On the other hand recursive segmentation methods, such as CBS and Rank Segmentation, do not require such restrictive assumption and have performed well in comparison studies, but have at least quadratic time performance. The FASST2 approach achieves a balance between these previous methods by using an HMM model (not to estimate the copy number or allelic event states) but rather to quantitate a large number of possible segment levels that might fall between the expected states. Subsequent processing is performed to combine these basic segments into copy number and allelic event calls.

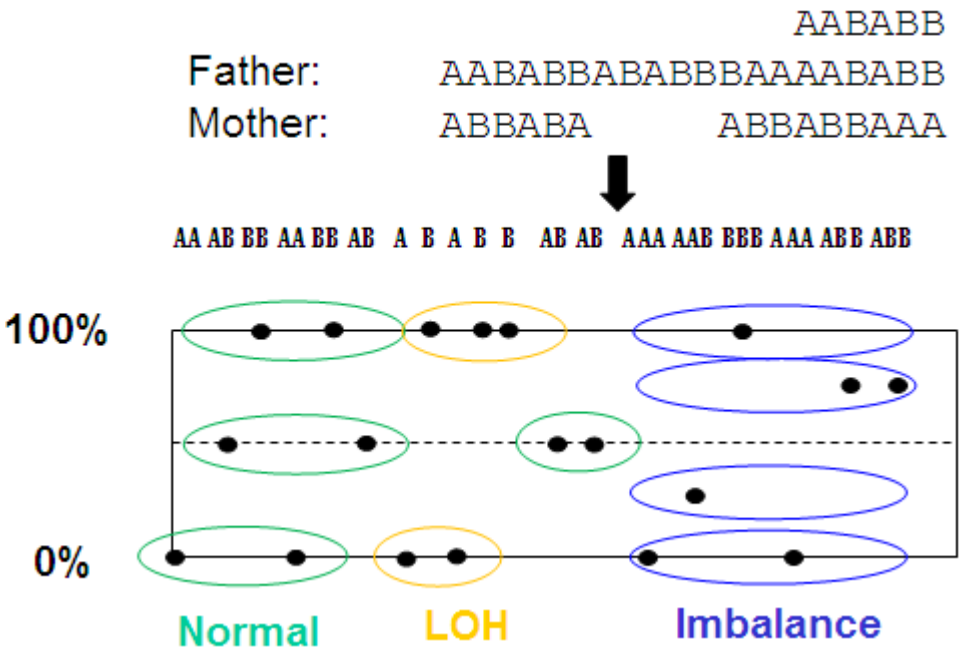
Allele-Specific Computation

B-allele frequency data from SNP arrays is used to provide allele specific results such as LOH, imbalance, or no change. The **SNP-FASST2** algorithm makes segmentation calls based on both the log ratio and B-allele frequency data.

B ALLELE FREQUENCY BANDS

With the B Allele Frequency values we can determine if there is a loss of heterozygosity (LOH) or an allelic imbalance. Let's take a look and see what this means. In the diagram below, we have alleles from both the father and mother. The 0% and 100% refer to the B Allele Frequency. If we have one allele contribution from each parent, we have a normal genotype (circled in green). These are normal genotypes for which we see three bands of probes on the plot. The areas below where the allele contribution was only from one parent (in this case the father as there is a deletion on the mother's chromosome), we have an LOH. These will show two bands on the plot (at 0% and 100%). The areas circled in blue are where we have one allele from each parent as well as an extra allele which causes an allelic imbalance. In the example below, there is an amplification of a segment of the allele from the father. Such areas will typically show four bands on the plot.

The first point in the graph is at the 0% because the genotype is AA so therefore the B Allele frequency is 0%. The next point is an AB and sits on the 50% line because the B Allele frequency is not 100% B but a mix of A and B. In the last part of the graph we have three alleles and therefore the clusters are around the 100% and 0% if there are 3 copies of the same allele. If all three copies are not of the same allele then we also see clusters around the 33% and 66% marks.

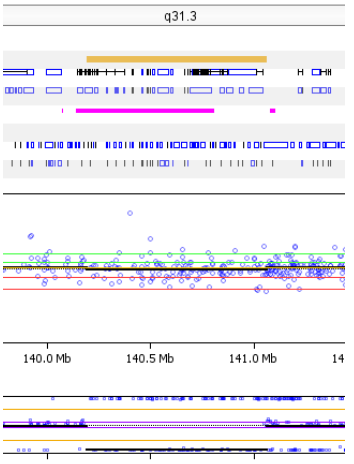


Using this information Nexus makes calls based on both copy number and the allelic ratios. In the figure below we have copy number results vertically and a second dimension of allelic events along the horizontal has been added. . The Xs in the table show what we would typically expect to see. E.g. a one or high copy gain which would cause an allelic imbalance. Sometimes, with a high copy gain, the amplification of one allele can be so high as to dwarf out the presence of the other allele and therefore it may look like an LOH. A homozygous loss would show a total allelic loss and a single copy loss would show as an LOH. In some cases, one can also see a copy neutral LOH where there is no copy number gain or loss but an LOH is seen. This can happen if one receives two copies of a chromosome from one parent, therefore preserving the copy number even though there's LOH.

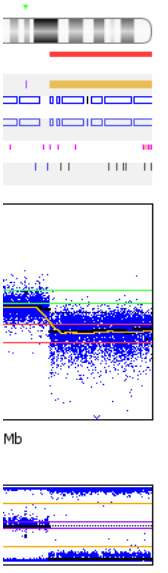
Typical results combining copy number with allelic events (remaining cells could have X's but is not so typical):

		Allelic Loss		Allelic Imbalance	Normal
		LOH	Total Allelic Loss	Allelic Imbalance	
Gain	High Copy Gain	X		X	
	One Copy Gain			X	
Loss	One Copy Loss	X			
	Homozygous Loss		X		
Normal		X			

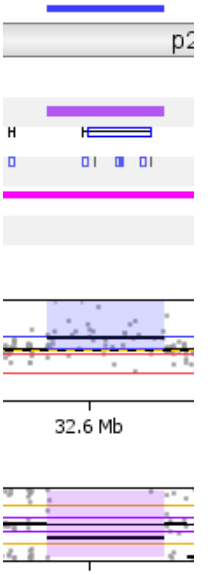
Here are some examples of what can be seen in the B Allele Frequency plots:



Copy neutral LOH



LOH with one copy loss



Allelic imbalance with high copy gain

NEXUS EXPRESS PROJECT

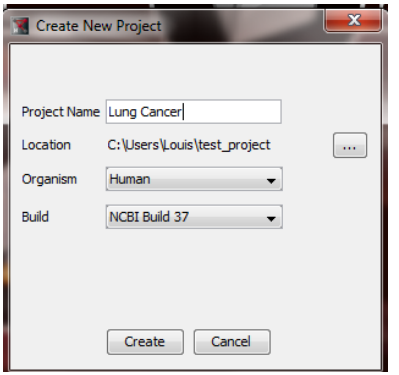
All analyses using Nexus Express are organized into Projects. The concept of a Nexus Express **Project** is explained in the section on *Nexus Express Basics* (page 13). Only one Project can be open at a time within Nexus Express. If a Project is currently open in Nexus Express and a new project is opened via **File->New Project** or **File->Open Project** then the currently open Project will close to open up the newly selected Project.

Upon starting Nexus Express, there are simply two options to process and look at your data: **Create A New Project** or **Open Existing Project**.

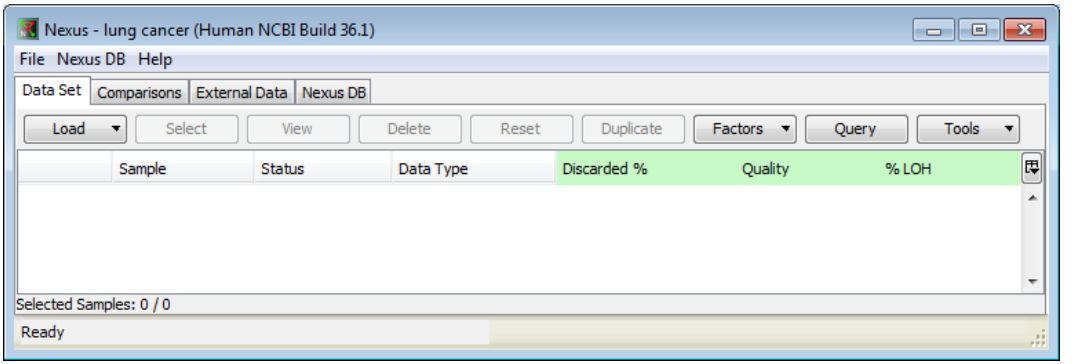
CREATING A NEW PROJECT

Clicking on the **Create a New Project** icon opens up the **Create new Project** window. Alternatively, select **File->New Project** from the menu bar to create a Project. Enter a

name in the **Project Name** text box. Click  to specify a location where the project data will be stored. Next, select the appropriate organism from the **Organism** drop down menu. Once the organism has been selected, the next field is populated with the appropriate genome builds available for that organism. Choose the build appropriate for your samples.

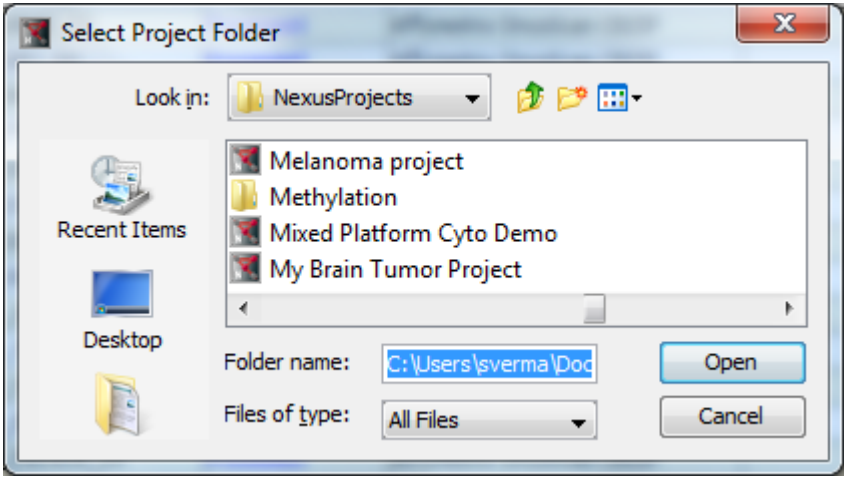


Once the Project is created, the following screen will be displayed in Nexus. Since this is a new Project and data hasn't yet been loaded into it, the **Data Set** tab is empty. The process of loading data is covered in the next section.



OPENING AN EXISTING PROJECT

To open a Project that you have already created, click **Open Existing Project** from the main screen. Alternatively, select **File->Open Project** from the menu bar. The **Select Project Folder** window opens up. Navigate to the location of the Project folder, select the Project folder, and click **Open**. The Projects are indicated by the Nexus Copy Number icon.



Data can be loaded into a new project that contains no data or into an established project that already has some samples.

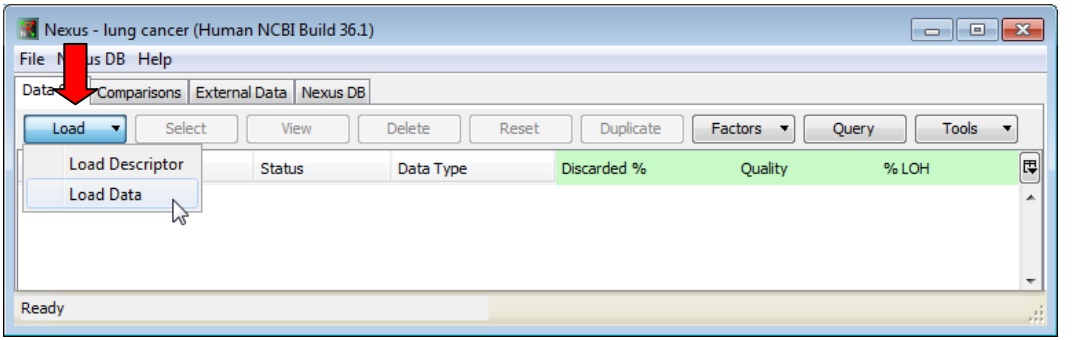
There are two ways to load data:

1. DIRECT DATA LOADING - Loading data directly via the file chooser
2. DATA LOADING VIA A SAMPLE DESCRIPTOR - Creating and loading a sample descriptor (a tab delimited text file listing sample names, where to find the data files, and associated clinical data).

Method 1, in combination with use of the **Load Factors** button (discussed later), is the simplest and preferred way for loading data and clinical attributes. In rare cases the sample descriptor method may be needed; for example, if loading replicate data, performing matched paired analysis, or using batch operations via the command line, you must use the sample descriptor method.

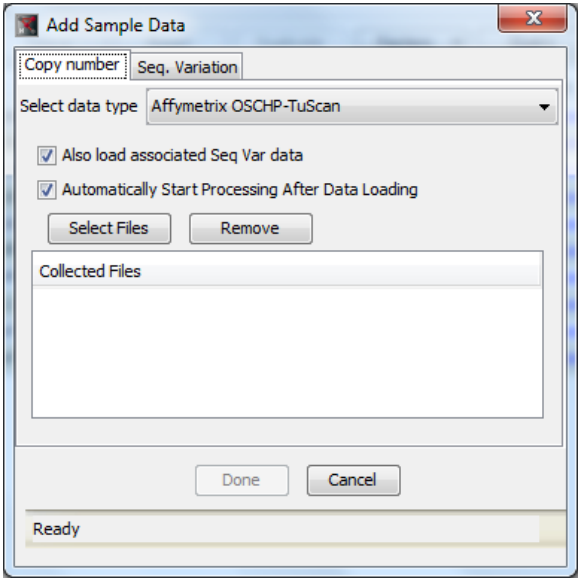
DIRECT DATA LOADING

In the **Data Set** tab, click the **Load** button and select **Load Data**.



LOADING ONCOSCAN ARRAY DATA (OSCHP FILES)

The **Add Sample Data** dialog looks like the following figure when loading OSCHP file and processing with the TuScan algorithm:

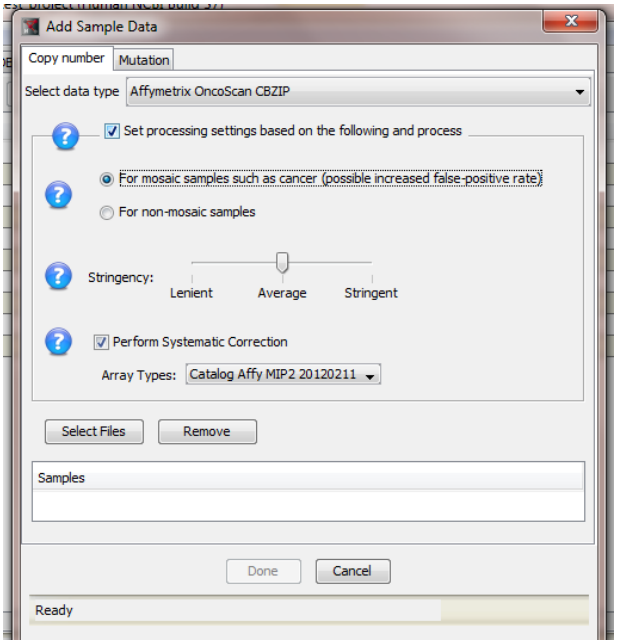


To load the somatic mutation data in addition to the copy number data select the check box “Also load associated Seq Var Data”. Note, no corrections should be applied to OSCHP file data, as the TuScan algorithm normally handles this for the user.

If you would like to process the data using SNP-FASST2 instead of TuScan, select **Affymetrix OncoScan-SNP-FASST2** in the Data Type dropdown. The resulting options are the same as that shown in the figure above.

LOADING ONCOSCAN V2 DATA (CBZIP FILES)

The **Add Sample Data** dialog looks like that in the following figure for OncoScan V2 data.



If you do not want to use the guide, uncheck “Set default processing settings based on the following”. When this is unchecked, all the options are grayed out and data loading proceeds without use of this guide:

To use the guide, select Affymetrix OncoScan CBZIP from the “Type” dropdown menu at the top and check off **Set processing settings based on the following and process**. Using the guide sets the appropriate settings for the Data Type you are loading. Adjustments are made to the “Analysis” settings in the Settings window as well as “Systematic Correction” if you choose to use systematic correction. The **Help** buttons on the side give further information on how to choose among the options presented.

Next select one of the radio buttons depending on the type of data you are loading: **For mosaic samples such as cancer (possible increased false-positive rate)** or **For non-mosaic samples**. The latter option is generally used for constitutive samples. The selection here adjusts the thresholds used to make calls (High Gain, Gain, Loss, and Big Loss).

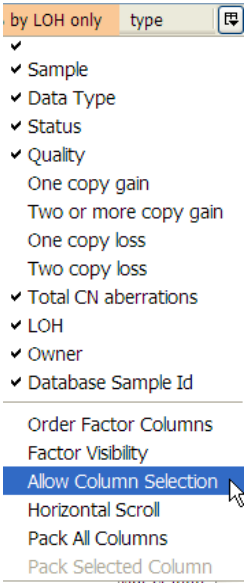
Next, adjust the **Stringency** slider to one of 3 positions: Lenient, Average, or Stringent. This adjusts the sensitivity of the algorithm in making calls. “Lenient” will generate more calls and “Stringent”, less calls (more specificity). It adjusts the **Significance Threshold** value in the Settings.

Next decide if you want to perform systematic correction (please see the section on *Systematic Correction*, page 201, for more information on what this setting does). This performs a “GC wave correction” to the data. Since the correction is based on probes used, the specific array is needed. It is strongly recommended that you perform a GC correction OncoScan v2 data.

Click on **Select Files** to add your data files. Once you click **Done** the appropriate settings will be applied to the Data Type you selected and the data will be processed automatically (you won’t need to click the **View** button in the Data Set tab to process samples). You can see the Settings parameters used by opening the **Settings** window. After reviewing your results, you may need to go back and further refine your settings for your particular data. You can always see exactly what settings parameters were used for each sample by clicking on the blue **Processed** text in the **Status** column of the Data Set table.

ADDING FACTORS

After your data has been loaded, if you need to add any factors you can do so using the **Factors** button and selecting either **Add Factor** button or **Load Factors**. **Add Factor** will prompt for a factor name and then will add a new column with the specified factor name. Then you can either manually enter values into each cell, enter one value into several cells at once, or copy and paste columns from a spreadsheet. If you are copying and pasting, make sure **Allow Column Selection** is selected in the menu displayed using the Column Control button  (see figure below). With the **Allow Column Selection** on, you can assign the same value to several cells without having to type it into each cell by highlighting some cells and then right-clicking and selecting **Set Value** which brings up a text box where the value can be typed in.



Selecting **Load Factors** prompts for a tab delimited file containing sample names and factors that can be loaded into the project (see section on *Loading Multiple Factors*, page 51, for the specific format of the tab delimited text file). Data from this table can be easily exported in a tab delimited text file using the **Export TXT** button.

Factor values can be numerical or textual and can even be a hyperlink. If a factor value is entered as a URL (e.g. http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM215420), after it is entered, the displayed hyperlink is "External Link". In the example below, URLs were entered into the GEO column pointing to the GEO site but the displayed link in blue is termed "External Link".

General Tumor Classification	GEO
Astrocytic tumors	External Link
Astrocytic tumors	External Link
Astrocytic tumors	
Astrocytic tumors	

The values can be edited by right-clicking in the cell with **Allow Column Selection** on and selecting **Set Value** or **Clear Value**.

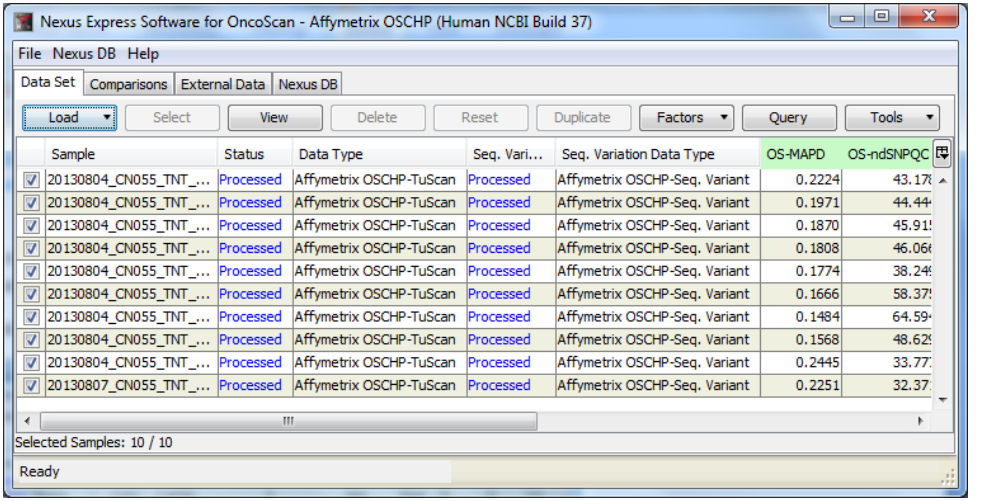
DATA LOADING VIA A SAMPLE DESCRIPTOR

If you performing paired analysis, you must create a sample descriptor to load data and you cannot use the direct data loading method. In some cases you may already have a spreadsheet containing file names, sample names, and clinical attributes and you can use this to create your sample descriptor. See the *Sample Descriptor* section, page 258, on how to create a sample descriptor. After your sample descriptor is ready, click on the **Load** button in the **Data Set** tab and select **Load Descriptor** to open the **Select Sample Descriptor** window. Navigate to the location of the Sample Descriptor file and select it. The Sample Descriptor file need not be located in the folder containing the input data files. It is usually convenient to locate it with the input data files since the path to the input file locations can be relative to the Sample Descriptor file and need not be prefixed by the file path, making it easier to specify file names such as GSM226862-787.txt instead of C:\Melanoma-OncoScan\GSM226862-787.txt. Sample Descriptor formats are detailed in the next section and *Appendix A: Data Type Format and Sample Descriptor Templates*, page 261.

DATA COLUMNS

Once data is loaded a minimum of four columns will be displayed (checkbox column, **Sample**, **Data Type**, and **Status**). Additional columns may be present depending on what other columns are present in the Sample Descriptor files (e.g. Factor columns) and on the columns that are selected to be displayed via the column control button on the top right of the table.

Data Set tab showing some processed samples:



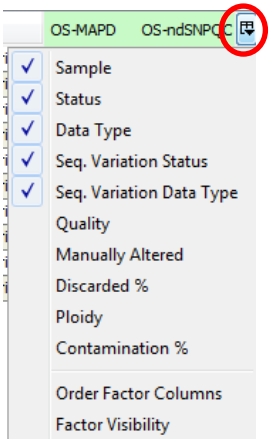
DATA SET

The **Data Set** Tab consists of several buttons used for loading, organizing, and processing the data within a project.

It also contains a table listing all the samples that are in this Project and associated clinical information (biological factors) about the samples.

Sample	Status	Data Type	Seq. Vari...	Seq. Variation Data Type	OS-MAPD	OS-ndSNPQC
20130804_CN055_TNT_...	Processed	Affymetrix OSCHP-TuScan	Processed	Affymetrix OSCHP-Seq. Variant	0.2224	43.178
20130804_CN055_TNT_...	Processed	Affymetrix OSCHP-TuScan	Processed	Affymetrix OSCHP-Seq. Variant	0.1971	44.44
20130804_CN055_TNT_...	Processed	Affymetrix OSCHP-TuScan	Processed	Affymetrix OSCHP-Seq. Variant	0.1870	45.91
20130804_CN055_TNT_...	Processed	Affymetrix OSCHP-TuScan	Processed	Affymetrix OSCHP-Seq. Variant	0.1808	46.06
20130804_CN055_TNT_...	Processed	Affymetrix OSCHP-TuScan	Processed	Affymetrix OSCHP-Seq. Variant	0.1774	38.24
20130804_CN055_TNT_...	Processed	Affymetrix OSCHP-TuScan	Processed	Affymetrix OSCHP-Seq. Variant	0.1666	58.37
20130804_CN055_TNT_...	Processed	Affymetrix OSCHP-TuScan	Processed	Affymetrix OSCHP-Seq. Variant	0.1484	64.59
20130804_CN055_TNT_...	Processed	Affymetrix OSCHP-TuScan	Processed	Affymetrix OSCHP-Seq. Variant	0.1568	48.62
20130804_CN055_TNT_...	Processed	Affymetrix OSCHP-TuScan	Processed	Affymetrix OSCHP-Seq. Variant	0.2445	33.77
20130807_CN055_TNT_...	Processed	Affymetrix OSCHP-TuScan	Processed	Affymetrix OSCHP-Seq. Variant	0.2251	32.37

There are a set of basic columns that will be present with each data set. Columns can be hidden or displayed by selecting them from the menu accessed via the column control button on the top right of the table:



Columns displaying quality or copy number statistics will have a green background for the header.

Columns in the Data Set table:

Column	Description
Checkbox	To select a subset of samples in the project to use in order to view in the browser and for further analyses.
Sample	The name provided to the sample (name of the data file loaded if using Load Button or as specified in the Sample Descriptor file if using the Load Descriptor button).
Data Type	Specified when loading data either in the Load Samples dialog or in the Sample Descriptor file. Indicates the type of data based on the input file. This can be from a specific array platform or particular quantification software. See the Sample Descriptor section for more information on data types.
Status	Values are **Processed** or **Unprocessed**. Indicates whether copy number computation has been performed on this sample. Clicking on **Processed** will bring up the **Settings** window to show what settings were used in processing this sample. The settings cannot be edited here.
Total CN Aberrations	Total number of copy number aberrations in this sample

Column	Description
OS-MAPD	MAPD is a global measure of the variation of all microarray probes across the genome. It represents the median of the distribution of changes in log2 ratio between adjacent probes. Since it measures differences between adjacent probes, it is a measure of short-range noise in the microarray data. Lower MAPD values are better. MAPD values above 0.3 are out of bounds.
OS-ndSNPQC	The metric, ndSNPQC is a measure of how well genotype alleles are resolved in the microarray data. Larger ndSNPQC values are better.
OS-CelPairCheckStatus	CelPairCheck is a test that inspects each pair of intensity (*.cel) files to determine whether the files have been properly paired and assigned to the correct channel. In addition to accidental mispairing of intensity files while setting up the analysis, a tracking problem during the assay may result in a sample being assigned to the wrong GeneChip array. As a result CelPairCheck ignores file names, and instead inspects the genotypes in the two intensity files to detect file mispairings. If the CelPairCheck Status is not *Pass*, then you should pair your sets of intensity files differently and repeat the analysis to see if the problem disappears.

Column	Description
OS-ndwavinessSD	ndWaviness-SD is a global measure of variation of microarray probes that is insensitive to short-range variation and focuses on long-range variation. As for all OncoScan QC metrics, ndWsd is computed on normal diploid markers. ndWavinessSD should be used along with LowDiploidFlag, ndCount (the actual number of diploid markers identified) BAFs and log2 ratio to assess if the log2 ratio is centered correctly. ndWavinessSD can thus help assess if log2 ratios need to be recentered. In addition when ndWavinessSD is high, the log2 ratios should be examined for clear breakpoints as opposed to a gradual drift of the log2 ratio. When the latter is observed small aberrations should be examined carefully. When breakpoints are sharp and the ndCount is large a high ndWavinessSD can be ignored.

Column	Description
OS-%AberrCells	Algorithmic estimation of the percent of aberrant cells in a sample. When this is "NA" it means that the % aberrant cells could not be estimated because the percent is either too low or the sample is heterogeneous and hence is composed of several types of aberrated cells or the percent varies from one genomic location to another. When %AC is not NA, % e.g. "60%" this means that across all aberration ~60% of the cells were aberrated and contributed to the elevated (or reduced) Copy number.
OS-Ploidy	Ploidy of tumor when determined or not available
OS-low Diploid Flag	Flag indicating there are not enough normal diploid markers identified in the sample. Signals were not calibrated by normal diploid markers and therefore algorithmic re-centering was not performed.
LOH	For SNP arrays only. Total number of LOH aberrations in this sample
%LOH	For SNP arrays, displays % LOH call across the genome, excluding the sex chromosomes (total length of LOH regions/total length of autosomal chromosomes). Good measure of consanguinity of the sample.

Column	Description
Discarded %	Count of the % of probes that are eliminated based on a flag specified in the Settings
One copy gain	The number of one copy gain events in this sample. Column header will have a green background.
Two or more copy gain	The number of two or more copy gain events in this sample. Column header will have a green background.
One copy loss	The number of one copy loss events in this sample. Column header will have a green background.
Two copy loss	The number of two copy loss events in this sample. Column header will have a green background.
Manually Altered	Values “YES” or “NO”. Field is automatically filled if a call on this sample was manually altered (e.g. addition of a new call by the user via the chromosome view in the sample drill down window)
Optional Factor columns	If any Factor columns are defined in the Factors file or Sample Descriptor file, they will be added to the table.

Data in this table can be exported into a tab delimited text file using the **Tools->Export TXT** button.

FORMATTING THE TABLE

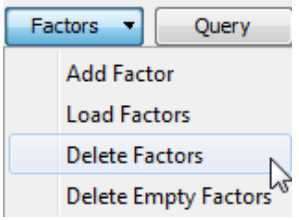
The columns in the table can be sorted and rearranged and the rows can be highlighted as described in the section on *Nexus Express Basics* (page 16). In addition, other functionality, described below, is available to this table.

EDITING CELL VALUES

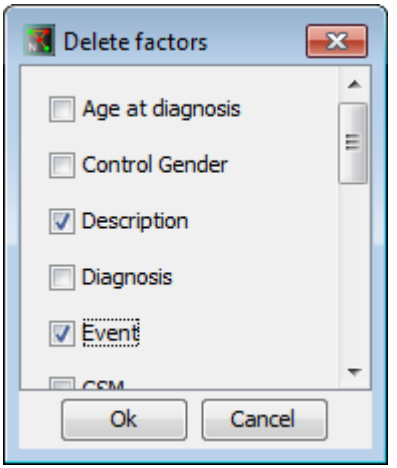
Some columns allow you to edit values by double clicking in the cell. This allows you to change Factor values without re-loading the factors file and without running the calculations again. Rather than manually editing individual cells, values can be copied from a spreadsheet column and pasted into selected cells in the **Data Set** tab. Also, with **Allow Column Selection** on (via the column control button), multiple cells can be selected and then right clicking and selecting **Set Value** brings up a text box in which a value can be typed. Now all the selected columns will have this new value. This feature can be used to enter values into empty cells or to replace existing values.

DELETING COLUMNS

Certain columns, mainly the Factor columns can be deleted. Right clicking on the column header brings up a menu with a **Delete** option. Select this to delete the column. A **Confirm** dialog will pop up asking you if you really want to delete this column. You can also delete multiple columns by selecting **Delete Factors** from the Factors button.



Simply mark off the checkboxes of the columns you want to remove and click **OK**.



You can select **Delete Empty Factors** to delete any factor columns that don't have a value in even a single cell.

RENAMING COLUMNS

Certain columns, mainly the Factor columns, can have their headers changed. Right clicking on the column header brings up a menu with a **Rename** option. Select this to rename the column. An **Input** dialog will pop up asking you to specify the column name in the input box.

SELECTING SAMPLES USING THE CHECKBOX

All samples contained in a Project are listed in the table within the **Data Set** tab. All or a subset of these samples can be selected via the checkbox for viewing and further analyses. Checking off the checkboxes allows the user to view these selected samples in the **Results** tab and perform any further analysis on these samples only. The checkbox can be selected or de-selected by clicking in the checkbox cell. To select multiple samples at once, highlight the samples first and then click on the **Select** button to check off the checkboxes. Please note that highlighting multiple samples and then clicking on **Select** will select the highlighted items but it will de-select any items that were

previously selected via the checkbox and were not highlighted when the **Select** button was pressed (see figure below).

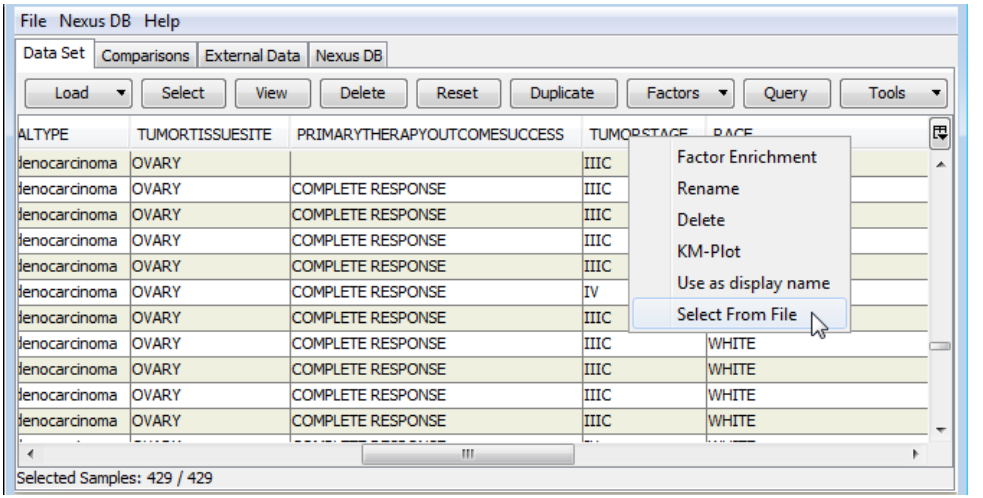
Initially AA3, AO1, and AO2 are selected with the checkboxes

The first four samples are highlighted and the Select button is pressed

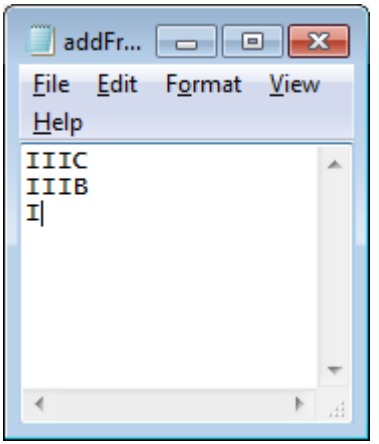
Now the first four samples are selected via the checkboxes but sample AO2 that was selected previously is no longer selected.

The column can also be sorted by samples that are checked off or not by left-clicking on the column header. At least one sample needs to be selected via the checkbox for the **View** button to be active.

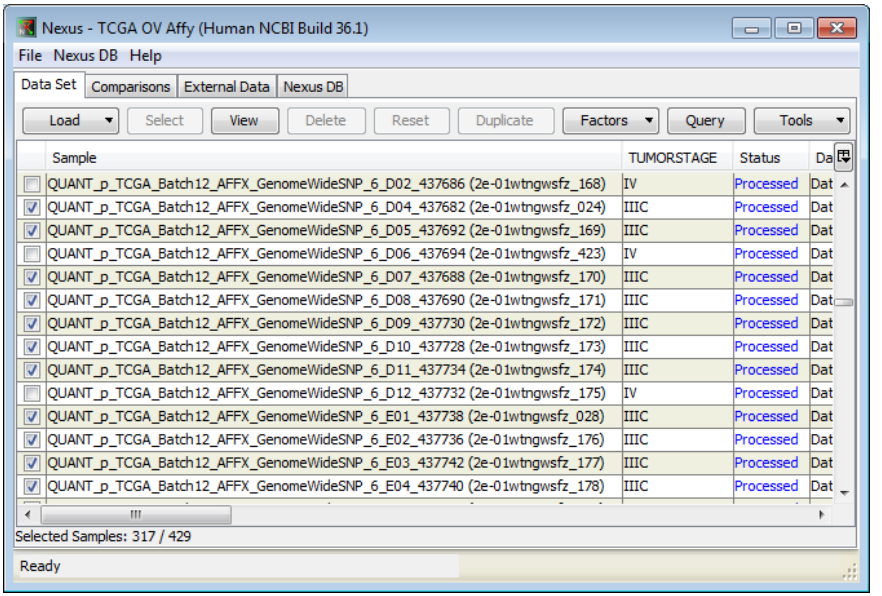
Samples can also be selected based on values in an external text file. The values need to be in one column. Right click on a column header and choose **Select From File**. A dialog box asks you to select a file. This file must contain the values in one column. If a value in the file is found in the chosen data set column, that sample will be selected. If any values in the text file are not found in the data set, then an alert will state which values were not found. This feature is very handy when the project contains hundreds of samples and selecting individually one by one or even by sorting is not feasible. In the example below, we want to select samples with certain tumor-stage values. Right clicking on the TUMORSTAGE column and selecting **Select From File** bring up a dialog box to select the file containing the values.



The file we select looks like this:



After selecting the file, if there are any values in the file that are not present in the factor column, a warning dialog box indicates which values were not found in the column. After the selection is performed, only the samples containing the indicated TUMORSTAGE are selected:



Deleting Samples

Samples can be deleted from the project by highlighting the appropriate rows and clicking on the **Delete** button.

Resetting Processing

A great feature of Nexus Express is that different settings can be used to process different samples of the same data type within a project. Once samples have been processed, the **Status** column has the value **Processed**. To reset (clear processing), simply highlight the rows for the samples that you want to reset and click on the **Reset** button. The value in the **Status** column for these will now indicate **Unprocessed**. To reprocess these, select **File->Settings** and make your changes here. Now when you click on the **View** button to process and see the results, these unprocessed samples will be processed using this new setting. So if you had a few samples with a high background,

you can correct for this by choosing to do a local background correction, for example, for these samples only. Note that Nexus Express uses “smart” processing. If only the threshold values have changed, the data is not completely reprocessed so the computation is very fast. In some cases the original input data files are needed to re-process samples. If the file is not found in the location where it was originally uploaded from, an alert will state that the sample cannot be reprocessed. In such cases, and the sample won’t be reset.

Duplicating a Sample

If you would like to process a particular sample using different algorithms, the **Duplicate** button will help perform this function. Highlighting one or more sample rows and clicking on the Duplicate button, adds a sample which is a copy of the selected sample. The name given to the duplicate sample has “copy" appended to the end of the sample followed by the sample name. The duplicated sample is **Unprocessed** and you can adjust processing settings, to select a different segmentation algorithm, for example, before processing the sample.

Viewing Results

Clicking on the **View** button allows you to see the results of the processed data in the **Results** tab. If samples have already been processed, clicking on the **View** button will lead you to the **Results** tab where the frequency data will be displayed against the genome browser. If some or none of the samples have been processed yet, then clicking on the **View** button will first process all the unprocessed samples and then bring forward the **Results** tab. Please note that only samples which have the checkbox marked will be displayed in the **Results** tab. The section above, *Selecting samples using the checkbox*, describes how to select samples using the checkbox column.

Adding Factors Individually

After the initial loading of data into a Nexus Express project and the Factors file, you can add additional factors from directly within the table. Click on the **Factors** button, select **Add Factor** and an Input Dialog will open asking you to give the factor a name. Once the factor name has been entered and the **OK** button is clicked, the new column shows up at the end of the table. You can either enter values manually by clicking on each cell and typing or you can copy and paste a column from a spreadsheet. You can also edit a group of cells at one time by highlighting some cells in one column and right clicking and selecting **Set Value.** In the resulting dialog, enter the factor value. Now all the selected cells have this value.

Loading Multiple Factors At Once

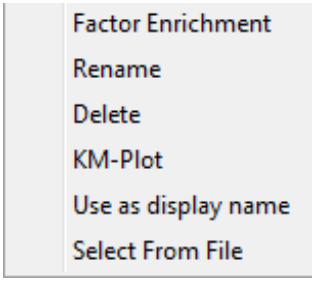
Selecting **Load Factors** via the **Factors** button can be used to load a tab delimited text file containing sample names and any number of factor columns. A column called **Sample Name** or **Sample** containing names matching the sample names in the project is required. If both a column called **Sample Name** and a column called **Sample** are present, then the **Sample** column will be used to identify the samples and the **Sample Name** column will be added as a factor. The Factor column headers do not have to be prefixed with **Factor:**. If the Data Set tab already contains a column with the same header as that in the Factors file being loaded and the Data Set tab contains at least one value in that column, then an alert box will inform that this column already exists and will ask if data should be overwritten. If the Factor column exists in the Data Set tab but is empty, then there will be no alert prompt and values will be filled in from the Factors file.

A tab delimited factors file opened in MS Excel:

	A	B	C	D	E	F
1	Sample	Gender	General Tumor Classification	Specific Tumor Classification		
2	AA1	Female	Astrocytic tumors	Anaplastic astrocytoma		
3	AA2	Female	Astrocytic tumors	Anaplastic astrocytoma		
4	AA3	Male	Astrocytic tumors	Anaplastic astrocytoma		
5	AO1	Female	Oligodendroglial tumors	Anaplastic oligodendroglioma		
6	AO2	Male	Oligodendroglial tumors	Anaplastic oligodendroglioma		
7	AO3	Male	Oligodendroglial tumors	Anaplastic oligodendroglioma		

Column Context Menu

Right clicking on the column header brings up the context menu for the column as seen in the figure below.



Depending on the column type, different items will be available in this menu.

Possible items in the context menu

Factor Enrichment – allows you to quickly identify those factors that are highly enriched in each column value subgroup. For example if performing factor enrichment on a clustering results column, if one of the clusters contains all high grade tumors, this tool would quickly identify such rare events.

KM-Plot – allows you to use this column to create a Kaplan-Meier survival plot. See the section on *Kaplan-Meier Curves* (page 143).

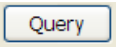
Delete – allows you to delete this column from the table.

Rename – allows you to rename this clustering analysis resulting in the column header changing to the new cluster name.

Use as Display Name – Clicking on a column header and selecting this will show the column header in bold indicating that values in this column will be used as the sample display name in all further analysis and visualization. If a cell in this column is empty, then the value from the original "Sample" column will be used as the display name. Renaming this column again will remove this as the Display name column. This is particularly useful to further identify the sample by adding details to the sample name, to de-identify a sample, or to shorten a sample name so the full name can be viewed in the individual sample tracks.

Select From File – Allows you to select samples based on either sample names or values found in an external text file. If performing this on the Sample Name, those names that appear in the file will be the samples that will be selected in the Data Set tab. If performing this on a Factor column, then the samples matching the factor values in the text file will be the selected samples.

QUERYING

The **Query** tool 

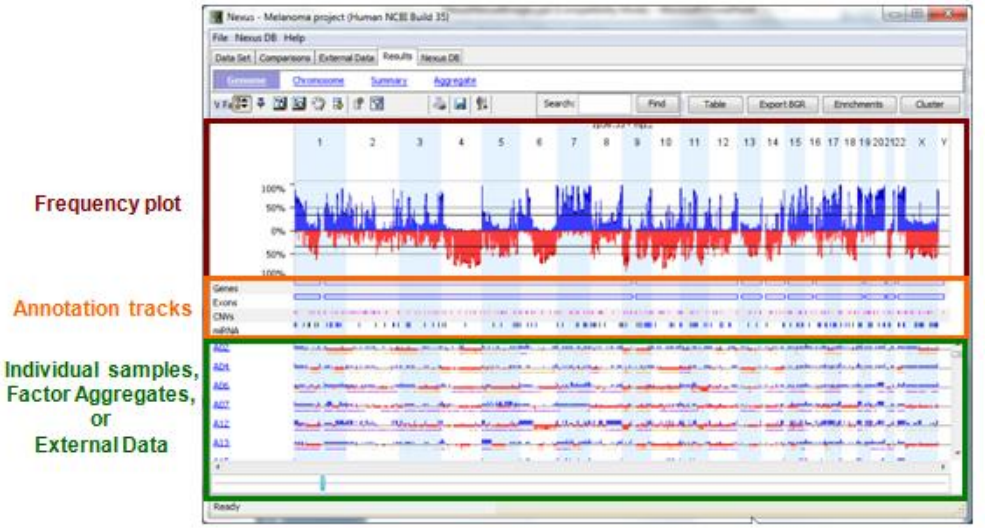
allows you to query your project for aberrant events for specified genes or regions. Please see the section on *Querying Data* for more information.

RESULTS

The **Results** tab is utilized to view and further analyze the processed data within a project. Only the samples that were selected by selecting the checkboxes in the **Data Set** tab are displayed here. If a sample in the **Data Set** tab is marked as **Processed** but is not checked off then it will not be displayed in the **Results** tab. There are several sub-tabs within the **Results** tab and their functions are described below.

GENOME AND CHROMOSOME VIEWS IN THE RESULTS

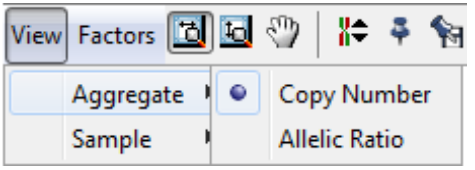
The **Genome** page displays information on all of the chromosomes at once whereas the **Chromosome** tab displays information on one chromosome at a time. Horizontally along the top panel are the visualization tools. Below this are the results of the analyses and this display is divided into three main parts as shown in the figure below.



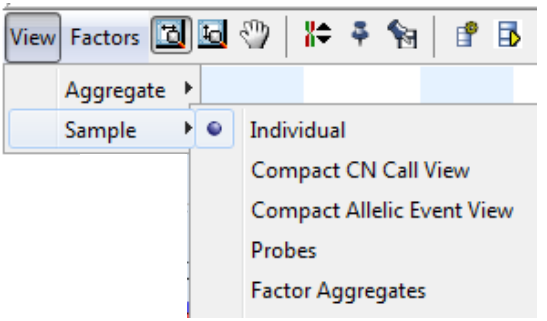
The top part of the window shows the frequency (or aggregate) plot for the samples. Blue indicates copy number gain events and red indicates copy number loss events.

Below this are the annotation tracks. By default, the **Genes, Exons**, **CNVs,** and **miRNA** tracks are displayed but can be turned off by choosing to do so in the **File->Options** window. Additional tracks for probes can be added as described in Appendix B.

Top panel display (copy number frequency or allelic events frequency) is dependent on what is selected under **View->Aggregate**:



The bottom panel display is dependent on what is selected under **View->Sample:**



The bottom part of the window can show one of the following:

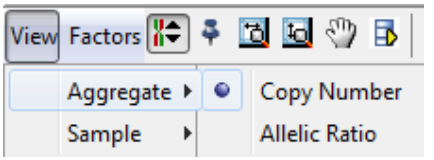
1. Copy number and allelic results for individual samples (one sample per row)
2. A compacted copy number calls view where the height of the calls is the same
3. A compacted allelic events view where the height of the calls is the same
4. Heat map of the probes intensities
5. Frequency plot of copy number or allelic ratio results for Factor Aggregates (sample groups depending on whether **Individual** or **Factor Aggregates**, respectively is selected in the **View** dropdown in the tool bar.

6. External data such as miRNA, methylation, or gene expression results if external data has been loaded into the project.

These items will be covered in more detail below.

THE FREQUENCY PLOT PANEL

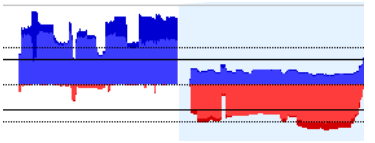
This top most area is used to display the occurrence frequency and is dependent on what is selected under the **View->Aggregate** menu.



If **Copy Number** is selected, then the plot on the top panel will display copy number change frequency. If **Allelic Ratio** is selected, the top panel will display allelic ratio plot.

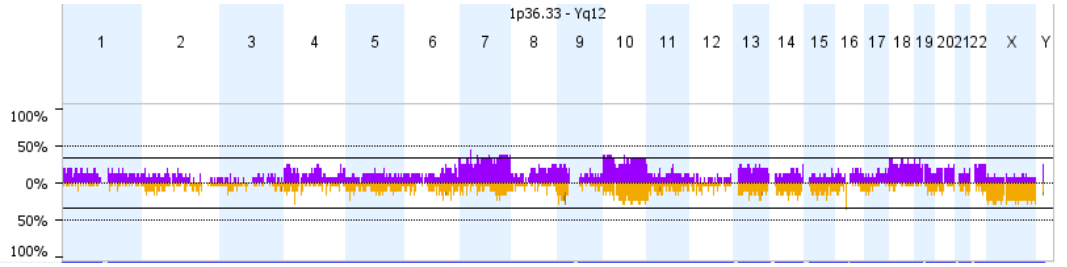
COPY NUMBER CHANGE FREQUENCY

The y-axis indicates the percentage of the population in the selected samples having an aberration at a specific point along the genome. Amplifications are plotted in blue above the 0% baseline. Deletions are plotted in red below the 0% baseline. Dark blue and dark red in the plot indicate two or more copy gain and homozygous loss, respectively. Some of the gain areas in the plot below are in dark blue indicating two or more copy gains.



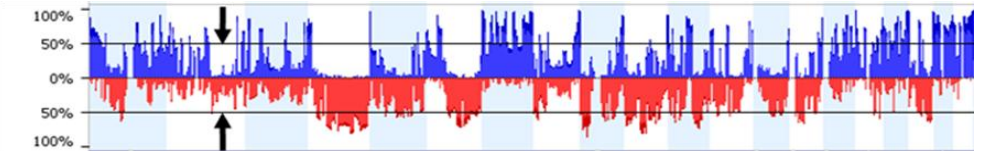
ALLELIC RATIO PLOT

The y-axis indicates the percentage of the population in the selected samples having the allelic event at a specific point along the genome. Allelic imbalance events are plotted in purple above the 0% baseline. LOH are plotted in brown below the 0% baseline.



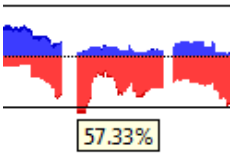
The figures below show copy number results but the features below apply to both the copy number and allelic ratio displays.

The black horizontal lines (marked with blue arrows in the figure below) delineate the **Aggregate % cut-off** value chosen.



This value is specified in the **Aggregate % cut-off** input box found at the bottom of the **Aggregate** tab.

Moving the mouse over the aberrant regions in the browser displays the percent of samples having this particular aberration at that point on the genome. The figure below shows that 57.33% of the samples have a loss at that point in the genome.

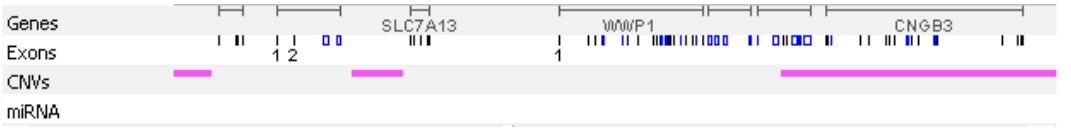


If the **Sort** tool is selected, clicking here will sort the individual samples below in order of smallest aberration of one type to the largest aberration of this type followed by aberrations of the other type. Please refer to the section on *Visualization Tools*, page 69 for more information.

ANNOTATION TRACKS

Tracks just below the frequency plot indicate genes, exons, CNVs (copy number variations from The Center for Applied Genomics' Database of Genomic Variants), and miRNA data (when available). Additional tracks for various probe types may be visible if they are selected under **Options->Tracks Selection**. By default, **Genes**, **CNVs, Exons,** and **miRNA** tracks (when available) are displayed but can be removed by editing the **Tracks Selection** options menu. See the *Options* section, page 220, for more information.

Four annotation tracks are displayed in this zoomed in region:



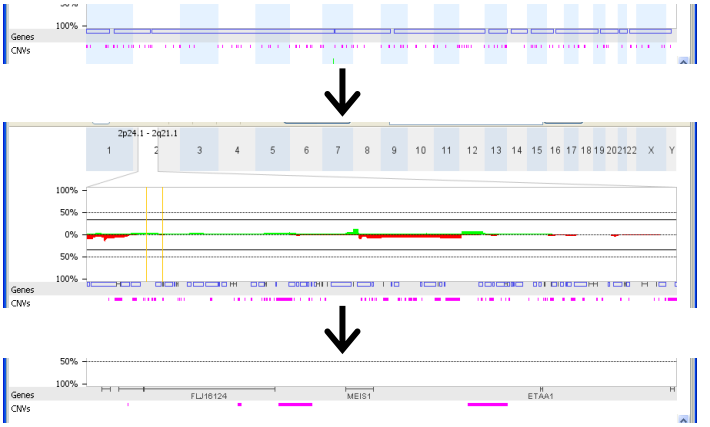
GENES TRACK

Genes displayed here are the longest variant of the gene. When completely zoomed out, only blue rectangles are displayed indicating regions containing genes. These blue rectangles are genes that are compacted together at the zoomed out level. As you zoom in on a region, individual genes are resolved. The blue rectangles convert to black lines and the gene names begin to appear as you increase the resolution (seen in the figure below).

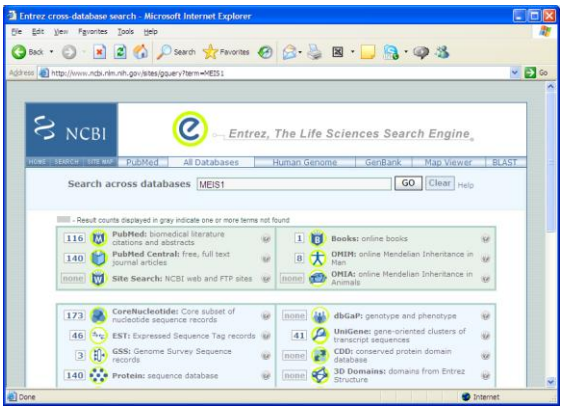
Genes are seen as Blue rectangles when completely zoomed out

Zooming in on area indicated by yellow vertical lines using the horizontal zoom tool

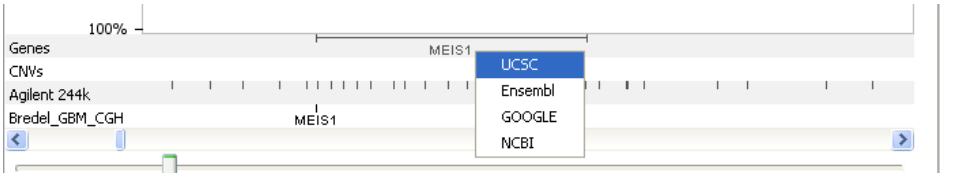
Zoomed in on area of chromosome 2. Gene names are visible in the Genes annotation track.



Clicking on a gene name (**MEIS1** for example) opens up NCBI's Entrez database search page (this is the default search page when Nexus Express is installed).



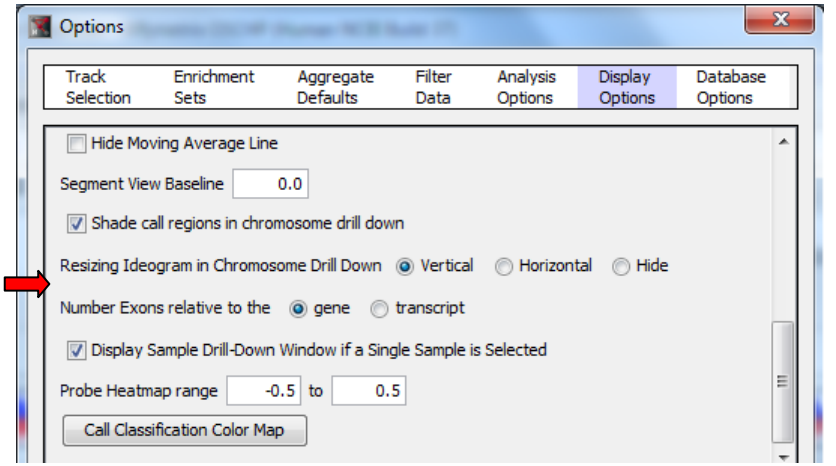
Right clicking on the gene name brings up a drop down menu giving you a choice of search engines (**UCSC**, **Ensemble**, **GOOGLE**, **and NCBI**).



Nexus Express remembers this choice and makes it your default web search until you change it again. Either the Ref Seq genes only or the Ref Seq genes *and* the UCSC genes can be displayed here. One needs to select the appropriate option in **Options->Track Selection**. For more information on how to select this, see the section *Track Selection*, on page 221.

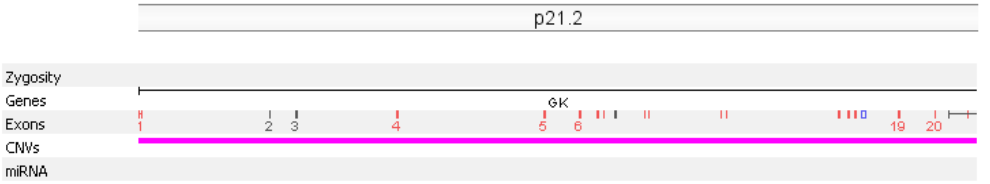
EXONS

The **Exons** track is displayed below the **Genes** track. When completely zoomed out, only blue rectangles are displayed indicating exonic regions. As you zoom in on a region, individual exons are resolved and the blue rectangles convert to black lines indicating exons. Right-clicking on an exon and selecting a transcript will reveal the exon numbers. If the gene has more than one transcript, the displayed transcript will have a bullet mark next to its name in the dropdown Numbering of the exons is based on the numbering option selected under **Options->Display Options.**

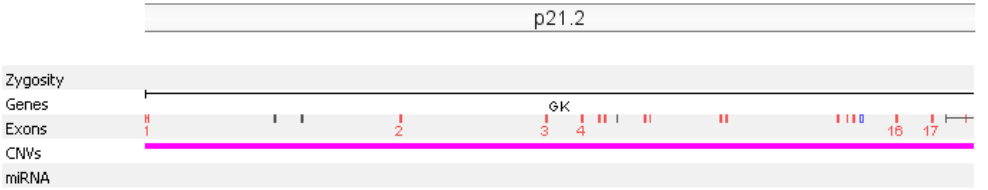


If **gene** is selected in the options, exons are numbered based on all exons in this gene and those comprising the selected transcript are displayed in red. If transcript is selected in Display Options, genes displayed in the **Genes** track are the longest variant of this gene and numbered sequentially starting at 1.

Exon numbering with respect to the gene (exons in this gene are numbered sequentially starting at 1 and those comprising the transcript are indicated in red):

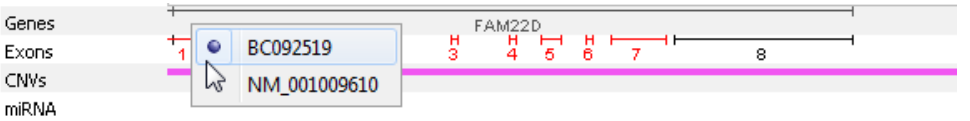


Exon numbering with respect to the transcript (the exons comprising this transcript are indicated in red and numbered sequentially from 1):

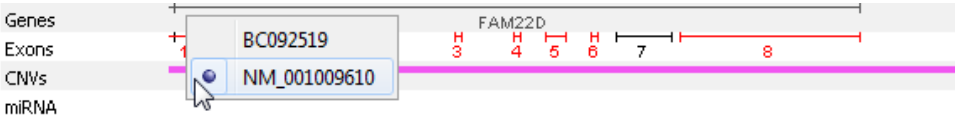


Multiple transcripts of the FAM22D gene with the exons making up the transcript indicated in red:

Transcript BC092519 excludes exon 8



Transcript NM_001009610 excludes exon 7



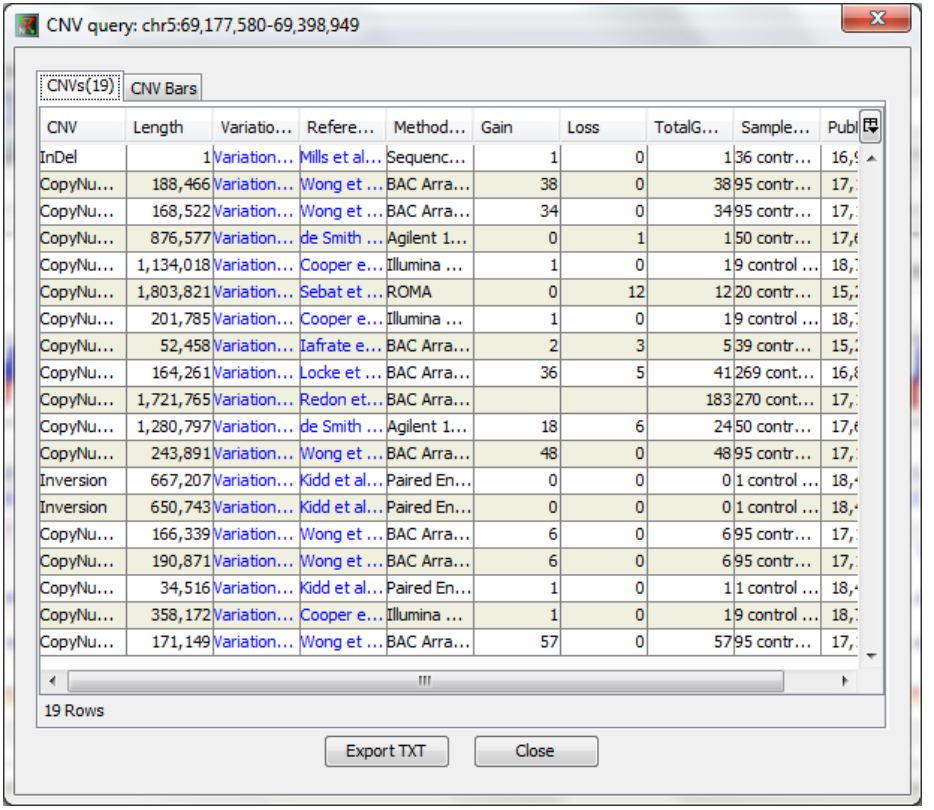
CNV TRACK

Known copy number variations are indicated in this track by colored bars spanning the CNV region. Magenta bars indicate copy number variations, pink bars indicate inversions, cyan bars indicate inversion breakpoints, and brown bars indicate Indels. The inversion breakpoints come in pairs because the chromosomal variation is due to inverted repeats spanning a region. When completely zoomed out, colored vertical lines are displayed in this track. As you zoom in on a region, the colored bars widen to indicate regions of known CNVs. Clicking on the CNV opens up a window providing more details about the CNV including the reference paper. CNV data is obtained from the Toronto Database of Genomic Variants (Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW, Lee C: Detection of large-scale variation in the human genome. Nat Genet. 2004 Sep; 36(9):949-51). There are two tabs in the window: CNVs and CNV Bars.

CNVs Tab

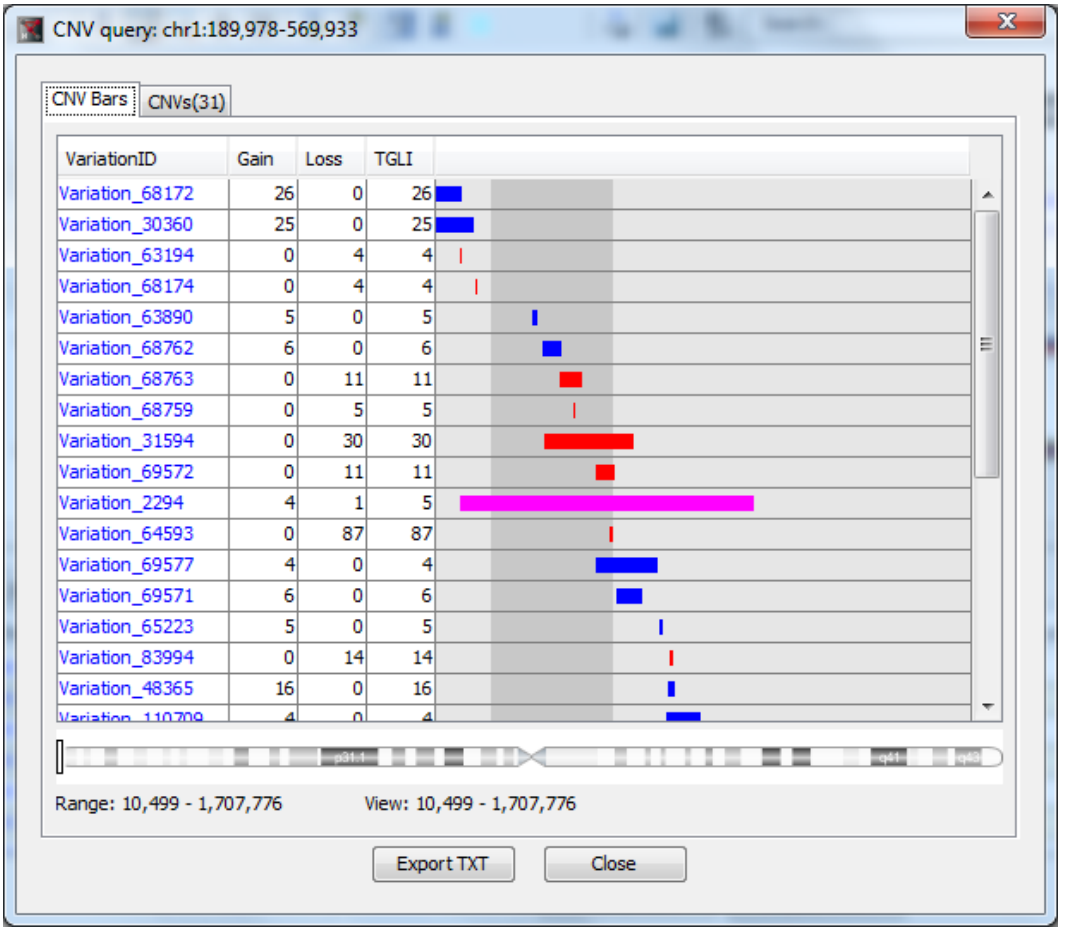
The CNVs tab contains table with further details on the CNVs. The **Gain** and **Loss** columns list the number of samples that showed the particular aberration in the cnv data set examined. The **TotalGainLossInv** column sums up the **Gain** and **Loss** column values. The **SampleSize** column shows the total number of samples that were examined in this reference. The **Methods/Platforms** column shows which technology was used to discover this cnv. Clicking on the blue hyperlinked **VariationID** opens the Toronto Database webpage for the CNV. Clicking on the blue hyperlinked **Reference** opens the PubMed entry for the reference and the PubMed ID is found in the **PubMedID** column.

CNV details:

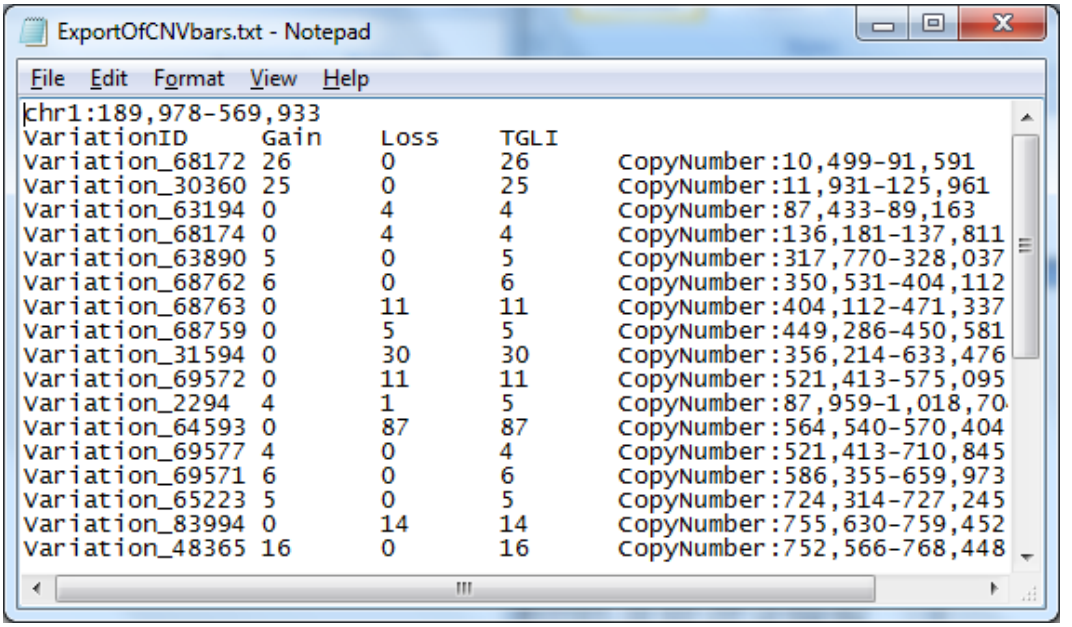


CNV Bars

The **CNV Bars** tab displays a graphical view of CNVs from each data set examined. It shows all CNVs comprising the entire CNV that was clicked upon in the CNV track. The point clicked upon is indicated with dark gray shading in the CNV bars tab; adjacent areas are shown in lighter gray. If all samples showed a gain the bar is represented in blue. If all samples showed a loss, the bar is displayed in red. If a mix of gain and loss were reported for a particular CNV, the bar will be displayed in magenta.



Moving the mouse over a track brings up the tool tip indicating the type of CNV (indel, inversion, etc.) and the chromosomal range it encompasses. To zoom in on a particular point, left click on that position. To zoom out, right click. To zoom out completely, quickly right click twice. The ideogram at the bottom has a rectangle indicating the chromosomal location of the CNV. The **Range** value indicates the CNV start and stop positions. The **View** value indicates the current range of the graphical view (when you are zooming in and out, this value changes). The **TGLI** is the **TotalGainLossInversion** column from the CNVs tab. This data can be exported in txt format using the **Export TXT** button. Following is the exported text version of the figure above:

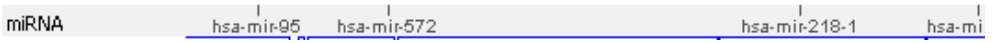


CUSTOM CNV TRACKS

You can create and display your own CNV data as well. Please see the section on *Adding a Custom CNV Track,* page 265.

MIRNA TRACK

If miRNA data is available for a particular organism build, then it's possible to display this data in its own track. The miRNA information for Human Build 36.1 is from the miRBase sequence database (http://microrna.sanger.ac.uk/) maintained by the Sanger Institute. A single vertical gray line marks the location of the miRNA transcript and the transcript ID is displayed below the tick mark. The tick marks are clickable links pointing to the sequence information from the miRBase database.



PROBE TYPE AND CUSTOM TRACKS

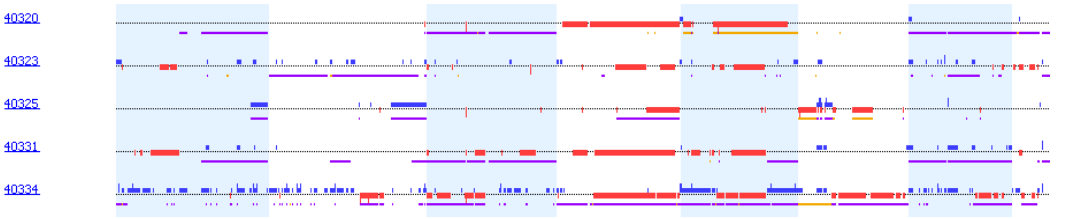
Additional tracks displaying probe locations can be displayed below the **CNVs** track. A separate track will appear for each probe type that is selected under **Track Selection** in the **Options** menu. See the *Options* section (page 220) for more information.

SAMPLES PANEL

Below the Annotation tracks is the area showing call results for individual samples, frequency plots for Factor Aggregates, or plots for External data. Which one of these types of data will be visible is determined by what is chosen under **View->Sample**. Two options are always present under the **View** menu: **Individual** and **Factor Aggregates**. Additional options are present if any External Data is loaded.

INDIVIDUAL SAMPLES/FACTOR AGGREGATES

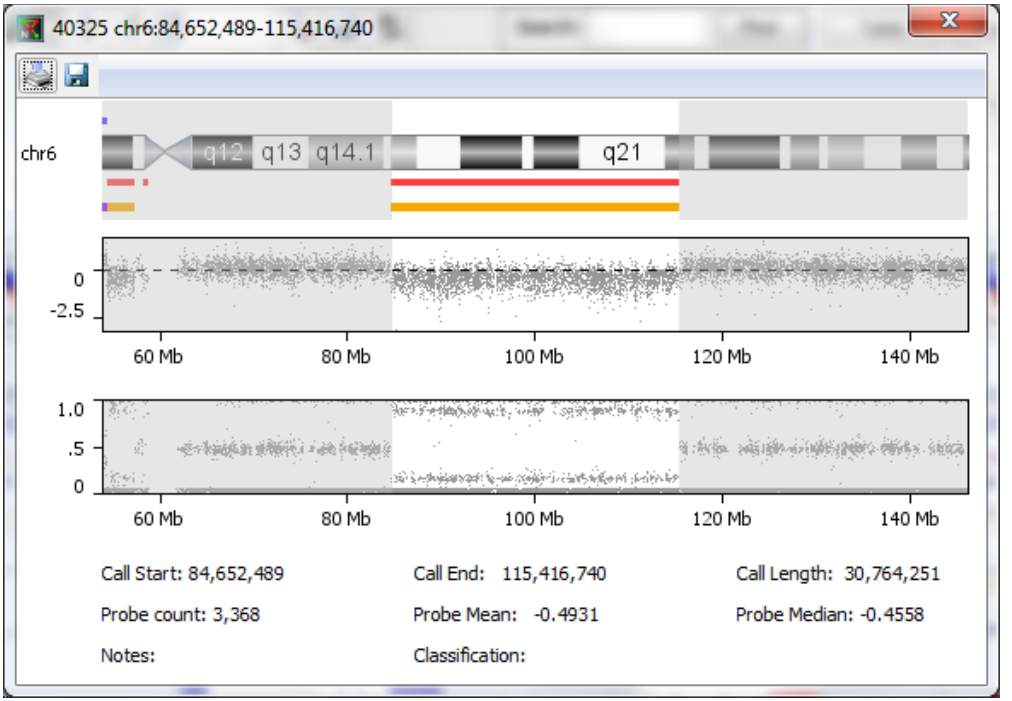
Frequency data for copy number aberrations as well as allelic ratio occurrence is displayed for each sample individually on separate rows. Copy number aberrations are displayed above the dotted line (blue for gain and red for loss). Allelic events are displayed below the dotted line (purple for allelic imbalance and brown for LOH).



A horizontal slider at the bottom can be used to adjust the number of visible samples in the window. Sliding it to the left compacts the samples vertically to allow more samples to be displayed in the viewable area.

When individual samples are displayed, the height of the red and blue bars represents the extent of loss or gain. Taller bars for copy number events indicate a higher gain or two copy loss than shorter bars. Clicking on the sample name will launch a new window displaying the probe level data for that sample (see section on *Individual Sample Drill Down*, page 104). Right clicking on the individual sample plot and selecting **Open Region**

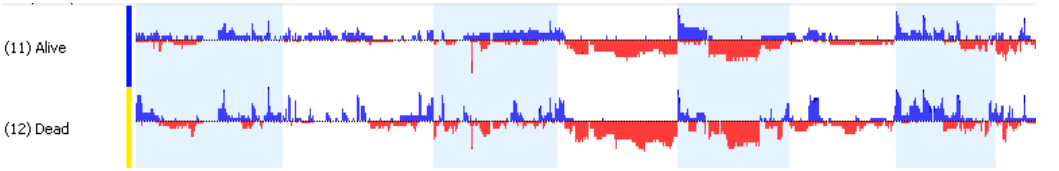
in Sample Window will zoom in on that region by launching the Sample Drill Down window with the **Chromosome** tab on top. Left-clicking on an aberration in the individual sample tracks brings up a window with a zoomed in probe level view of the call and information such as start, end, length etc. of the aberration. This is much faster way to take a quick look at the aberration vs. opening a drill down window.



The image can be saved or printed using the respective icons in the top left of the window. To open up a sample drill down window, right click on the aberration and select **Open Region in Sample Window.**

When Factor aggregates are viewed, samples with the same factor values are grouped together and are displayed as a frequency plot. If the **Aggregate** display is selected to be **Allelic Ratio** as in the figure below, then allelic events will be displayed in this panel (figure below).

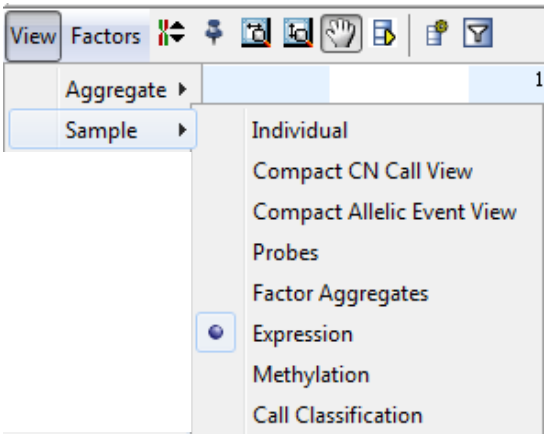
If **Copy Number** is selected, then the display will show copy number change frequency plots.



Moving the mouse over a specific location on the plot displays the frequency of gain/loss or LOH/allelic imbalance at that position and the base pair location. For more on this feature, please review the information on **Factors** tool in the *Visualization Tools* section, page 69.

External Data

If any External Data such as Expression, miRNA, or Methylation is loaded into the project, an option with the external data type is available under the **View->Sample** menu.



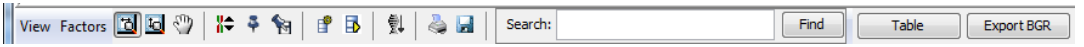
Selecting this will show plots for the appropriate external data that has been loaded into the project. One track is displayed for each external data input file that is loaded. Either **Up** or **Down** must be checked off in the **External Data->Expression** tab for plots to appear in the **Results** window.

In the expression plot, colored bars represent genes that are up or down regulated. The bars span the length of the gene. Blue bars are used for up regulated genes and these bars appear above the horizontal median line. Magenta bars represent down regulated genes and these bars appear below the median line. When zoomed out, the bars look like single vertical lines. Upon zooming in, they convert to bars. Moving the mouse over the bars displays the gene symbol and additionally the p-value and/or log ratio if this data was present in the input file.

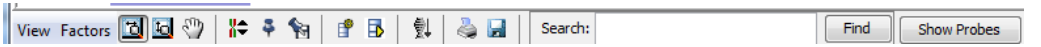
VISUALIZATION TOOLS

The tool bars for the **Genome** and **Chromosome** pages have only a few differences and the differences are in the buttons located to the right of the Search panel. The **Chromosome** page lacks the **Table,** and **Export BGR,** buttons but has a **Show Probes** button.

Tool bar for the Genome page:

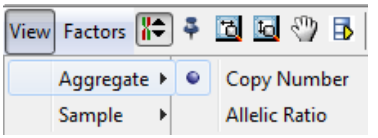


Tool bar for the Chromosome page:

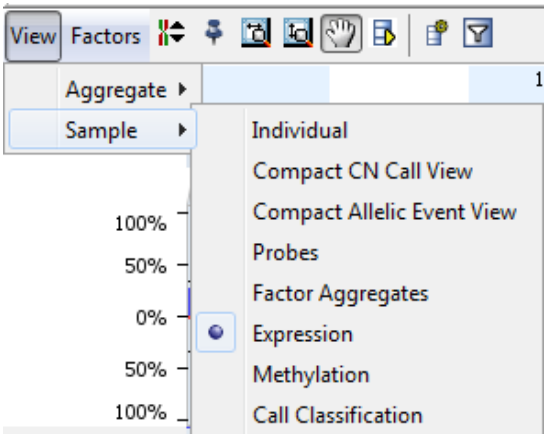


View: Allows you to see the data within the project in different forms in the top and bottom panel of the window. Options in the menu are

- **Aggregate –** refers to the plot in the top panel which can either display copy number or allelic ratio

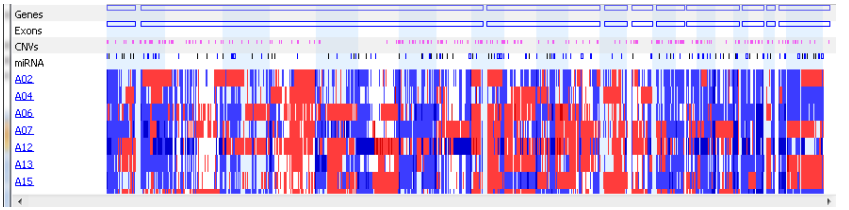


- **Sample –** refers to the bottom panel which can display the items listed below. If no Expression data was loaded then the Expression option will not be visible in the menu. If Expression data was loaded, you must click on **View** from the **Data Set** tab for the Expression option to show up in the menu.

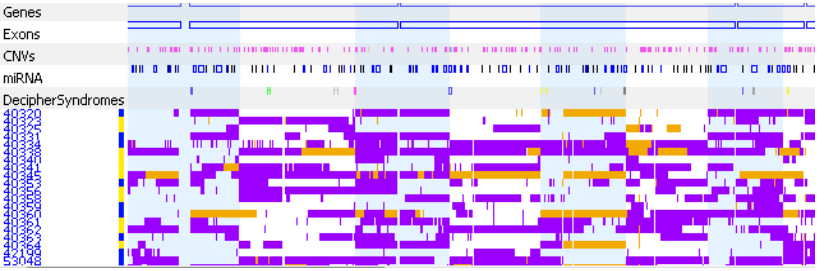


a. **Individual** – Each sample is displayed in its own track and depicts gain and loss information if the Aggregate view is **Copy Number** or displays allelic imbalance/LOH data if the Aggregate view is **Allelic Ratio**.

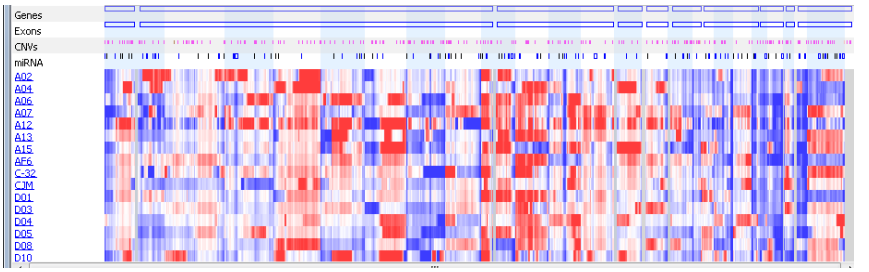
b. **Compact CN Call View** – Displays a condensed view of the copy number calls where the events are drawn at the same height (useful for displaying focal events).



c. **Compact Allelic Event View -** Displays a condensed view of the allelic events where the events are drawn at the same height (useful for displaying focal events).



d. **Probes** – Probe intensities are displayed as a heat map

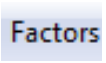


e. **Factor Aggregates** – Data is grouped together by Factors and the frequency plots are displayed for these Factor Aggregates. Once **View->Factor Aggregates** is selected, you need to specify the Factors you

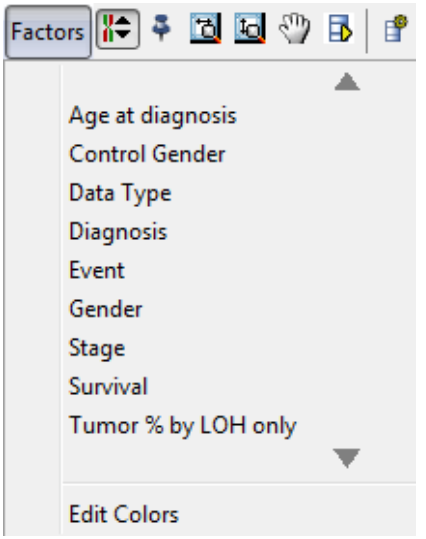
want to see using the **Factors** drop down menu before any factor aggregate data is displayed.

f. **Expression** – Any expression data that was loaded into the project will be visible in the bottom part of the window. Each loaded expression file will be displayed in its own track.

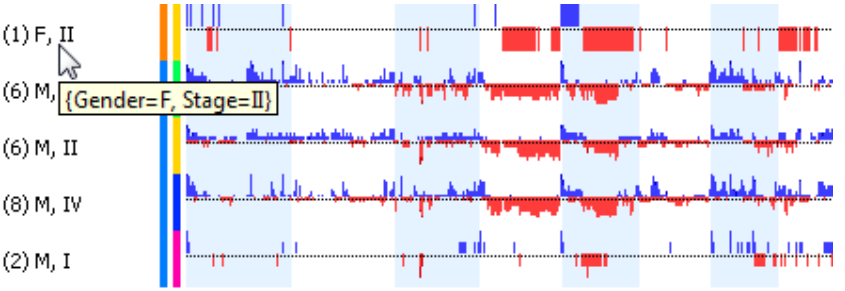
g. **Methylation** - Any methylation data that was loaded into the project will be visible in the bottom part of the window. Each loaded methylation file will be displayed in its own track.



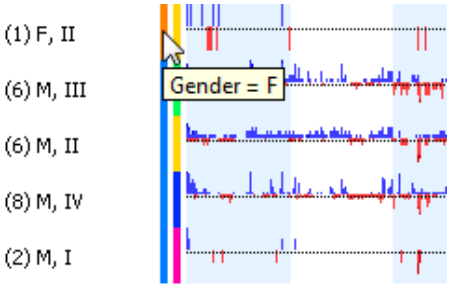
Factors: This menu lists all the Factors that are associated with each sample, any Clusters that were generated, the Data Type column, and an option to **Edit Colors**. Once you choose to view by **Factor Aggregates**, you'll need to select some Factors so that the data can be displayed in the bottom of the window. Once the specific Factors are selected in this menu, a check mark appears next to the Factor name within the menu drop down. If more than one Factor is selected, aggregate plots for all combinations of the selected Factor names are displayed.



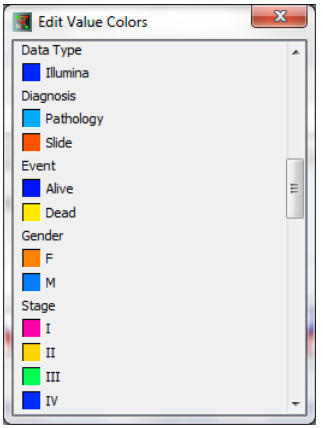
In the figure below, on the left side of the panel are the names for each Factor group. Moving the mouse over this will show exactly which Factors are included in the particular aggregate (see figure below). In parentheses towards the left of the name is a number indicating how many samples are in this group. When factor aggregates are created, vertical colored bars appear on the left of the plot indicating the factor groupings. There is one column of vertical bars per factor chosen.



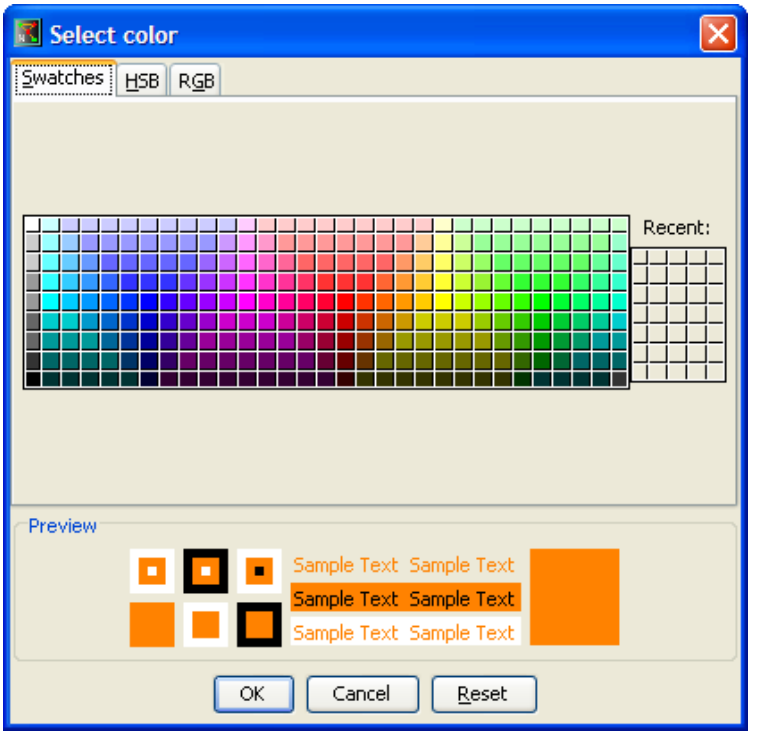
Moving the mouse over the colored bars displays the factor name. In the figure below, positioning the mouse over the orange bar indicates that this color represents the **Gender** factor where the value is **F**.



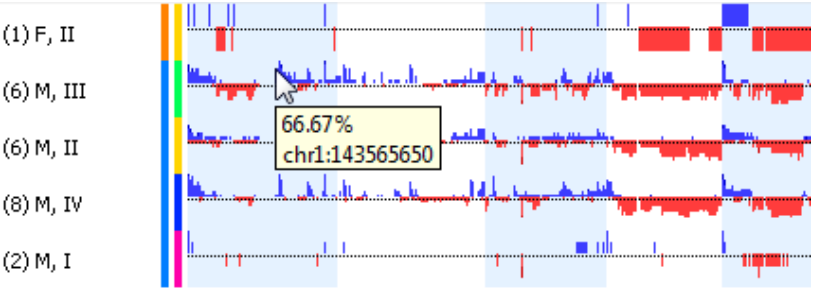
To see the full legend for the color coding, select **Factor->Edit Colors**. The **Edit Value Colors** dialog appears which shows the color assigned to each Factor name.



The colors can be edited here by clicking on the colored box and selecting another color in the resulting **Select Color** dialog via the color swatch, specifying RGB (Red/Green/Blue) values, or specifying HSB (Hue/Saturation/Brightness) values.

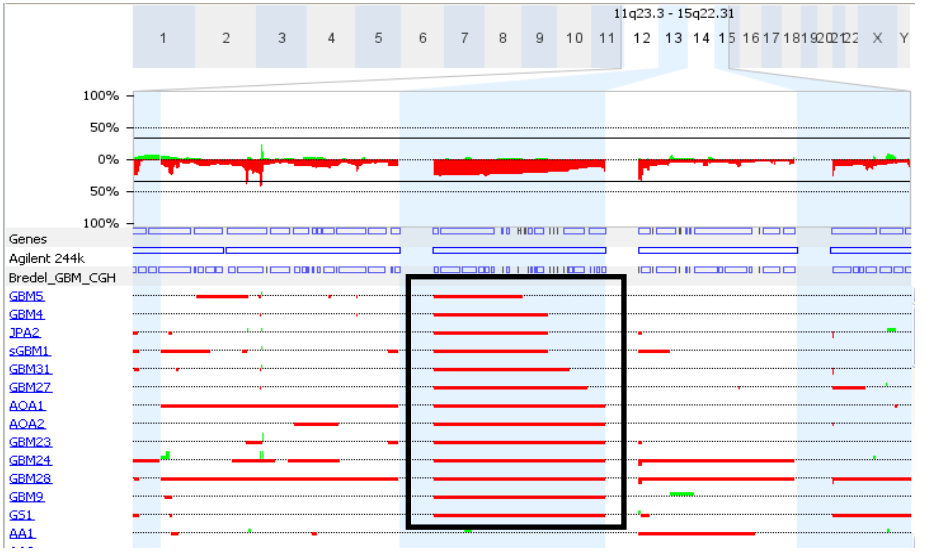


Moving the mouse over a specific position on the plot displays the occurrence frequency at that point and the base pair location on the chromosome.



 **Sort:** The sort tool allows sorting the individual samples in the bottom panel based on the length and type of aberration. This tool is selected by clicking on the **Sort** button. If the **Sort** tool is selected, clicking it will sort the individual samples below in order of smallest aberration of one type to the largest aberration of this type followed by aberrations of the other type. For example, clicking on a loss (red) region will sort the samples so that the samples with the smallest loss are placed at the top, followed by samples with larger losses, and finally samples with gains. Sorting samples at this point will also sort the sample list in the **Data Set** tab in the same order.

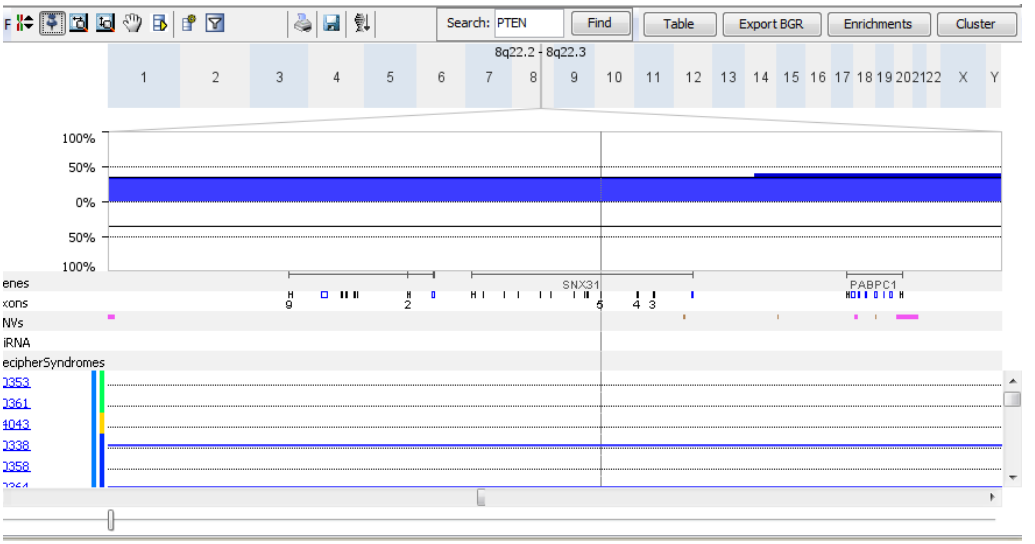
Individual samples sorted on an area of deletion on chromosome 13:



Mark Tool: This tool allows marking a location with a vertical line running through the frequency plot, annotations and sample tracks. It is particularly useful when you find something in the zoomed-in-view and would like to remember that location when you zoom out or when you open an individual sample drill down. The mark is retained throughout the session and will be removed the next time you open the project, unless the marks were saved (see next item below). The marks will appear in sample drill down views as well as the aggregate view. You can change the color of the mark as well as annotate it by right clicking on the mark and choosing from the menu.

Save Marks: This tool allows you to save marks by specifying a file name. The marks can later be selected to be displayed via the **Track Selection** tab in the **Options** window. A **Select Custom Marks** section will be displayed under the **Select CNV track** section in the **Track Selection** tab. Custom marks files saved into the **custommarks** folder (default

save location) of the genome build will automatically be loaded the next time the project is opened.



Horizontal Zoom: Allows you to zoom in along the x-axis on the genome or chromosome frequency plots. This tool is selected by clicking on the **Horizontal Zoom** icon. Once selected, the mouse cursor converts into a magnifying glass when the cursor is over the plots indicating that this tool is active.

Vertical Zoom: Allows you to zoom in along the y-axis on the genome or chromosome frequency plots. This tool is selected by clicking on the **Vertical Zoom** icon. Once selected, the mouse cursor converts into a magnifying glass when the cursor is over the plots indicating that this tool is active.

To zoom in at a specific position

Left click on a specific position on the plot to zoom in.

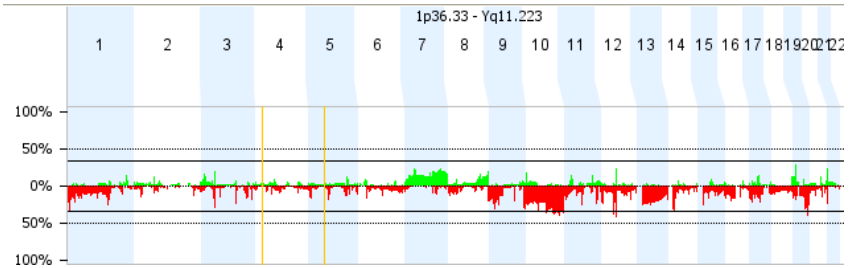
To zoom in on a region

Left click the mouse and hold while dragging the mouse to the right or left for the Horizontal Zoom or up or down for the Vertical Zoom and then release the mouse button. Two vertical yellow lines (for Horizontal zoom) or two horizontal yellow lines (for Vertical Zoom) on the plot will indicate the region that will be zoomed into.

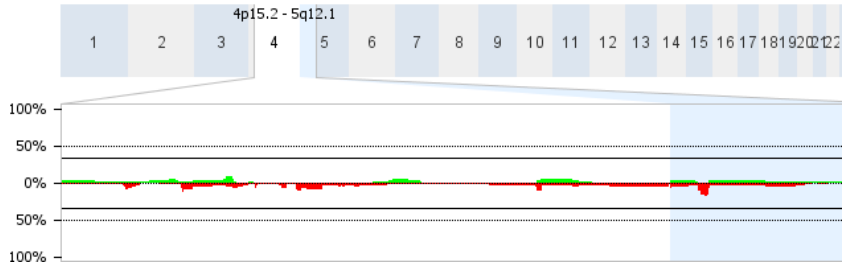
To zoom out

Double click the right mouse button to zoom out completely. To zoom out incrementally, click the right mouse button repeatedly.

In the figure below we are zooming in along the x-axis using the **Horizontal Zoom** tool, on a region spanning chromosome 4 and 5. Yellow vertical lines indicate the region that will be zoomed into.



Zoomed in view of the chromosome 4 and 5 region:

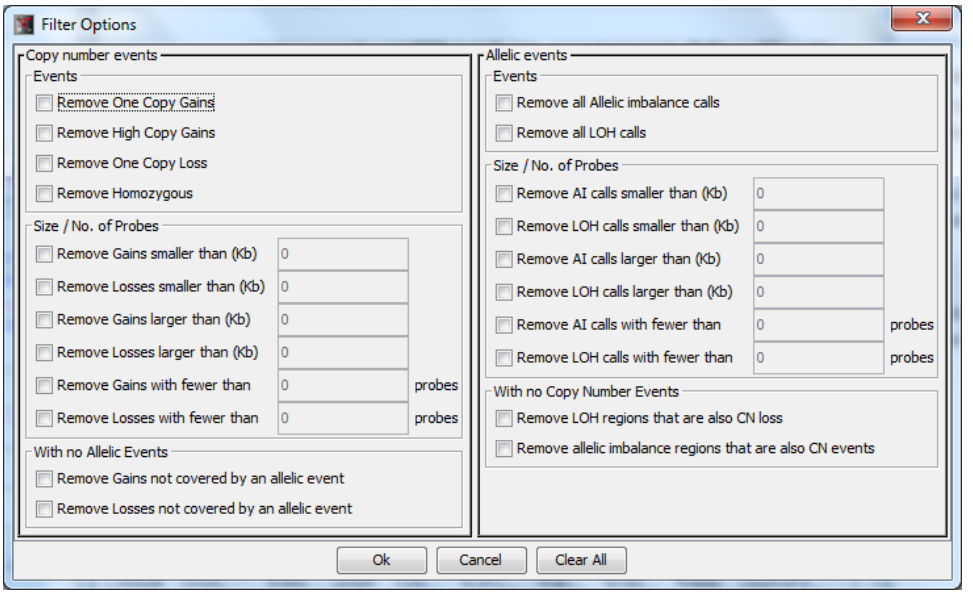


 **Pan:** Allows you to pan across the genome. Left click and hold down while moving the mouse left or right across the genome to pan. This function is also accomplished by using the horizontal scroll bar near the bottom of the window.

Query Database: This tool is displayed if the Nexus license has Nexus DB enabled and allows querying the repository aberrations in other projects at a specific region. See the Section on Nexus DB for more information on this tool.

Region Query Settings: Allows you to quickly access the filtering settings for making region queries against the Nexus DB repository. These settings are also available via **Options->Database Options**.

Filter: Provides quick access to the **Filter Data** options to filter out calls from the display. Clicking the button brings up a window with the Filter Data options accessed via the **File->Options** menu.



If any one of the checkboxes is marked, the button turns red indicating that the filter is active:

When a filter is used, any affected data columns in the Data Set tab will have a * next to the header indicating that the column contains filtered data. When “Remove LOH regions that are also CN Loss” is selected, **LOH** and **%LOH** have a * next to them as shown below.

LOH*	% LOH*
328	40.402
473	26.08
300	25.789

See the section on

Filter Data, page 225, for more information on filtering options.

 **Print**: Prints the contents of the active tab.

 **Save Image**: Saves the image visible in the active tab as an image file or to the clipboard. The resolution, image format, and whether to save as file to copy to clipboard can be modified from within the Save Dialog.

 **Drill Down**: Clicking on the drill down button brings up a window containing annotation tables.

Search tool: Allows you to search by gene name, base pair range, probe id, or cytoband.

Range search: Search for a base pair range. A base pair range is specified by giving the chromosome prefix along with the base pair range in the **Search** box in the **Genome** page. The chromosome prefix is not required when performing a search in the **Chromosome** page.

Example range search in **Genome** page: chr8:172000-180000

Example range search in **Chromosome** page: 172000-180000

Cytoband search: Search for a cytoband on a chromosome.

Example search in **Genome** page: **1:p36.33**

Example search in **Genome** page: **1p36.33**

Gene search: Search for a gene using its symbol. This search is case insensitive.

Example: **MECP2** or **mecp2**

Probe id search: Search for the location of a specific probe id by entering the probe id in the search box.

Example: **A_16_P21633817** (an Agilent probe name)

miRNA search: Search for a microRNA by entering the miRNA name in the search box.

Example: **mir-544a**

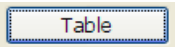
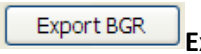


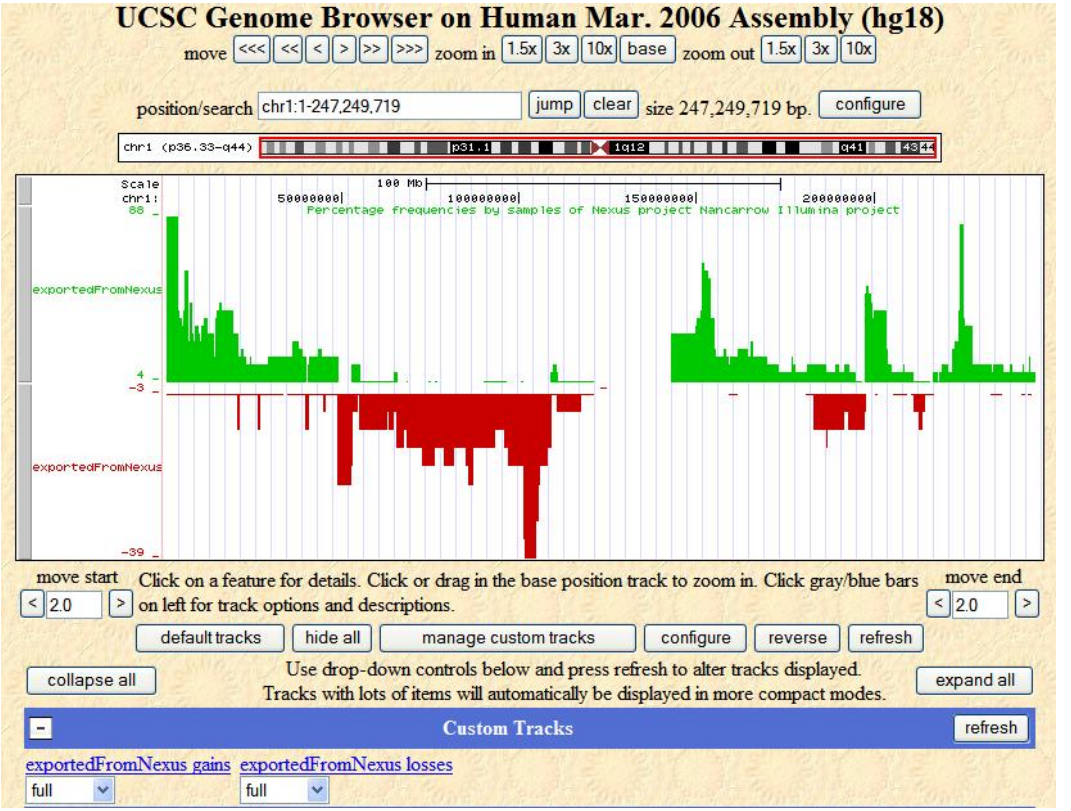
Table button: Only in the **Genome** page. The **Table** button will generate a new window listing samples along with each segmented region in all chromosomes, the aberration for this region and other information about the aberrant region. Please see the section on *All Samples and Aberrations Table,* page 100, for more information on this.

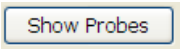


Export BGR button: Only in the **Genome** page. The **Export BGR** button will export the frequency plot values as a BEDGRAF (.bgr) file so that it can be used in other browsers such as the UCSC Genome browser. For more information on BEDGRAF format please see the UCSC Genome Browser help section: http://genome.ucsc.edu/goldenPath/help/bedgraph.html. When uploaded to the UCSC

Genome Browser, aberration frequencies will be displayed with gains in one track and losses in another track. The name of the tracks will be the file name of the exported data suffixed with “losses” or “gains”.

A BEDGRAF file called “exportedFromNexus” is uploaded into the UCSC Genome Browser:





Show Probes button: Only in the **Chromosome** page. The **Show Probes** button allows you to display probes from multiple samples in a common aberrant region

in one window. See section below on *Multi-Sample Probe View* for more information on this tool.

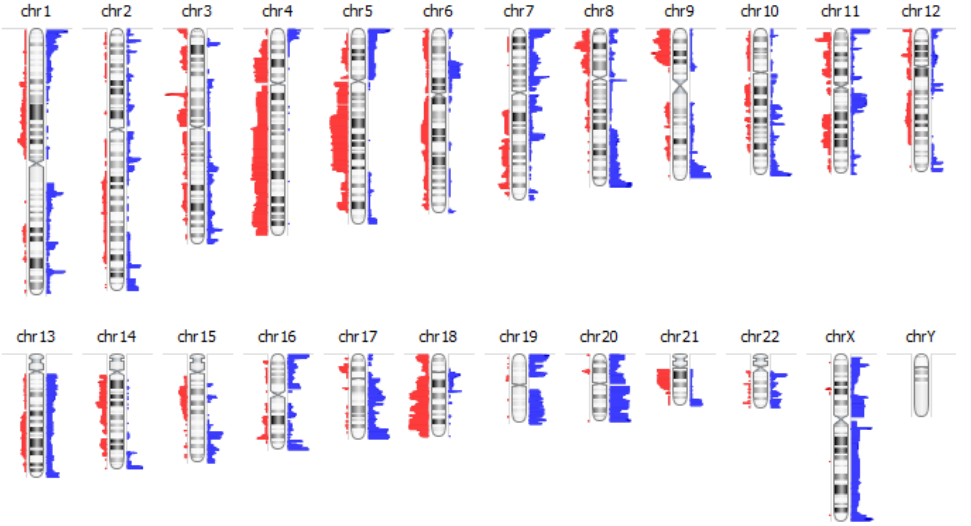
This drop down on the bottom left of the window is used to select which chromosome to view. This can also be used as a "Previous" and "Next" buttons to step through each chromosome in the genome. Highlighting the selected chromosome number and using the up and down arrows, displays the previous and next chromosomes, respectively.

SUMMARY – FREQUENCY PLOTS FOR ALL CHROMOSOMES

The **Summary** page in the **Results** tab displays the frequency plots for all chromosomes in one view. It is a nice compact view of gain and loss regions or LOH and allelic imbalance events across all chromosomes. The selection from the **View** menu determines what is displayed.

If the project contains SNP data with B allele frequencies, then Allelic Events can be displayed.

Copy Number aberration view:

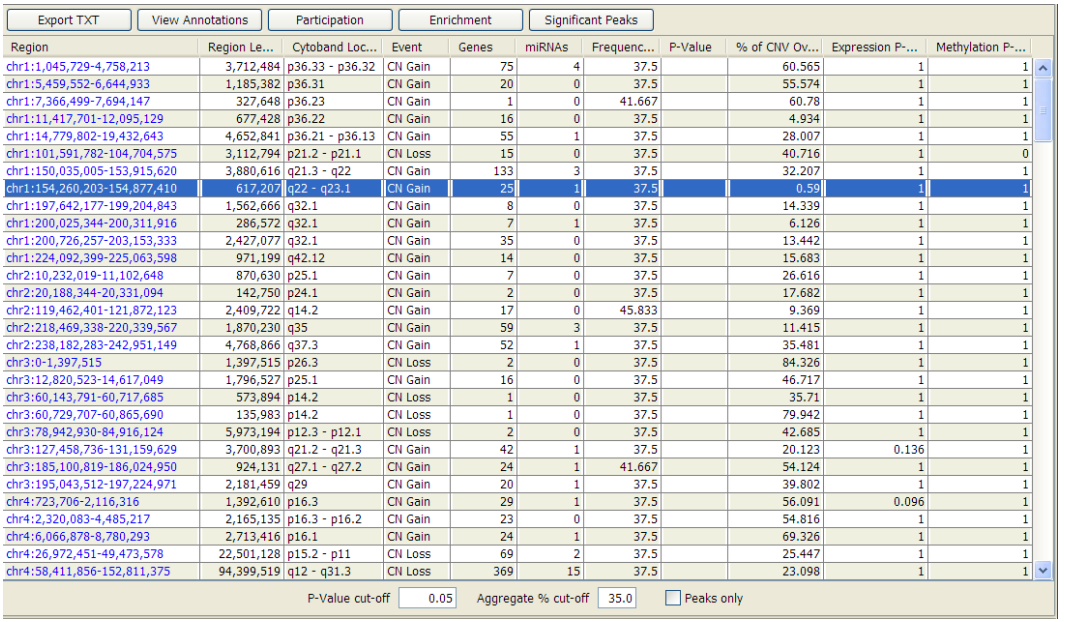


Allelic events view:

Clicking on the chromosome ideograms and chromosome numbers will bring forward the **Chromosome** page displaying data for the chromosome that was selected. You can print or save images by using the **Print** and **Save** icons at the top of the tab.

AGGREGATE – REGIONS OF COMMON ABERRATIONS

The **Aggregate** page consists of a data table and the following buttons: **Export TXT**, **View Annotations**, **Participation, Significant Peaks** and input fields for a few statistical parameters (**Aggregate % cut-off, P-value cut-off, Peaks Only)**.



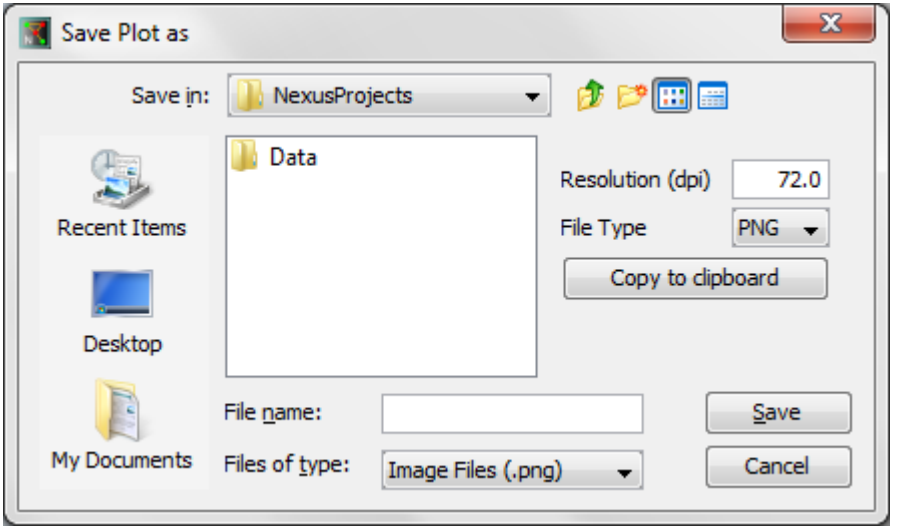
Region	Region Le...	Cytoband Loc...	Event	Genes	miRNAs	Frequenc...	P-Value	% of CNV Ov...	Expression P-...	Methylation P-...
chr1:1,045,729-4,758,213	3,712,484	p36.33 - p36.32	CN Gain	75	4	37.5		60.565	1	1
chr1:5,459,552-6,644,933	1,185,382	p36.31	CN Gain	20	0	37.5		55.574	1	1
chr1:7,366,499-7,694,147	327,648	p36.23	CN Gain	1	0	41.667		60.78	1	1
chr1:11,417,701-12,095,129	677,428	p36.22	CN Gain	16	0	37.5		4.934	1	1
chr1:14,779,802-19,432,643	4,652,841	p36.21 - p36.13	CN Gain	55	1	37.5		28.007	1	1
chr1:101,591,782-104,704,575	3,112,794	p21.2 - p21.1	CN Loss	15	0	37.5		40.716	1	0
chr1:150,035,005-153,915,620	3,880,616	q21.3 - q22	CN Gain	133	3	37.5		32.207	1	1
chr1:154,260,203-154,877,410	617,207	q22 - q23.1	CN Gain	25	1	37.5		0.59	1	1
chr1:197,642,177-199,204,843	1,562,666	q32.1	CN Gain	8	0	37.5		14.339	1	1
chr1:200,025,344-200,311,916	286,572	q32.1	CN Gain	7	1	37.5		6.126	1	1
chr1:200,726,257-203,153,333	2,427,077	q32.1	CN Gain	35	0	37.5		13.442	1	1
chr1:224,092,399-225,063,598	971,199	q42.12	CN Gain	14	0	37.5		15.683	1	1
chr2:10,232,019-11,102,648	870,630	p25.1	CN Gain	7	0	37.5		26.616	1	1
chr2:20,188,344-20,331,094	142,750	p24.1	CN Gain	2	0	37.5		17.682	1	1
chr2:119,462,401-121,872,123	2,409,722	q14.2	CN Gain	17	0	45.833		9.369	1	1
chr2:218,469,338-220,339,567	1,870,230	q35	CN Gain	59	3	37.5		11.415	1	1
chr2:238,182,283-242,951,149	4,768,866	q37.3	CN Gain	52	1	37.5		35.481	1	1
chr3:0-1,397,515	1,397,515	p26.3	CN Loss	2	0	37.5		84.326	1	1
chr3:12,820,523-14,617,049	1,796,527	p25.1	CN Gain	16	0	37.5		46.717	1	1
chr3:60,143,791-60,717,685	573,894	p14.2	CN Loss	1	0	37.5		35.71	1	1
chr3:60,729,707-60,865,690	135,983	p14.2	CN Loss	1	0	37.5		79.942	1	1
chr3:78,942,930-84,916,124	5,973,194	p12.3 - p12.1	CN Loss	2	0	37.5		42.685	1	1
chr3:127,458,736-131,159,629	3,700,893	q21.2 - q21.3	CN Gain	42	1	37.5		20.123	0.136	1
chr3:185,100,819-186,024,950	924,131	q27.1 - q27.2	CN Gain	24	1	41.667		54.124	1	1
chr3:195,043,512-197,224,971	2,181,459	q29	CN Gain	20	1	37.5		39.802	1	1
chr4:723,706-2,116,316	1,392,610	p16.3	CN Gain	29	1	37.5		56.091	0.096	1
chr4:2,320,083-4,485,217	2,165,135	p16.3 - p16.2	CN Gain	23	0	37.5		54.816	1	1
chr4:6,066,878-8,780,293	2,713,416	p16.1	CN Gain	24	1	37.5		69.326	1	1
chr4:26,972,451-49,473,578	22,501,128	p15.2 - p11	CN Loss	69	2	37.5		25.447	1	1
chr4:58,411,856-152,811,375	94,399,519	q12 - q31.3	CN Loss	369	15	37.5		23.098	1	1

The table consists of numerical data for the results of the selected samples (these are selected in the **Data Set** tab). This is the same data that is visualized in the **Genome** page. The data table consists of regions across the whole genome that have an aberration meeting or exceeding the **Aggregate % cut-off** value specified in the bottom of tab. For example, if 35% is the **Aggregate % cut-off** value then the table will only list those regions which are lost or gained in at least 35% of the population.

Column	Description
Region	The segmented region with an aberration listed in the following format chr8:172,199-300,002
Region length	The length of this aberrant region in base pairs
Cytoband Location	The cytoband on which this region resides
Event	The copy number or allelic event at this region. Possible values are **CN Loss, CN Gain, Allelic Imbalance**, and **Allelic Loss**
Genes	Number of genes in this region
miRNAs	Number of miRNAs in this region
Frequency %	Percentage of the samples in the data set having this event.
P-Value	If the Significance analysis was run, then a p-value would appear here.
% of CNV Overlap	Percent of this region that is covered with known CNVs (those listed in the CNV track)
Expression P-Value/Expression Q-Bounds	If Expression data is available in this project, then a p-value or Q-bounds for the expression data will appear here. If multiple expression data sets are checked off to be viewed, then the value here corresponds to the expression dataset with the lowest p-value/Q-bounds. This column will display either Expression P-Value or Expression Q-Bounds depending on whether the value has been corrected for multiple testing.

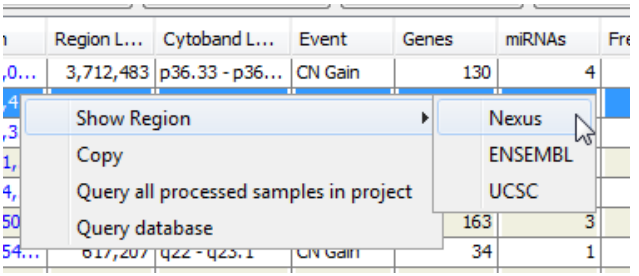
Column	Description
Methylation P-Value/Methylation Q-Bounds	If Methylation data is available in this project, then a p-value or Q-bounds for the methylation data will appear here. If multiple methylation data sets are checked off to be viewed, then the value here corresponds to the methylation dataset with the lowest p-value/Q-bounds. This column will display either Methylation P-Value or Methylation Q-Bounds depending on whether the value has been corrected for multiple testing.

Columns can be sorted and rearranged as described in the *Nexus Express Basics* section under

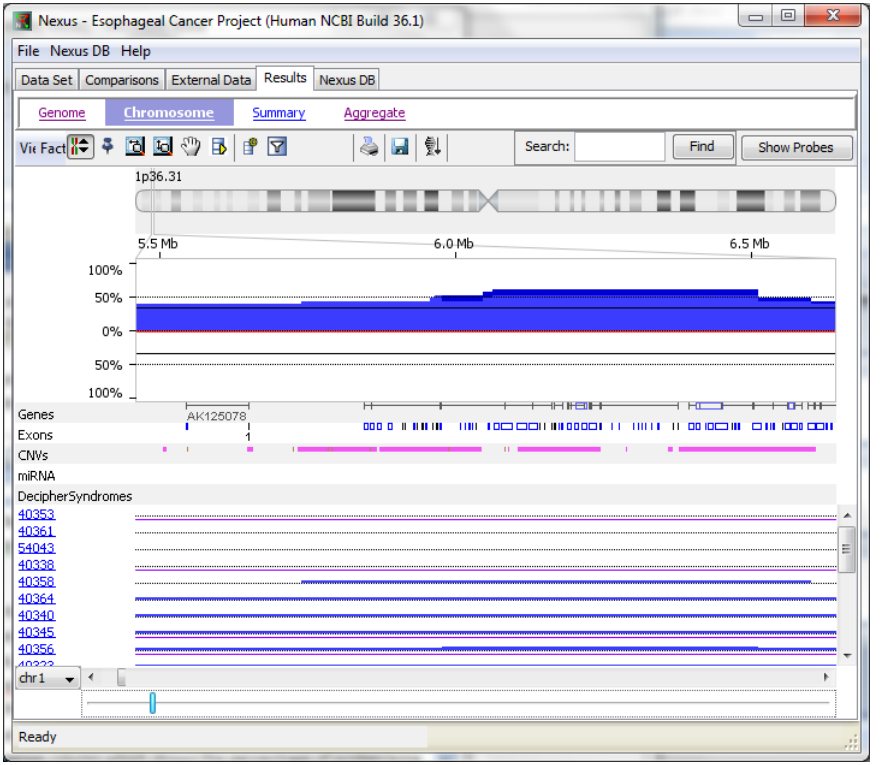


Data Table page *16*. Clicking on the hyperlinked region name brings up the annotation window. Please see the section on Annotations for more information. Right clicking on the region name opens up a drop down menu with choices: **Show Region**, **Copy, Query all processed samples in project**, and **Query Database** (if the Nexus DB module is

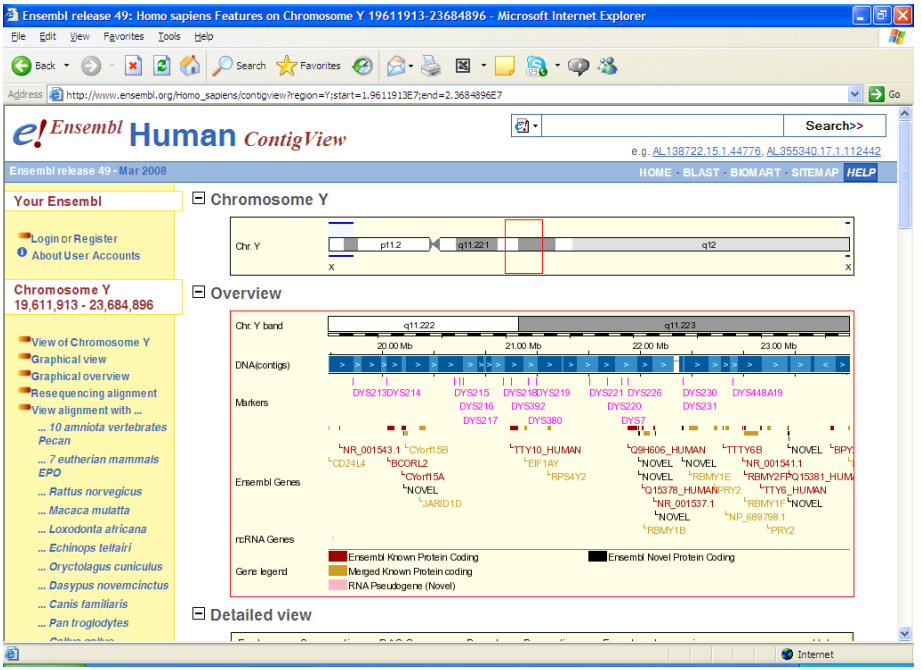
included in the license). Selecting **Show Region** brings up another menu with several options on where to view this region.



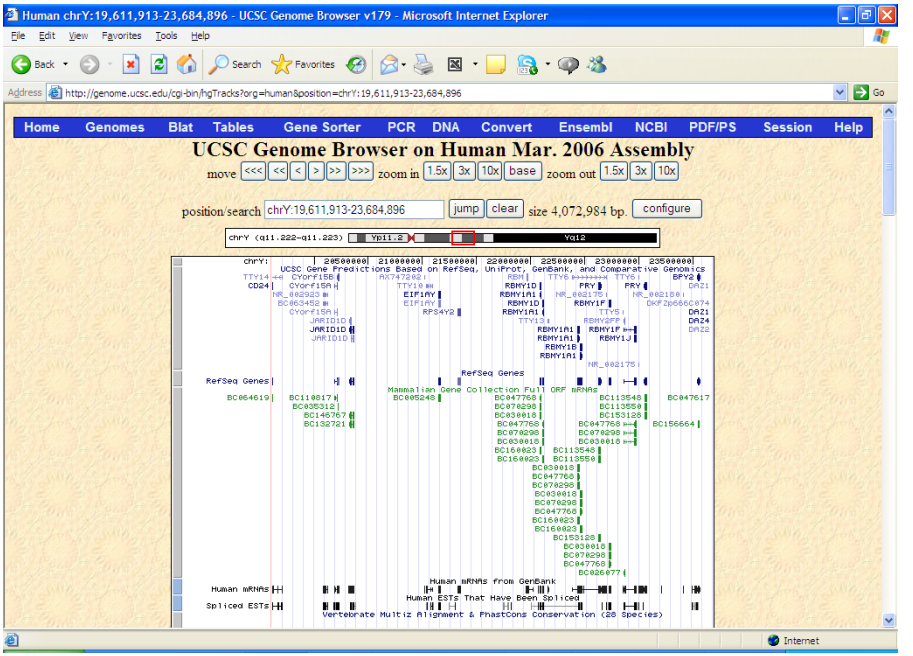
Selecting **Nexus** takes you to the zoomed in view of the region in the **Chromosome** page as seen in the figure below.



Selecting **ENSEMBL** opens up a browser window to the Ensembl Contig Viewer as shown below.

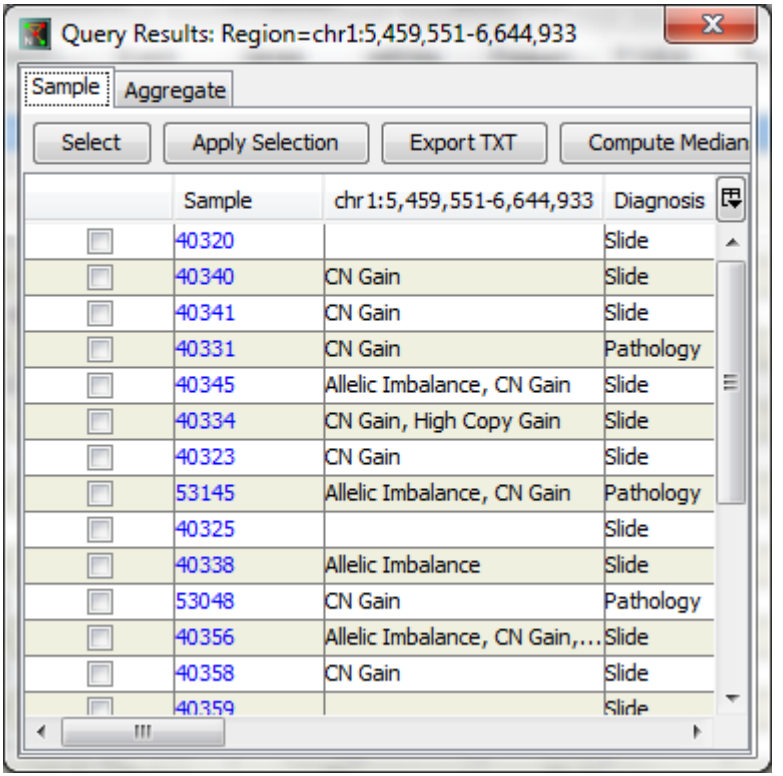


Selecting **UCSC**, opens up a browser window with the UCSC Genome Browser zoomed in on this region (see figure below).



Selecting **Copy** copies the region name (e.g. chr1:11,778,627-12,097,833) to the clipboard; now it can be pasted anywhere you would like.

Selecting **Query all processed samples in project** uses this region to query via the Query Tool (**Query** button in the **Genome** page) and immediately displays a window containing the query results.



Sample	chr1:5,459,551-6,644,933	Diagnosis
40320		Slide
40340	CN Gain	Slide
40341	CN Gain	Slide
40331	CN Gain	Pathology
40345	Allelic Imbalance, CN Gain	Slide
40334	CN Gain, High Copy Gain	Slide
40323	CN Gain	Slide
53145	Allelic Imbalance, CN Gain	Pathology
40325		Slide
40338	Allelic Imbalance	Slide
53048	CN Gain	Pathology
40356	Allelic Imbalance, CN Gain,...	Slide
40358	CN Gain	Slide
40359		Slide

Selecting **Query database** searches the Nexus DB repository. If the user is not already logged in, the login screen will first appear and then the results will be displayed.

Tools

Immediately after processing (after the **View** button is clicked on the **Data Set** tab), only the **Export TXT, Participation,** and **Significant Peaks** buttons are active in the **Aggregate** page. The **View Annotations** buttons is unavailable until some data is selected in the table.

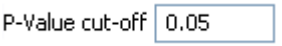
Export TXT button: The data in this table can be saved in an external file. Upon clicking the button, a **Save** window opens up asking you to specify the file name and location. The data is saved in a tab delimited text file.

View Annotations button: Annotations including gene name, description, and GO classification for a single region or multiple regions can be seen in a new window by highlighting some rows and clicking on the **View Annotations** button. Please see the *Annotation Drill Down* section, page 175, for more information.

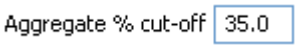
Participation button: This tool displays a new window with a table specifying how each individual sample's aberration overlaps with the aggregate aberrant region. Please see the section on *Participation Table* for more information.

Significant Peaks button**:** Used to establish significance of the copy number aberrations and allelic events in a region. Once a **P-Value** and **Aggregate % Cut-off** are specified, click on this button to gain significance information. This operation uses the STAC algorithm. See section on *STAC, page 140*, for more information on significance testing using STAC.

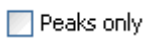
P-Value cut off: This input box is used when performing significance testing. See section on *STAC, page 140*, for more information on significance testing using STAC.



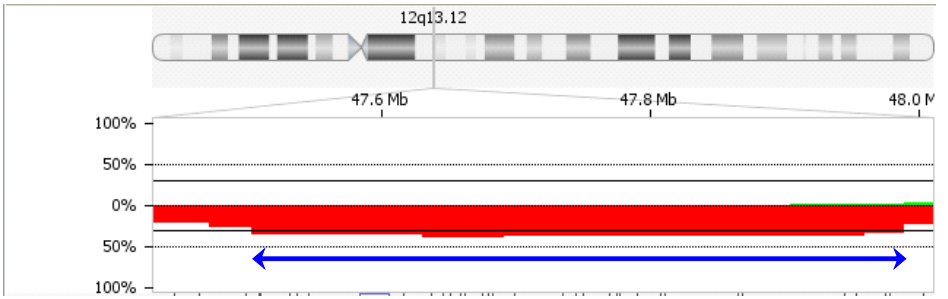
Aggregate % cut-off: The value in this input box is used to limit the results in the **Aggregate** page to only those regions of aberrations with frequencies equaling or exceeding that defined in the box. Specifying a value here will list only those aberrations that equal or exceed this cut-off value.



Peaks only: Marking this checkbox provides a more fine-tuned list of aberrations. With this option selected, the regions of aberrations listed in the **Aggregate** page consist of smaller regions above the **Aggregate % cut-off** value that are considered peaks. The **Aggregate** table will now list the smallest regions with the highest frequencies above the % cut-off line.

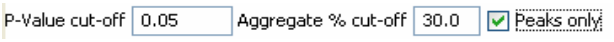


In the figure below, without **Peaks only** selected, chromosome 12 shows a loss of 484,960 bases (above the 30% **Aggregate % cut-off** threshold) indicated on the plot with the blue arrow.

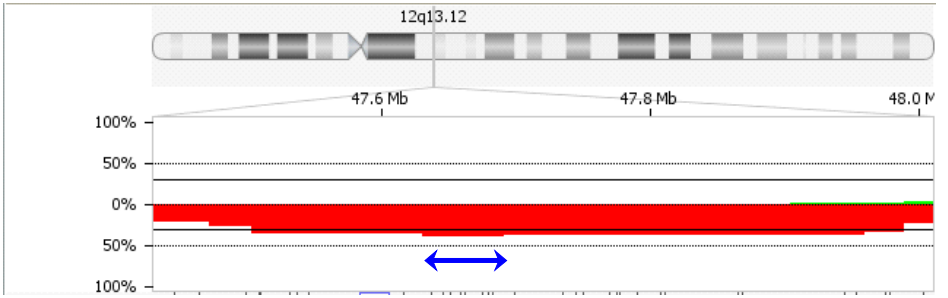


chr10:77,020,223-88,759,195	11,738,972	Loss	30.189
chr10:89,456,155-135,217,395	45,761,241	Loss	30.189
chr12:47,504,292-47,989,252	484,960	Loss	33.962
chr12:55,149,816-55,684,468	534,653	Loss	32.075

With the **Peaks only** checkbox selected



Chromosome 12 shows a region of only 61,252 bases indicated by the blue arrow.



chr10:120,816,409-120,937,196	120,787	Loss	41.509
chr10:134,991,656-135,190,856	199,200	Loss	37.736
chr12:47,630,825-47,692,077	61,252	Loss	39.623
chr12:55,402,061-55,676,753	274,692	Loss	43.396

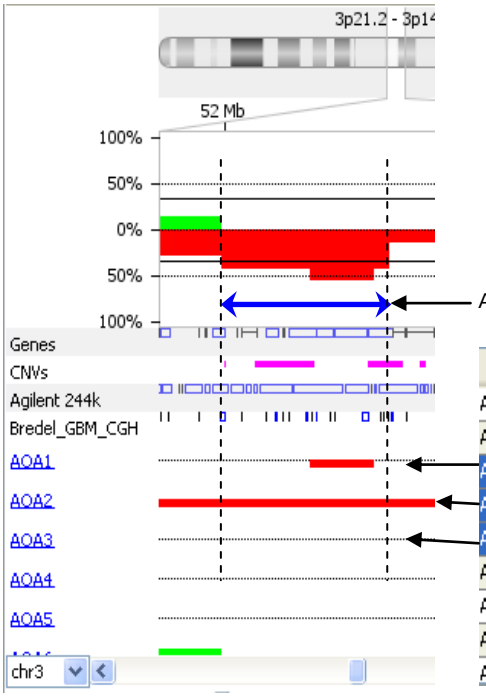
PARTICIPATION TABLE

This table is generated by clicking on the **Participation** button in the **Aggregate** page and specifies how each individual sample's aberration overlaps with the aggregate aberrant region. If an individual sample's gain, loss, allelic imbalance, or allelic loss overlaps completely with the aggregate region, then it is labeled as **Complete**. If an individual sample does not have an event in this region, then it is labeled as **Absent**. If the individual sample's aberration overlaps with the aggregate region only partially, then the **Participation** is labeled as **Partial**. The table contains four columns: **Region**, **Event**, **Sample**, and **Participation**.

Column	Description
Region	Chromosome and base pair coordinates of the segment.
Event	Possible values are **Gain**, **Loss, Allelic Imbalance, Allelic Loss**
Sample	Name given to sample in Sample Descriptor file
Participation	Possible values are **Complete**, **Absent**, **Partial**
Probes Median	Probe medians for the segment. This column is only present if medians were computed by clicking the **Compute Medians** button. This applies to log ratio values (gains and losses only). Since allelic events (Allelic Loss, LOH, Total Allelic Loss) are not based on log ratios, probe median values for these events will be blank or NA.

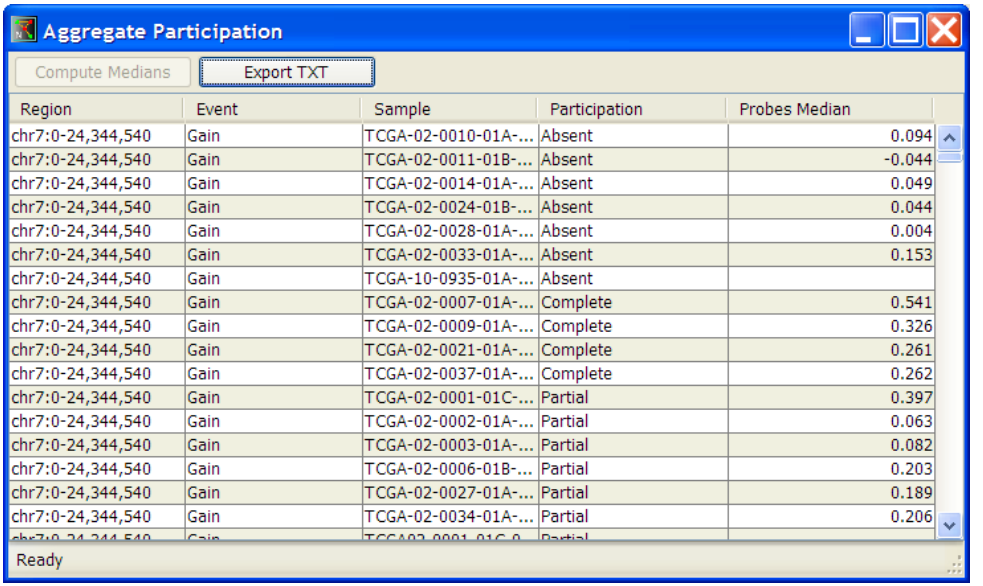
The figure below shows a region of loss on chromosome three. The Chromosome view is depicted on the left with the corresponding region and samples highlighted in the **Aggregate Participation** table on the right. The aggregate region computed is 51,988,363-52,832,112 bp (depicted by the blue arrow and vertical dashed black lines on the Chromosome view). It can be seen that sample AOA1's region of loss only overlaps with the aggregate partially (Participation is **Partial**). Sample AOA1's region of

loss completely covers the aggregate region (Participation is **Complete**). Sample AOA3 does not have a loss in this region (Participation is **Absent**).



Sample	Participation	Event	Region
AOA6	Absent	Loss	chr3:47,545,433-48
AOA7	Complete	Loss	chr3:47,545,433-48
AOA1	Partial	Loss	chr3:51,988,363-52
AOA2	Complete	Loss	chr3:51,988,363-52
AOA3	Absent	Loss	chr3:51,988,363-52
AOA4	Absent	Loss	chr3:51,988,363-52
AOA5	Absent	Loss	chr3:51,988,363-52
AOA6	Complete	Loss	chr3:51,988,363-52
AOA7	Complete	Loss	chr3:51,988,363-52

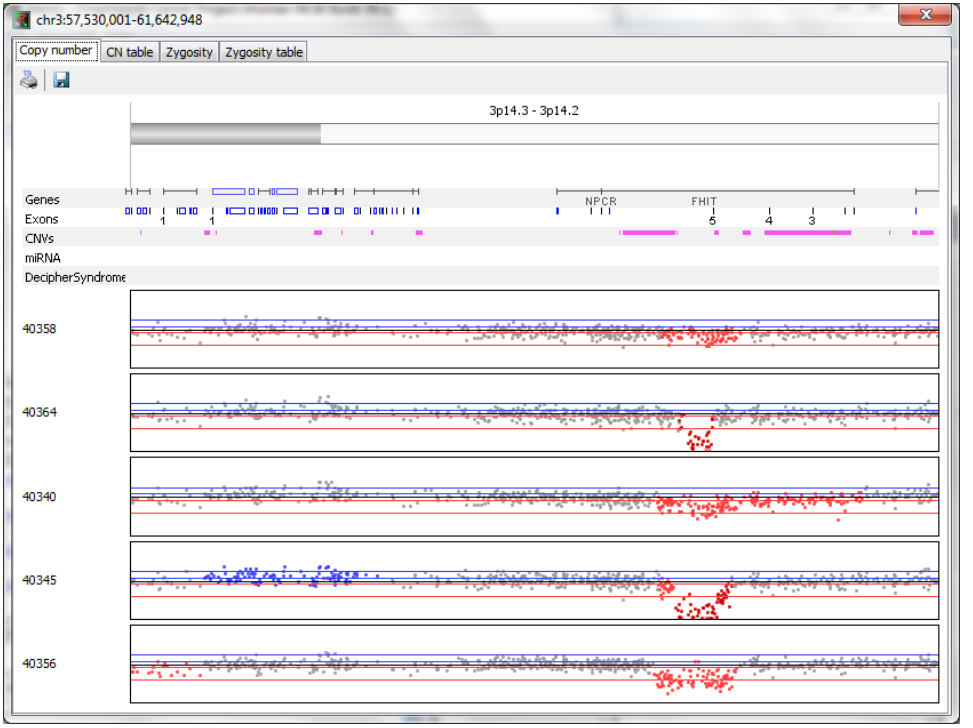
Data from this table can be exported in a tab delimited text format by clicking on the **Export Text** button. The **Compute Medians** button will compute and add to the table the median value of the probes for each region and event in each sample.



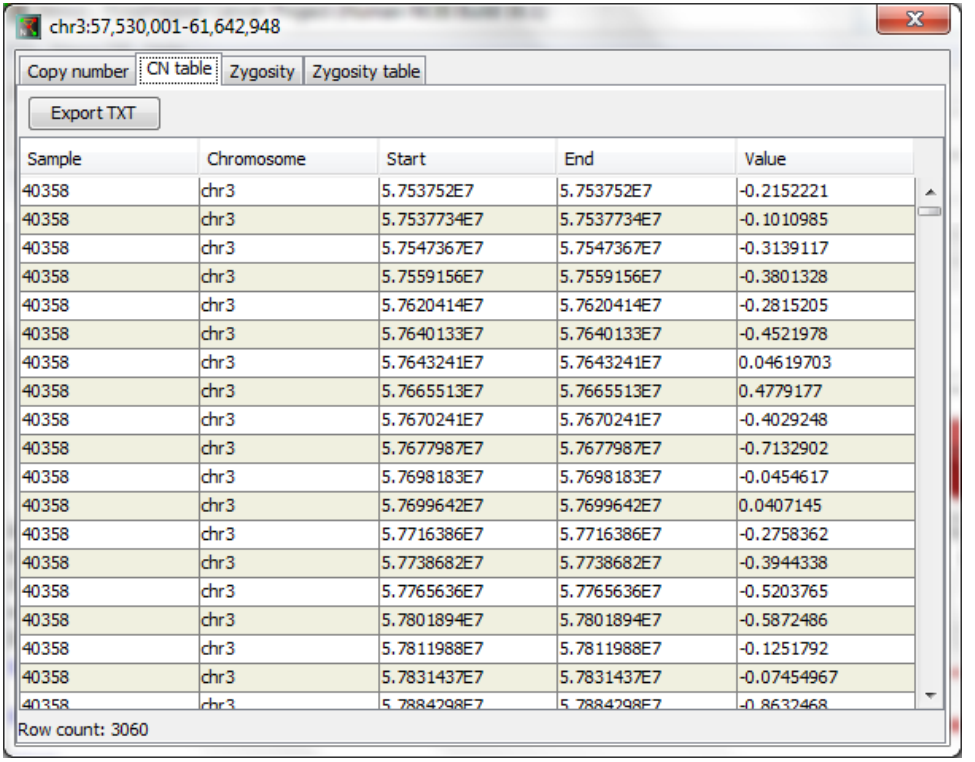
MULTI-SAMPLE PROBE VIEW

The multi-sample probes view displays in one window the scatter plots of multiple user-selected samples. Select the samples you want to view in the **Data Set** tab, zoom into a region and click the **Show Probes** button in the **Chromosome** tab. Two tabs will be displayed, one for log ratio and one for B-allele frequency. Please note that only the samples that are visible in the samples tracks will be displayed in the multi-probe view. For the copy number display, red indicates loss, and blue indicates gains. For the B-allele view, brown indicates LOH, and purple allelic indicates imbalance.

Copy Number tab showing probes for copy number change events:

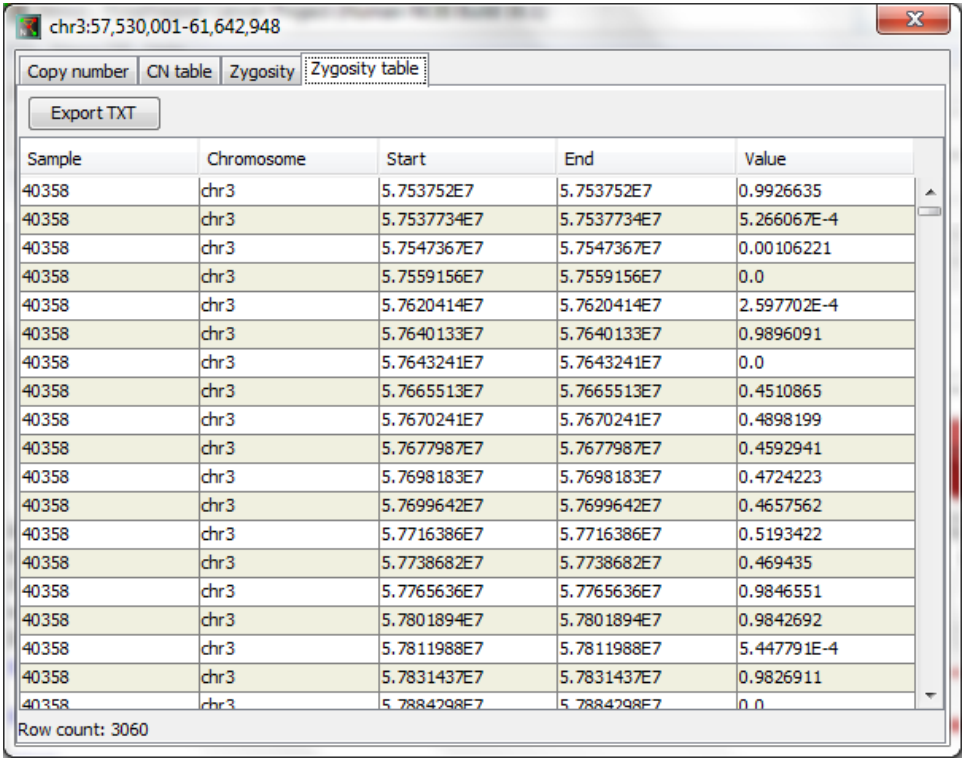


CN table tab listing log ratio values for the probes:



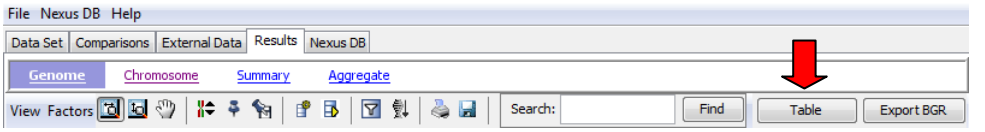
Zygosity tab showing probes for allelic events:

Zygosity table tab showing B-allele frequency values for the probes:



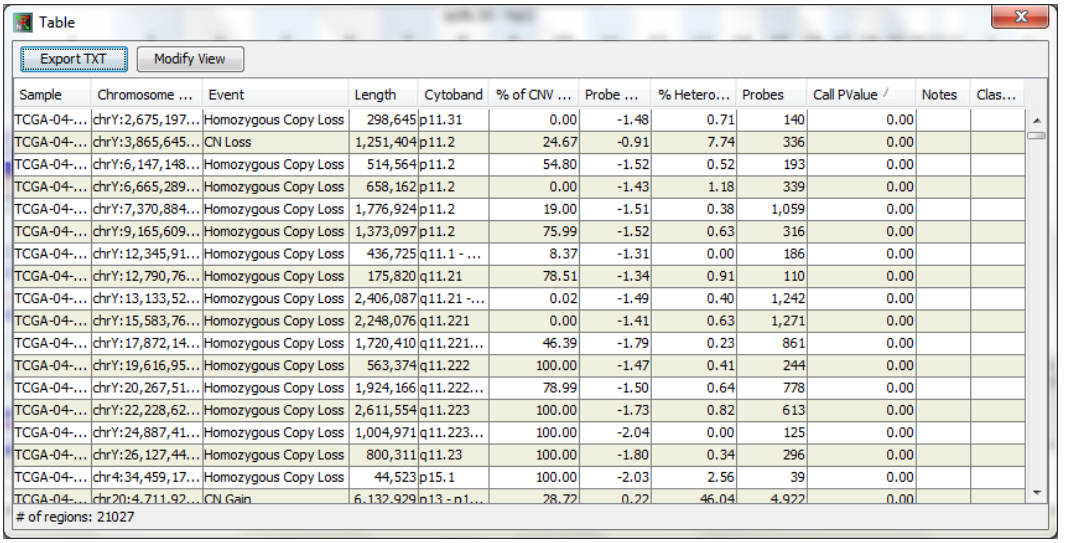
All Samples and Aberrations Table

The **Table** window lists the samples along with the segmented regions defined by the segmentation algorithm and whether there is a copy number or allelic event in that region. This table is generated by clicking on the **Table** button in the **Genome** page.



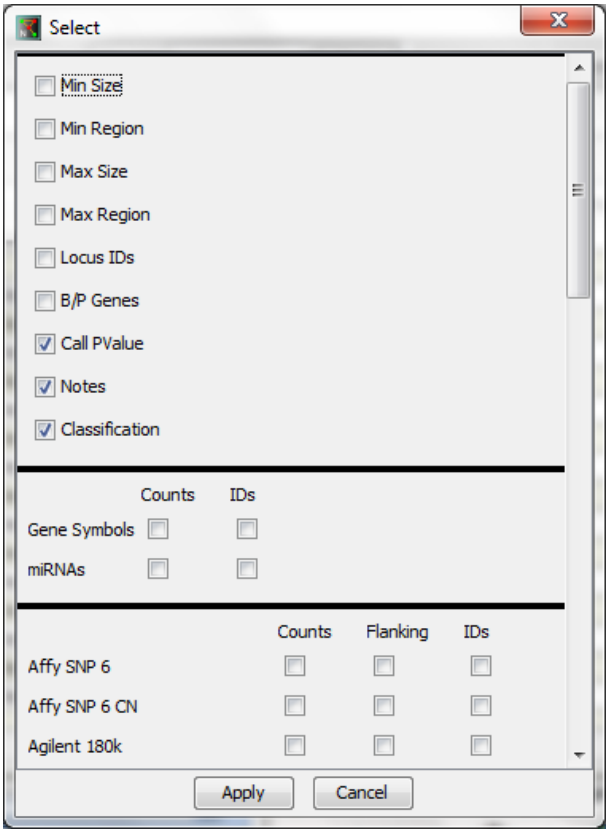
Column	Description
Sample	Name given to the sample in the Sample Descriptor file
Chromosome Region	Chromosome along with a base pair region in the following format: chr8:172,199-300,002
Event	Possible values are **CN Gain**, **CN Loss**, **High Copy Gain**, **Homozygous Loss**, **LOH**, **Allelic Imbalance**, **Total Allelic Loss** and **Unchanged**
Length	Length of the region
Genes	Number of genes in this region
Cytoband	Which cytoband this region covers
% of CNV Overlap	Percentage of this region covered with CNVs
Probe Median	Median value of the probes in the segment
% Heterozygous	Percentage of probes lying outside the Homozygous Value Threshold – yellow lines in the plot. Applicable only to SNP arrays.
B/P Genes	Breakpoint genes (genes that are only partially covered by the region – possible fusion sites)
Call PValue	Significance of obtaining this call at this location (one-tailed z-test) - the probability of obtaining the observed mean of the probes encompassing the call segment assuming the true mean is 0 and the distribution is normal. The value is corrected for multiple testing. If the p-value cannot be calculated for a call (e.g. for a sex chromosome), the value here will be NA.
Notes	Any notes the user wants to enter. This also by default states “manually altered” if a call was added by the user
Classification	The classification entered by the user for this region (e.g. benign, pathogenic, unknown)

Probes	Number of probes in the region
Min Region	The region encompassed by the two most external probes in the segment
Min Size	Length in bp of the Min Region
Max Region	The region encompassed by the closest probes on either end of the segment that are not part of the segment
Max Length	Length in bp of the Max Region



The columns can be sorted and rearranged as described in the *Nexus Express Basics* section under *Data table tabs*. This data can be exported in text format by clicking on the **Export TXT** button on the top. A **Save** window opens up asking you to specify the file name and location. The data is saved as a tab delimited text file.

The columns displayed can be modified by clicking the **Modify View** button. A list of columns is displayed where you can check off what to display or hide.



CALL PVALUE CALCULATION

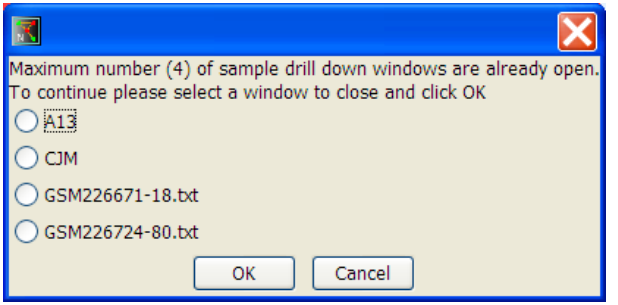
The Call Pvalue in the table is calculated by first taking the variation of the probe measurements over the whole genome. Next for each call, the mean of the probes in the call is computed. Then Nexus Express computes the probability of observing a mean with that many probes given that the true mean is 0 and the probes are normally distributed with the variance calculated earlier. Finally correction for multiple testing is applied by multiplying that probability by the number of places on the genome where one could have that large of a call to create the p-value.

INDIVIDUAL SAMPLE DRILL DOWN

Probe level data for each individual sample is obtained by clicking on the blue hyperlinked sample name or anywhere on the frequency plot of the individual samples in the bottom panel within the **Genome** and **Chromosome** pages of the **Results** tab. This will open up a new window with five tabs-**Sample Info**, **Overview**, **Whole Genome**, **Chromosome**, and **Summary**.

These tabs contain plots and ideograms that can be saved as image files or can be printed using the **Print** and **Save** icons at the top.

A maximum of four pop up drill down windows can be open at the same time. This will allow you to place samples side by side to do a visual comparison. If opening an additional drill down window is attempted, Nexus Express will bring up an alert box informing you that the maximum number of windows is open. It will list the samples and give you the option of closing one of them in order to open another one (see figure below).



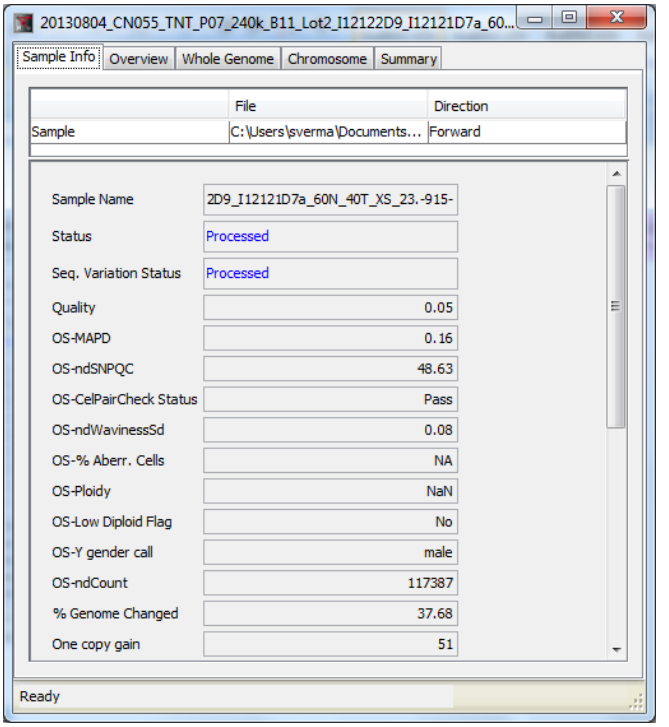
The **Zoom** tool is available in the **Whole Genome** and **Chromosome** tabs. You can zoom in on a section by clicking on the zoom tool, then clicking on the plot and holding and dragging to mark the area you want magnified and then letting go. In the **Chromosome**

tab, you can also zoom in by clicking on the zoom tool and then clicking and dragging on the chromosome ideogram. See the **Chromosome** tab section below for a figure showing how to zoom in on the ideogram. In both tabs over the ideogram or plot regions, right clicking zooms out and left clicking the mouse zooms in. Double clicking the right mouse button zooms out completely.

In the **Whole Genome** and **Chromosome** plots, if any log ratio values fall outside the default plot bounds, these values will be represented as a small "x" mark at the left or right edge of the plot area.

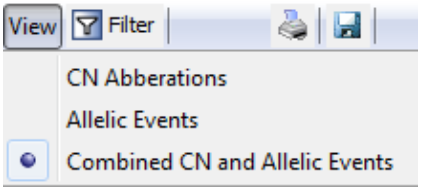
SAMPLE INFO TAB

The **Sample Info** tab conveniently displays information including quality metrics and factors from the Data Set table for the sample. The **File** field shows where the input file is located on the drive and the **Direction** field shows the array "orientation" as "Forward."

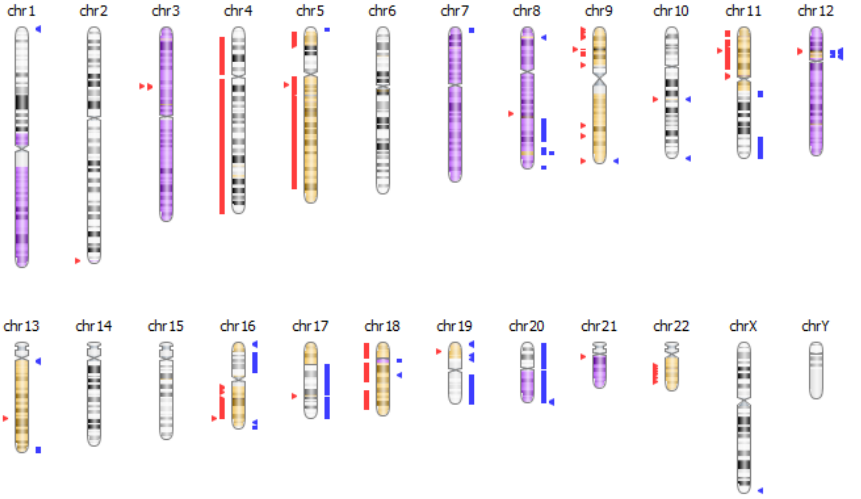


OVERVIEW TAB

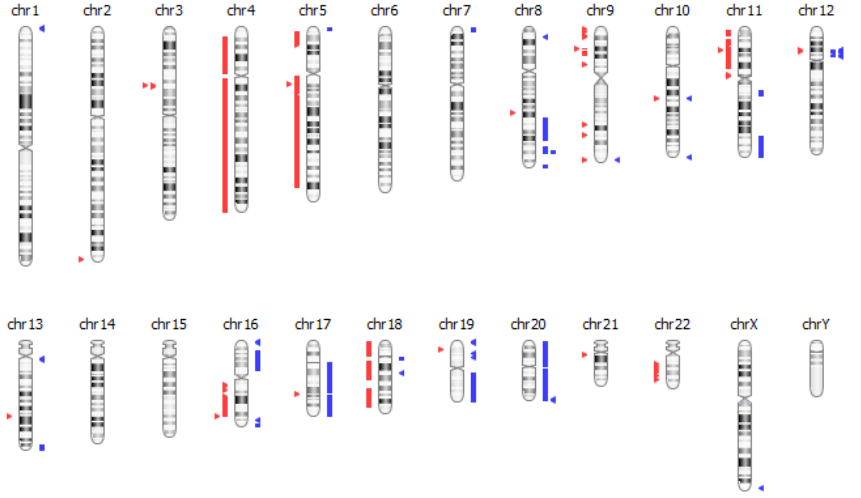
The **Overview** tab displays the organism's chromosomes, depending on the organism selected when creating the Nexus Express Project. The **View** button allows display of copy number events or allelic events.



The Combined CN and Allelic Events view displays both types of aberrations in a single view by coloring the ideogram in light brown/purple to indicate LOH/allelic imbalance and displaying the CN gain/loss next to the ideogram in blue/red.



CN Aberrations display:



Copy number gains are marked by a blue bar/arrow and deletions by a red bar/arrow, to the right and left side of the ideogram respectively.

Allelic Events display:

Allelic imbalance events are marked by a purple bar/arrow and LOH by a brown bar/arrow, to the right and left side of the ideogram respectively.

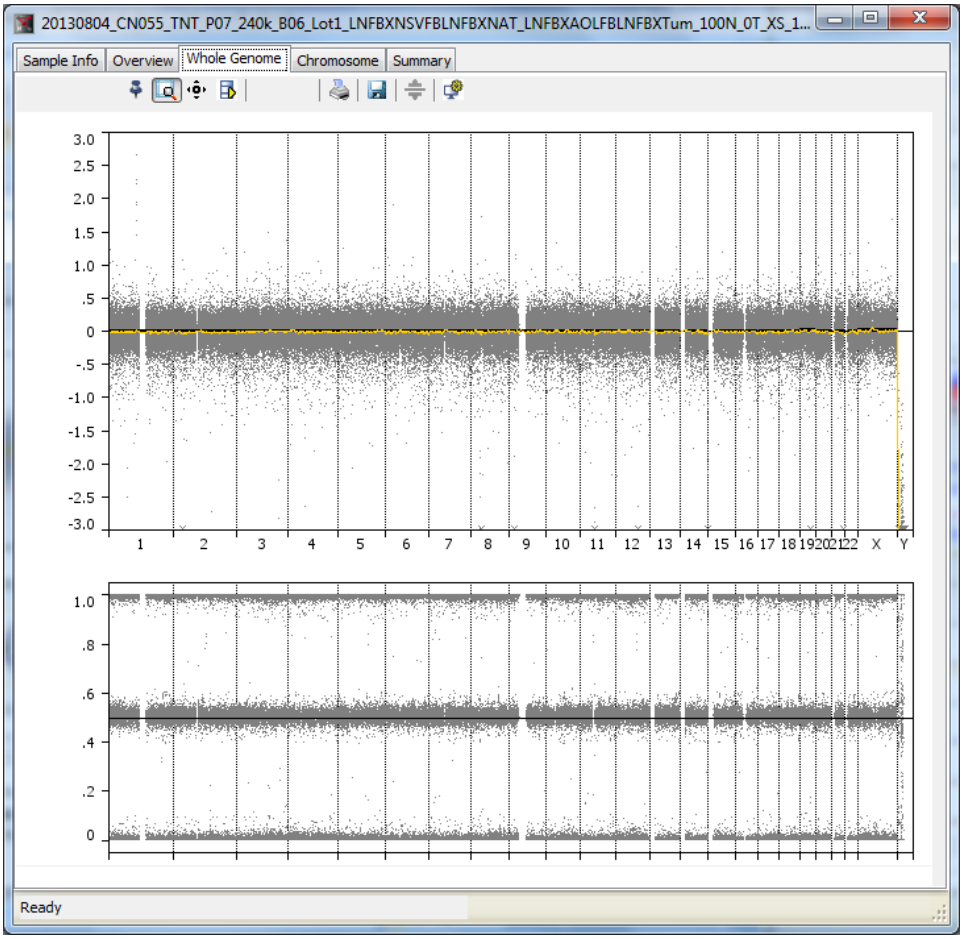
By clicking on any of the ideograms, the selected chromosome will be displayed in the single **Chromosome** view described later in this section. Small events are displayed with an arrow to make it more prominent on the plot.

The **Filter** button is a quick way to open up the Filtering options window (also available from the **Options** menu item):

WHOLE GENOME TAB

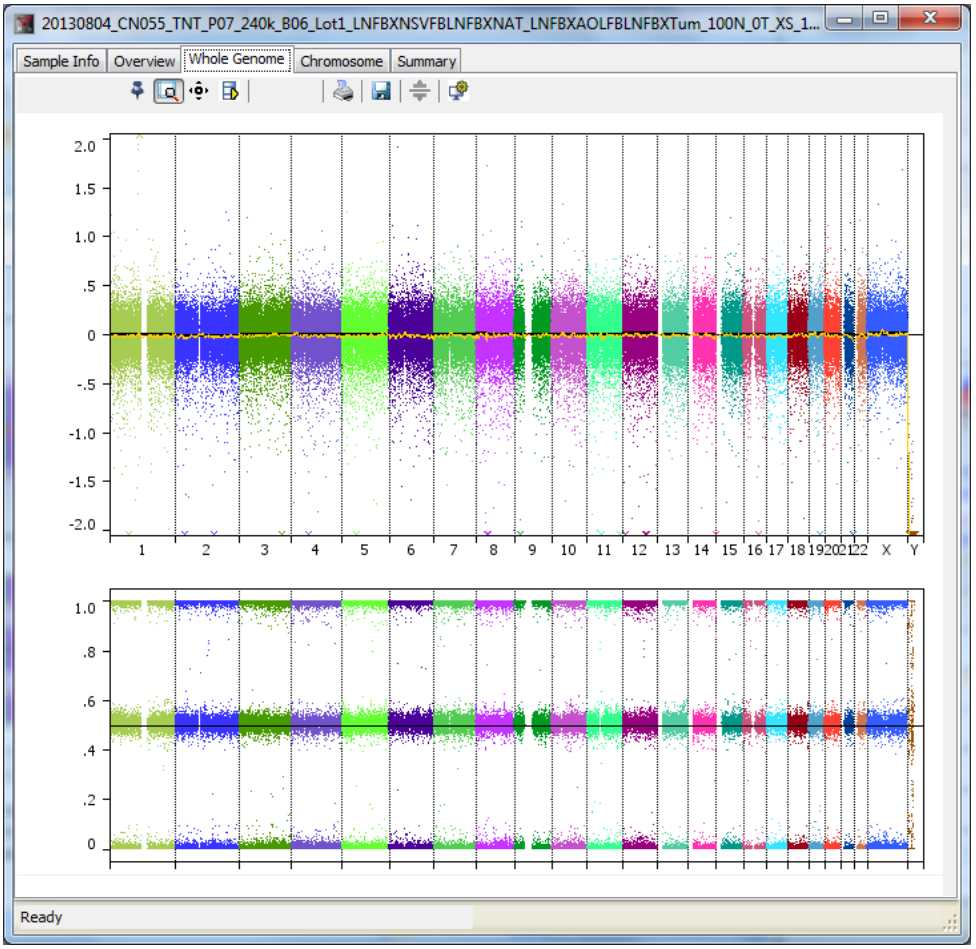
The **Whole Genome** tab displays the entire genome with chromosomes lined end to end as shown below. The top plot shows copy number aberrations and the bottom shows allelic events. The amount of space allocated to each panel can be adjusted via **File->Options->Display Options**. Increasing the percentage under "SNP probe plot area vs. Copy Number probe area height" will increase the vertical space taken up by this plot and decrease that of the log ratio plot at the top. The **Segment View Baseline** adjusts the baseline from the default of 0.

In the top plot (log ratio), each gray dot corresponds to a probe on the array. If the same probe is spotted in multiple locations (replicates) and if the **Combine Replicates** box is checked off in the **Settings** options, the dots shown in the plots below are the combined value of all the replicates. Note that **Combine Replicates** is not available for the Affymetrix OncoScan TuScan data type. If **Combine Replicates** is not checked off, then each dot represents an individual probe's value. The yellow line in the plot depicts a moving average value and the black lines are the output of the calling algorithm. The "call" thresholds for single and higher copy gain or loss are shown as blue and red horizontal lines, respectively in the copy number plots for SNP-FASST2 processed data. For the B allele frequency plot on the bottom, the brown and purple lines mark the thresholds for LOH and allelic imbalance regions, respectively visible only for SNP-FASST2 processed data.

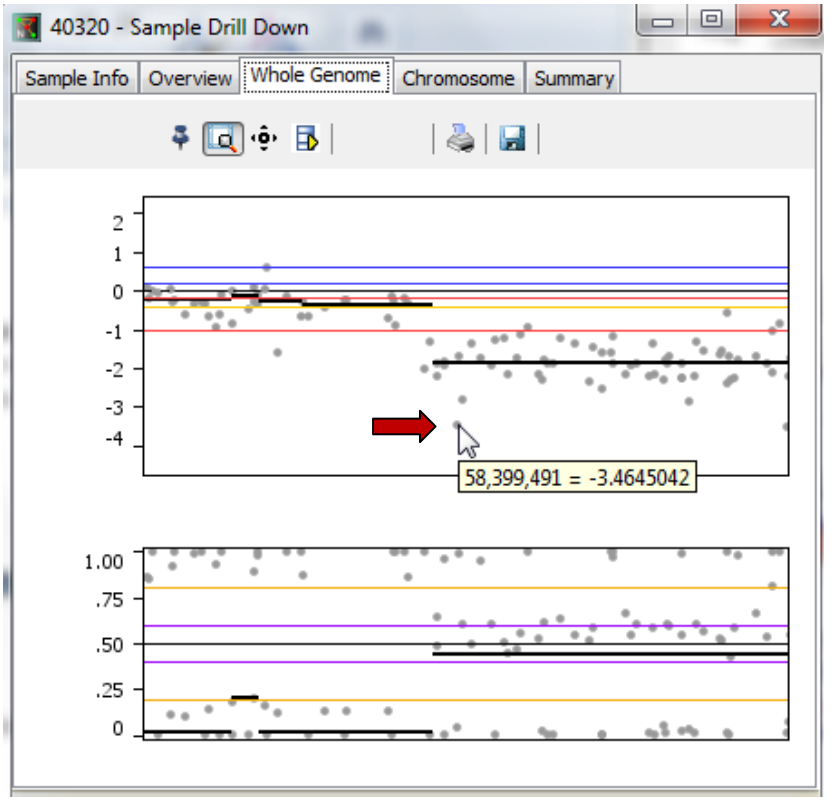


The color scheme of the plot (probes, moving average line, and segment color) can be adjusted from the defaults via **File->Options->Display Options**. Just click on the colored square and select a new color. Click on the **Restore Defaults** button to return these colors to the default.

Each chromosome can be colored in a different color. This option is available via the **Options->Display Options** as the **Genome Probes Rainbow** checkbox. Checking the box provides a more colorful whole genome view:



Moving the mouse over a probe displays the bp location and either log ratio intensity or B-allele frequency depending on which plot is used. In the figure below the location and log ratio intensity is displayed when the mouse is over a probe in the log ratio plot (near red arrow).



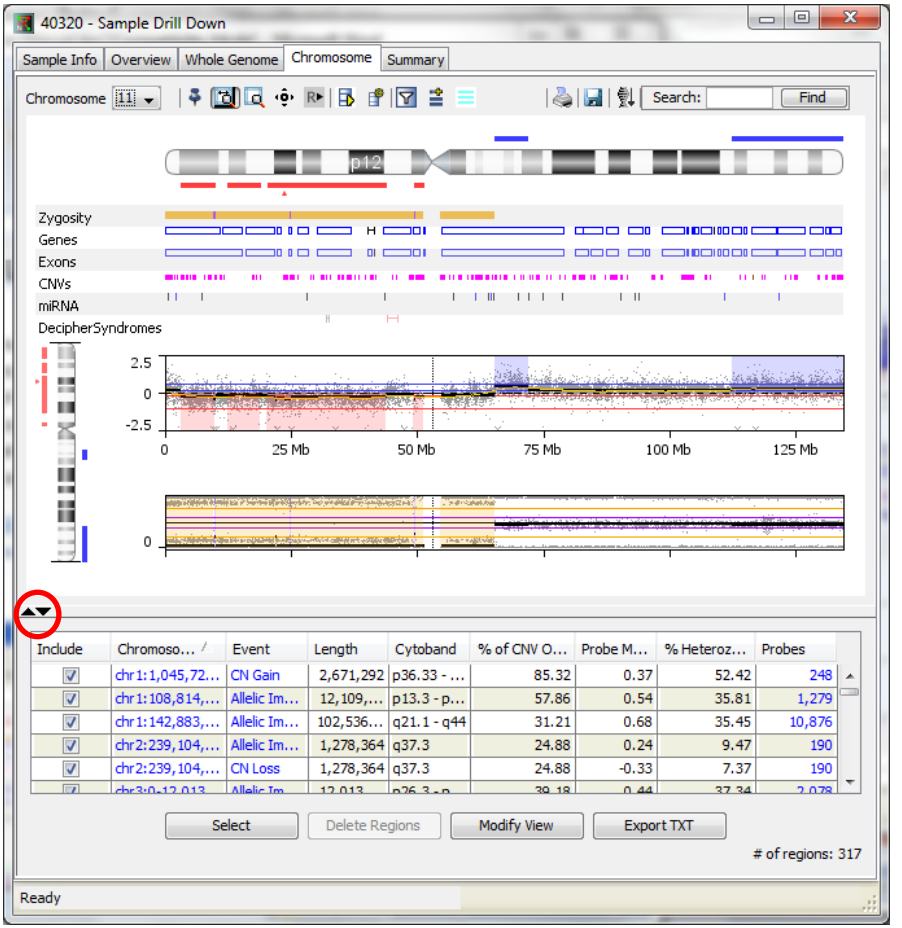
Mark Tool: This tool allows marking a location with a vertical gray line running through the plots. This line will persist throughout other windows during the current Nexus Express session but will be removed when the project is opened next.

Recentering tool: Using this tool and clicking anywhere on the plot will re-center the plot around that point.

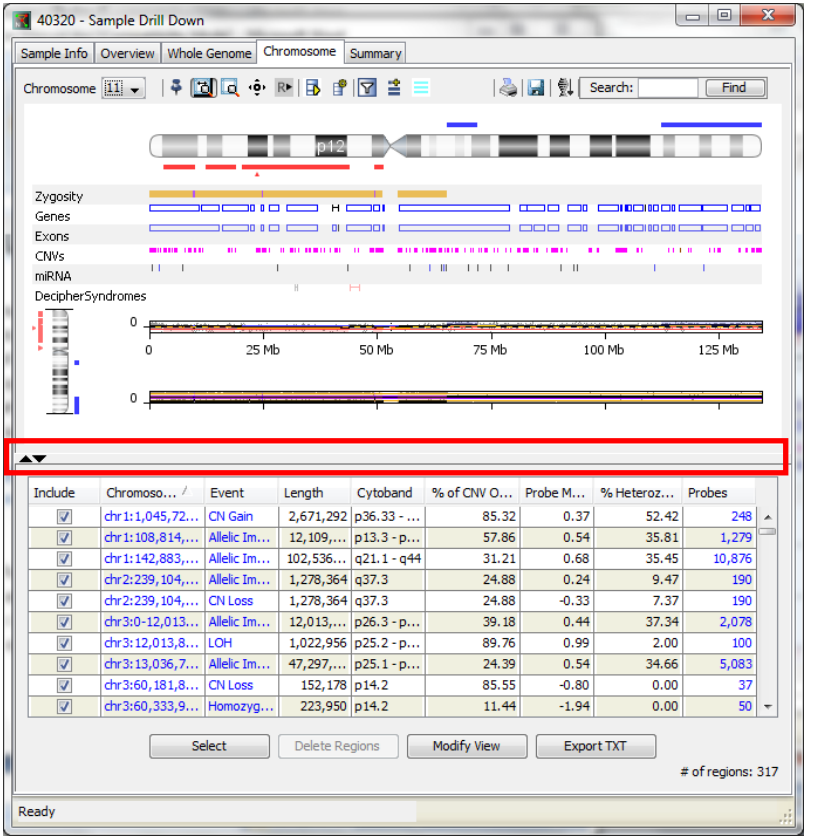
Query Database tool: May be present if the software has a license to Nexus DB. It allows querying the repository for region aberrations. See the section on *Nexus DB* for more details on this tool.

CHROMOSOME TAB

The **Chromosome** tab enables the user to zoom in and view individual chromosomes. It displays both the graphical as well as tabular data for the aberrations on each chromosome. The window is split horizontally and provides the chromosome view in the top portion and the data table in the bottom portion. The split window is resizable to vary the size of the split views and to show only the chromosome view or table view using the arrows circled in red in the figure below.



The chromosome ideogram is on the top followed by the annotation tracks, and then the copy number aberration plot, followed by the B Allele Frequency plot , and lastly the report table listing each aberration in the chromosome. The arrows on top of the tabular data pane (circled in red above) can be used to hide either the graphical view or the tabular view such that the full window space is taken by the data table or graphical display. The size of each pane can be adjusted by clicking and dragging the pane handle (indicated with the red rectangle below). Dragging up will allocate more space to the report table and dragging down will give more space to the graphical display. In the figure below the report table is fully expanded in the combined view.

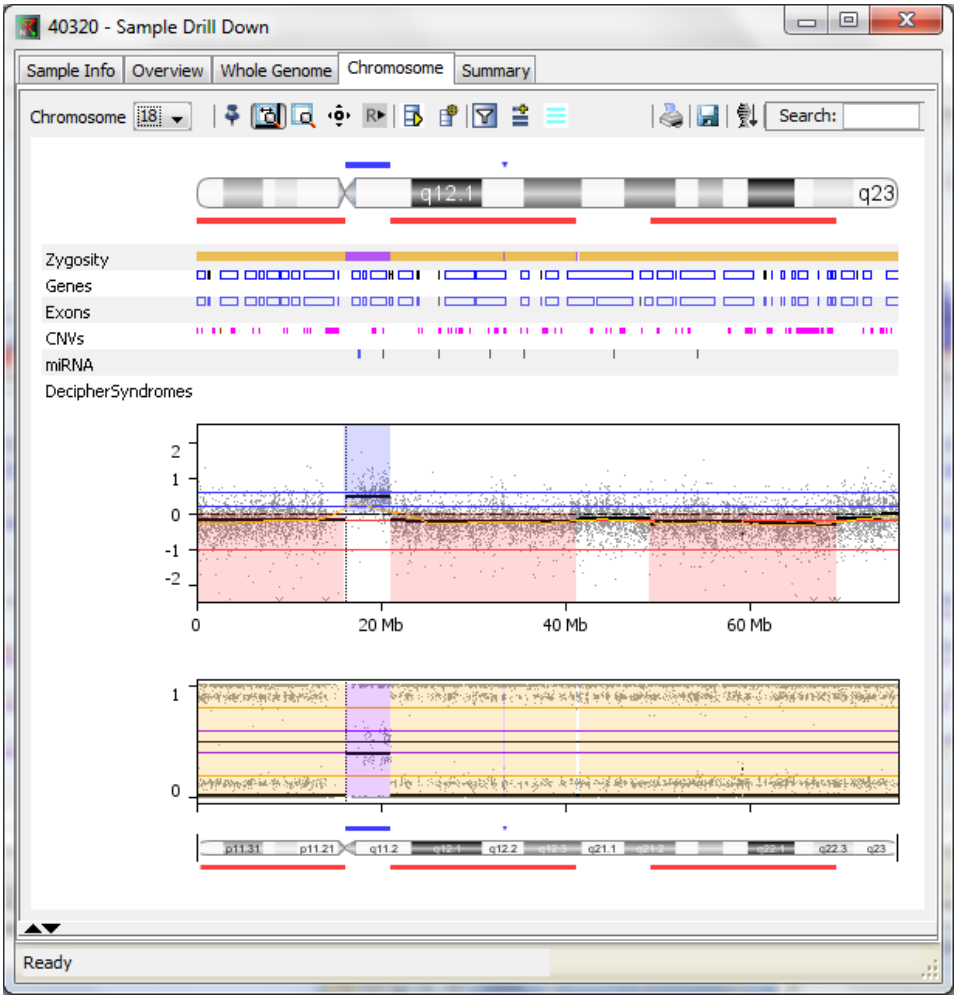


The default space allocation for the report can be defined in the **Options->Display Options** window using the **% of window for repost table in Sample Drill-down Chromosome tab** parameter. If you don't want to see the report table at all whenever a sample drill down window is opened, set this value to 0. Then when the window is opened the report table will be hidden. If you want to view it, you can expand the report window using the horizontal resize handle. If you only want to see the report table and not the graphical displays, set this value to 100. The split can always be resized using the horizontal handle when viewing a sample drill down but each time a new sample drill down window is opened, the report table will only occupy the percentage specified here.

GRAPHICAL DISPLAY

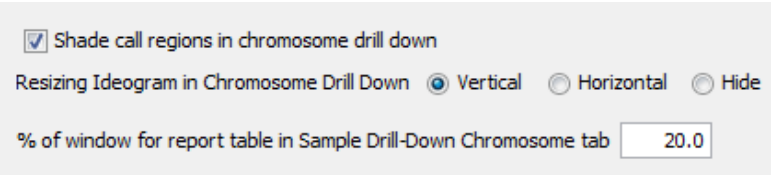
The top pane in the Chromosome tab contains the graphical display with the ideogram, annotation tracks, and frequency plots.

The plots can have shading in the aberrant regions to make the aberrant regions more prominent in the plots. The shading colors match those indicating aberrations (blue – gain, red/pink – loss, purple – allelic imbalance, light brown – LOH). The option to shade the regions is available via **Options->Display Options->Shade call regions in chromosome drill down**.

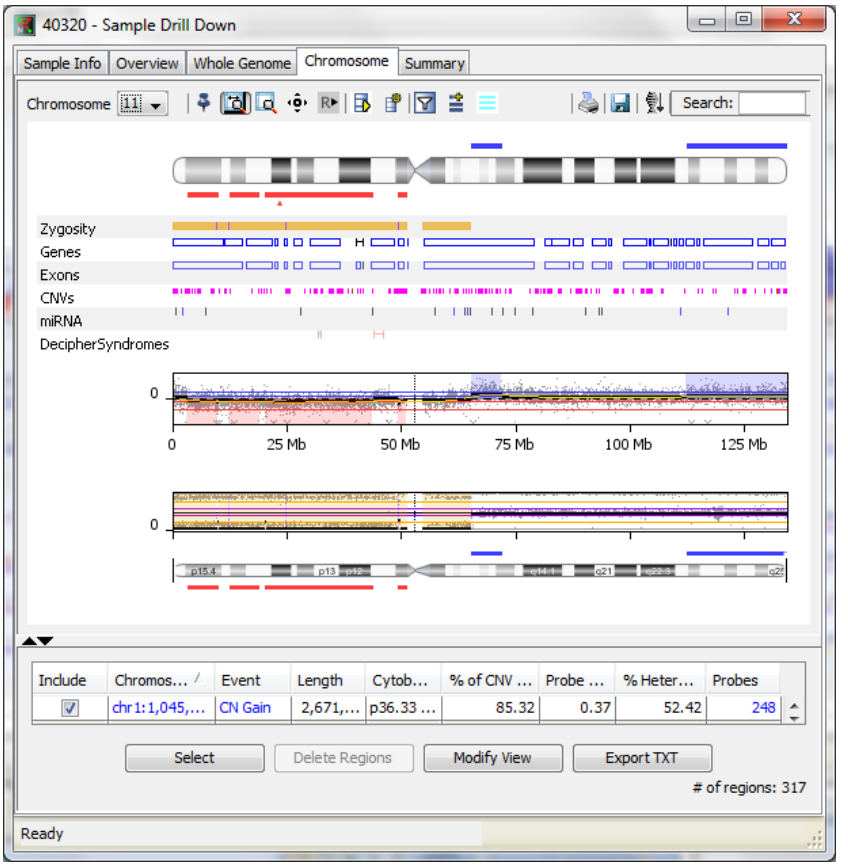


A smaller chromosome ideogram (for panning, adjusting zoomed region, and full chromosome overview) is present to the left of the probes plots. This is the default setting.

This ideogram can be moved to the bottom or removed completely via options in the **Options->Display Options** window:

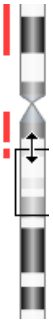


Horizontal ideogram on bottom:



After using the zoom tool to zoom in on the plot or on the horizontal chromosome at the top, the zoomed in region is depicted by a black rectangle (slider) on this smaller vertical (or horizontal) ideogram.

Moving the mouse over the slider brings up the hand icon and activates the panning feature. Clicking and dragging this slider up and down (or left and right) allows panning over the chromosome.



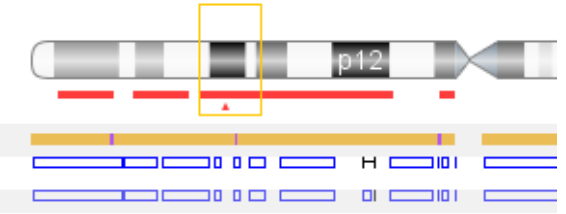
The slider can be resized by dragging the top or bottom (or left or right) edge to zoom in and out from a region. The cursor changes to an arrow when it is over the top or bottom edge to indicate that the resize tool is active.

Selecting the **Locate Region** tool and then clicking on an aberration along the ideogram will highlight the row containing that aberrant region in the data table pane. Clicking on the **Event** value in the data table will zoom in on the corresponding aberrant region in the graphical pane.

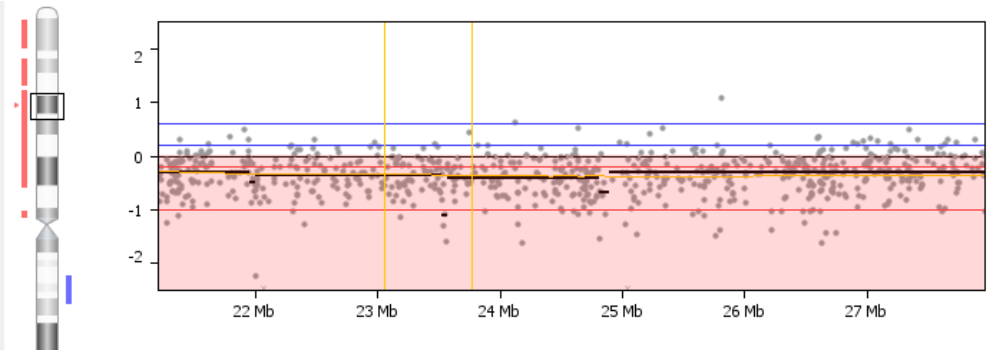
In the top pane, under the ideogram, an annotation track called **Zygosity** is present just under the chromosome. Purple bars in this track indicate areas with an allelic imbalance and brown bars indicate LOH. The copy number aberrations are marked just next to the chromosome image (with red and blue bars).

There are two zoom tools in this window: a horizontal zoom and a rectangular zoom .

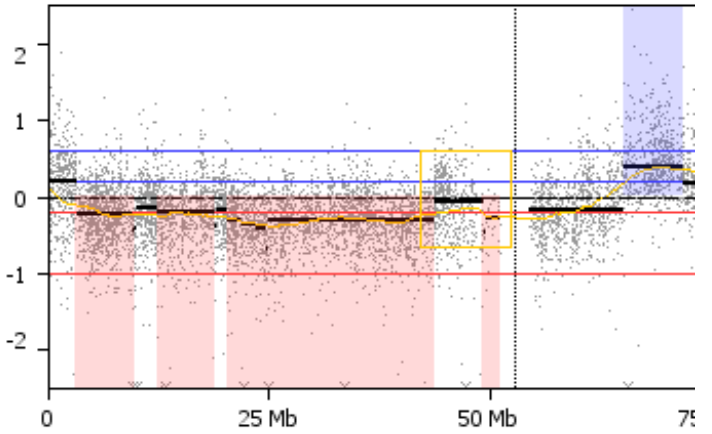
The **Horizontal Zoom** tool can be used to zoom along the x axis only by selecting the area on the plot or on the ideogram. It is easiest to click and drag on the ideogram to select a particular region to zoom into. The figure below depicts the zoom tool being used to zoom in on the larger deletion on the p arm. Clicking and holding down the mouse button on the ideogram and dragging the mouse creates a gold colored rectangle around the region to be zoomed into.



Releasing the mouse will zoom in on the region enclosed by the gold rectangle. On the scatter plot, the tool will create two horizontal lines showing the area that will be zoomed into.

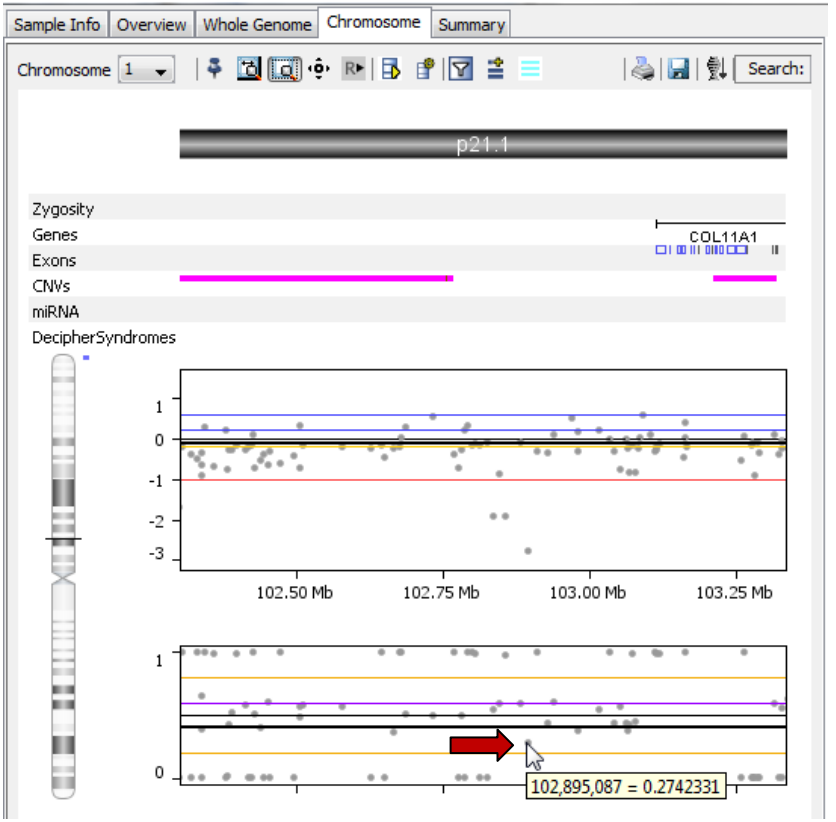


The rectangular zoom tool allows zooming along the X axis when used on the ideogram but zooming in both the X and Y axis when used on the scatter plot. This tool will create a gold rectangle around the area that will be zoomed into on the scatter plot.



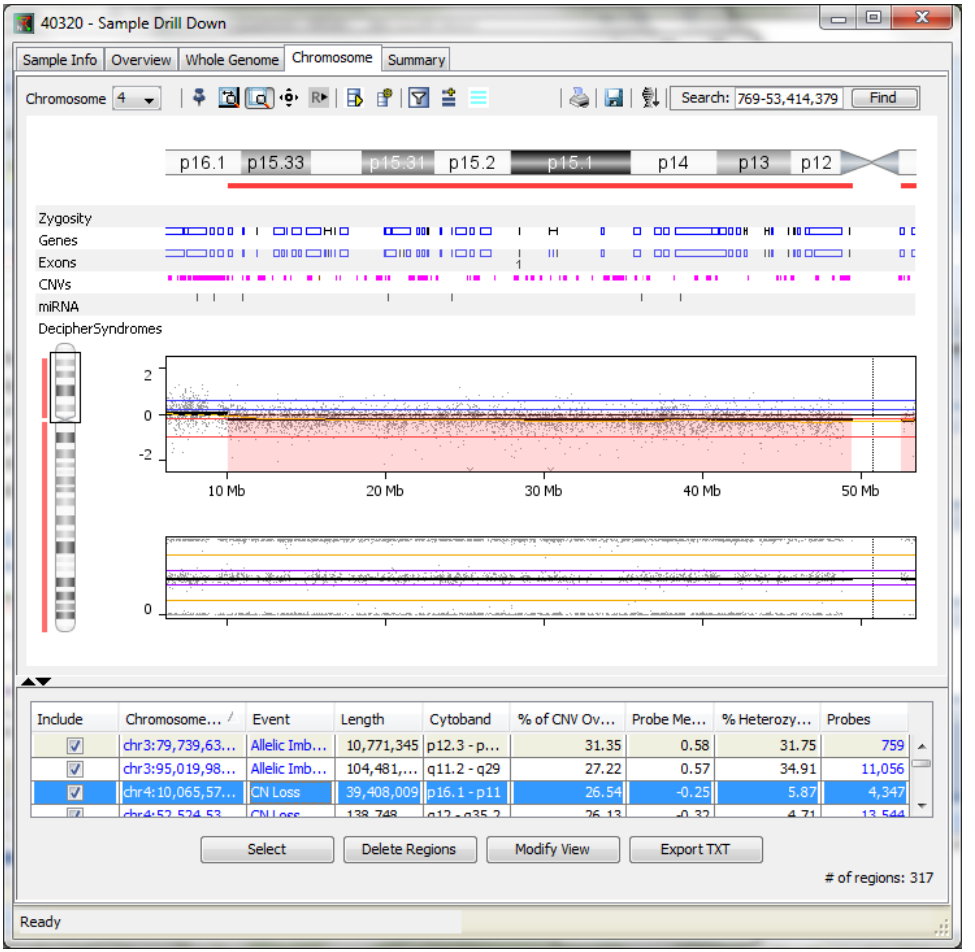
With both zoom tools you can also zoom in incrementally by left clicking on the ideogram or on the plot. Right click repeatedly to zoom out incrementally. Double clicking quickly on the right mouse button will zoom out the view completely.

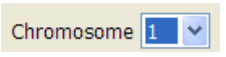
Moving the mouse over the probes in the plots will display the bp location and either the log ratio intensity or B-allele frequency depending on the plot. In the figure below, near the red arrow, the location and B-allele frequency value is displayed when the mouse over a probe in the B-allele frequency plot.



As in the annotation tracks in the **Genome** and **Chromosome** tabs, gene names, CNVs, miRNAs can be clicked and the browser window or pop up dialog will show the appropriate information.

Clicking on an aberration (red, blue, brown, purple bars) on the ideogram highlights the row containing the aberrant region in the report table in the bottom pane, allowing you to obtain details on that aberration (e.g. length, the number of probes in this region, etc.) . For example, clicking on the blue “CN Loss” text in the Event column in the row containing the aberration on the p arm zooms in on this aberration in the top graphical pane:





This drop down is used to select which chromosome to view. This can also be used as “Previous” and “Next” buttons to step through each chromosome in the genome. Highlighting the selected chromosome number and using the up and down arrows, displays the previous and next chromosomes, respectively.

 **Mark** Tool: This tool allows marking a location with a vertical gray line running through the annotations tracks, the log ratio and B-allele frequency plots. It is particularly useful when you find something in the zoomed in view and would like to remember that location when you zoom out or when you open an individual sample drill down. The mark is retained throughout the session and will be removed the next time you open the project. The marks will also appear in the aggregate view.

 **Horizontal Zoom**: Allows you to zoom in along the x-axis on the genome or chromosome frequency plots. This tool is selected by clicking on the **Horizontal Zoom** icon. Once selected, the mouse cursor converts into a magnifying glass when the cursor is over the plots indicating that this tool is active.

 **Rectangular Zoom**: Allows zooming along the X axis when used on the ideogram but zooming in both the X and Y axis when used on the scatter plot. This tool will create a gold rectangle around the area that will be zoomed into on the scatter plot.

 **Centering** tool: Using this tool and clicking anywhere on the plot will re-center the plot around that point.

 **Query Database** tool: It allows querying the repository for region aberrations. See the section on *Nexus DB* for more details on this tool.

 **Region Query Settings**: Allows you to quickly access filtering settings for making region queries against the Nexus DB repository. These settings are also available via **Options->Database Options** and changing the settings using this tool also changes the settings via the **Options** menu.

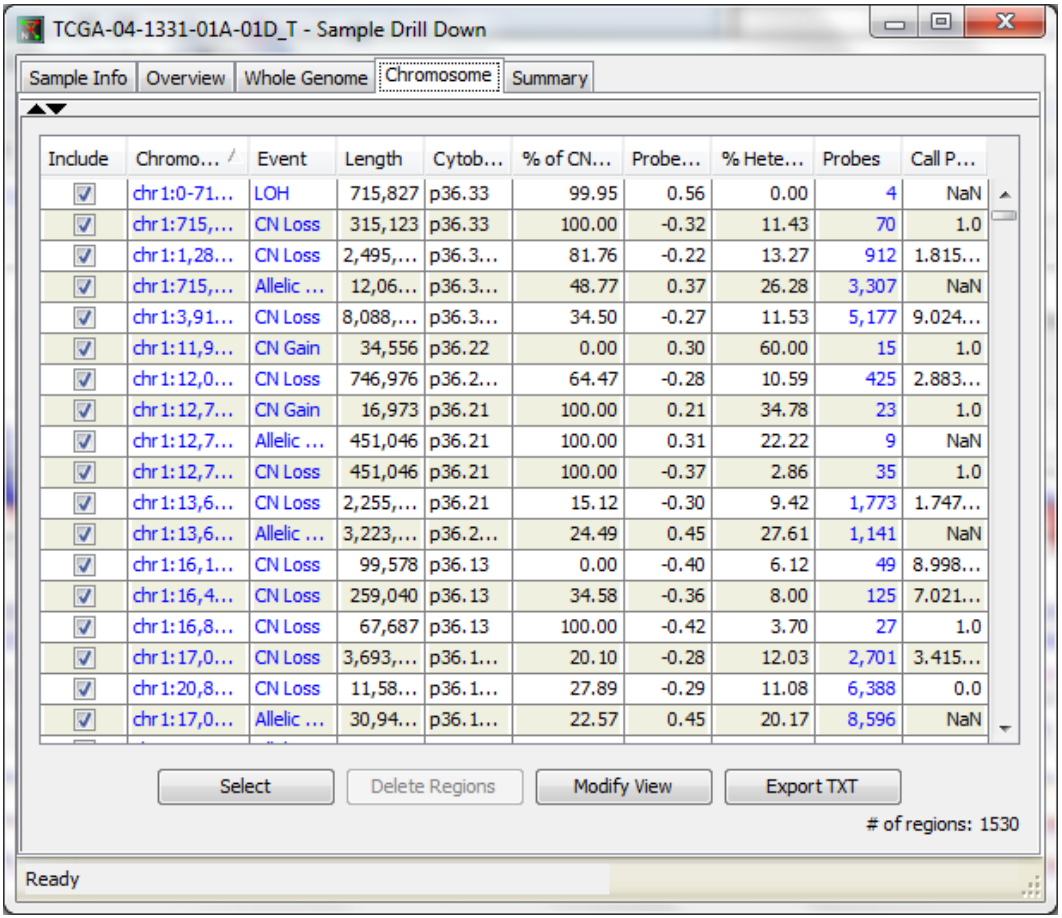
 **Add | Join Region**: Allows manual addition of calls joining of multiple short segments in a region.

 Modify Tracks: Allows you to hide or display specific annotation tracks in the active drill down window. This does not affect selections in the **Options->Track Selection** window.

 **Change Display Settings**: Brings up the display options window allowing you to adjust the display settings for the individual sample drill down. This will not affect selections in the **Options->Display Options** window.

REPORT TABLE

The data table shows numerical data on the aberrations for this sample. The table contains the chromosome and its aberrant region, the Event (CN Gain, CN Loss, High Copy Gain, Homozygous Loss, Allelic Imbalance, LOH, and Total Allelic Loss), Length of the aberrant region, cytoband location, percentage of the region covered with CNVs (% of CNV overlap), number of Probes in that region, probe median, and % Heterozygous. The **Probe Median** is the median value of the probes in the called segment. There is a **% Heterozygous** column which shows the percentage of probes lying outside the Homozygous Value Threshold that is specified in the Settings. Any gene, CNV, or probe that falls partially within an aberrant region will be displayed in this table.

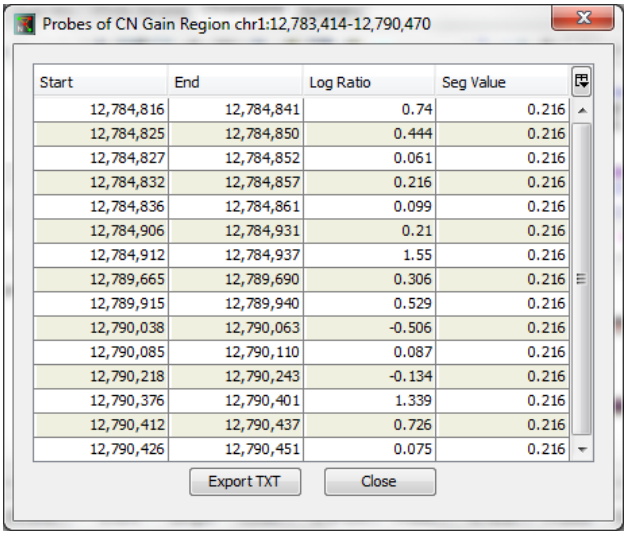


The **Include** column allows you to select which data you want exported so that you don't have to export the whole table and can just select to export some of the data. For example, if you only want to export the high copy gain regions on chromosome 20 for example, first sort on the **Event** column and while holding down the **Ctrl** key, then click on the **Chromosome Region** header to do a secondary sort by chromosome. Now you can easily highlight the High Copy Gains on Chromosome 20 and click on the Select button to mark off the checkboxes. Now just these selected rows will be exported when you click the **Export TXT** button.

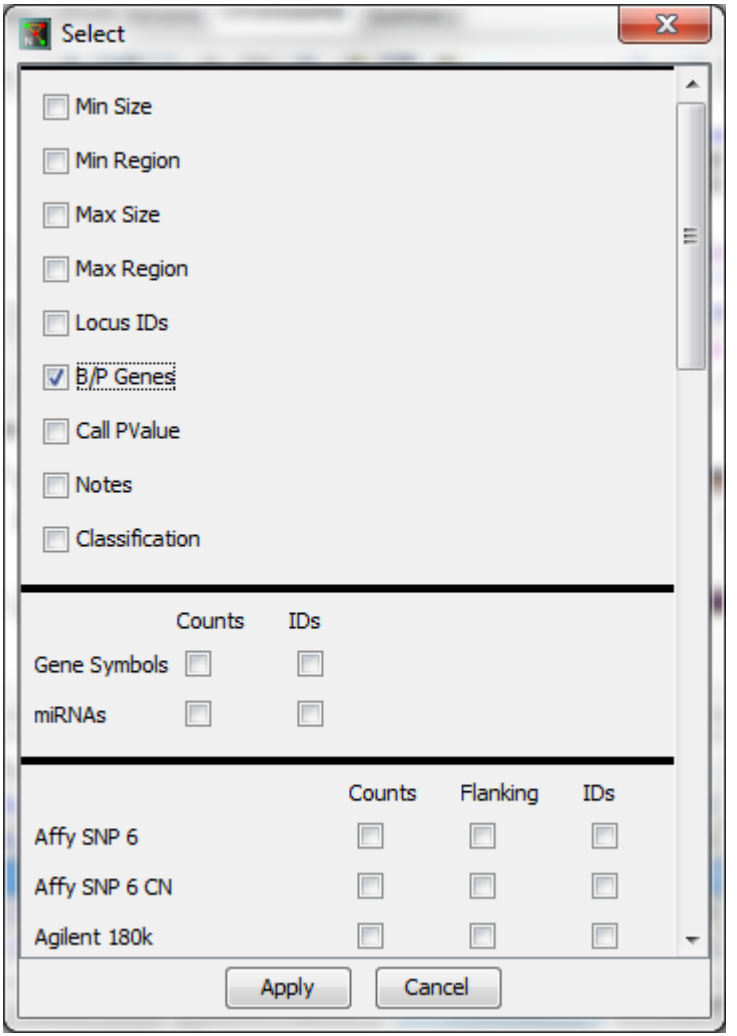
Selecting data rows for export:

Chromosome Region	Event	Len...	Ge...	Cyt...	% of C...	In...	Pr...
chr17:36,970,681-37,239,001	High Copy G...	268,321	11	q21.2	66.968	☐	28
chr17:40,409,058-41,284,023	High Copy G...	874,966	20	q21.31	85.31	☐	70
chr17:70,323,918-70,634,339	High Copy G...	310,422	14	q25.1	4.727	☐	35
chr19:0-3,324,608	High Copy G...	3,324...	112	p13.3	80.168	☐	430
chr19:43,273,039-44,149,288	High Copy G...	876,250	29	q13.1...	18.435	☐	87
chr19:49,834,798-50,058,758	High Copy G...	223,961	7	q13.3...	29.538	☐	26
chr19:53,459,456-55,770,399	High Copy G...	2,310...	127	q13.3...	87.377	☐	258
chr19:60,238,096-60,846,105	High Copy G...	608,010	32	q13.42	26.957	☐	89
chr20:17,093,425-17,558,362	High Copy G...	464,938	4	p12.1	17.255	☑	76
chr20:43,348,330-43,507,436	High Copy G...	159,107	8	q13.12	0	☑	22
chr20:46,732,112-47,091,289	High Copy G...	359,178	2	q13.13	0.435	☑	39
chr20:59,585,011-60,211,658	High Copy G...	626,648	6	q13.33	40.456	☑	98
chr20:60,211,658-60,575,726	High Copy G...	364,069	10	q13.33	99.091	☑	46
chr20:60,575,726-62,435,964	High Copy G...	1,860...	56	q13.33	97.943	☑	239
chr21:44,325,574-44,639,404	High Copy G...	313,831	9	q22.3	100	☐	50
chr1:186,569,223-186,875,6...	CN Loss	306,449	0	q31.1	12.699	☐	25
chr2:16,898,560-17,036,129	CN Loss	137,569	0	p24.3 ...	4.212	☐	14
chr2:139,451,720-140,339,2...	CN Loss	887,496	0	q22.1	6.313	☐	83
chr2:140,497,754-149,100,2...	CN Loss	8,602...	9	q22.1 ...	9.985	☐	831

Clicking on a value in the **Probes** column brings up a window with a drill down of the log ratio and segmentation values of each probe in the selected region.

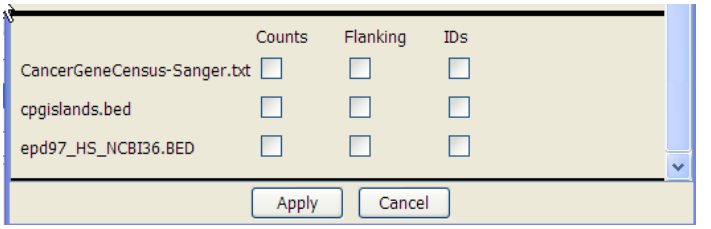


A **Modify View** button allows you to select what data columns to display in the table. After making selections, click **Apply** to update the table with the modified columns.

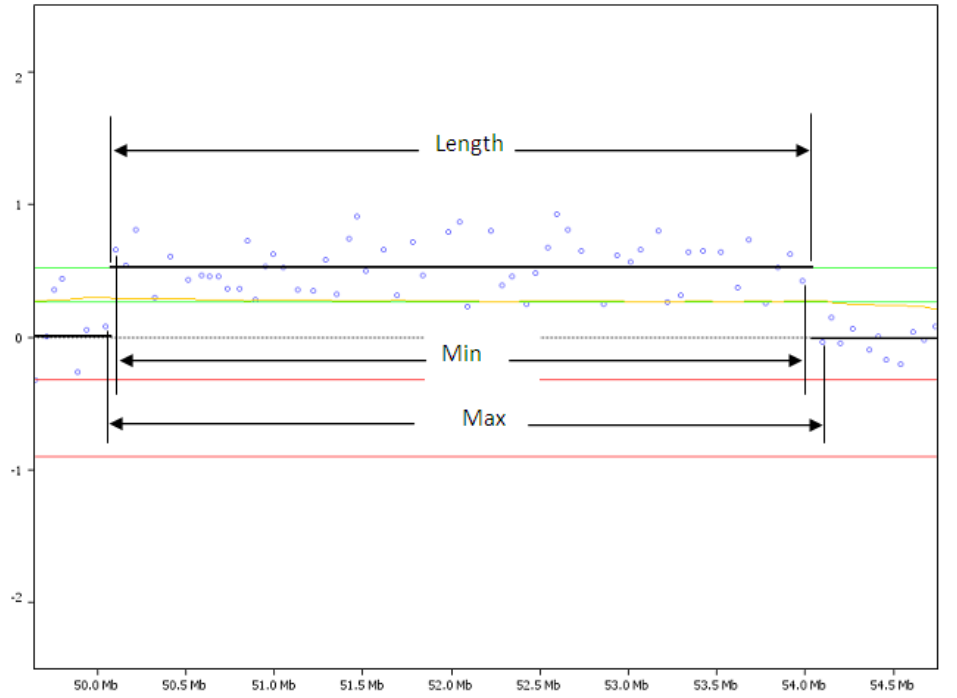


Checking **B/P Genes** will display the breakpoint genes (genes that are only partially covered by the region – possible fusion sites). You have the option to display counts and/or IDs for Genes and miRNAs. For annotation tracks, the count of the number of items in the region, flanking probes of the region, and IDs in the region can be checked

off to be displayed in the table. Data from other tracks such as those from the **customtracks** folder can also be selected to be displayed in the table.



The **Min Size** and **Max Size** refer to the length of a segment. **Min Size** is the length encompassed by the two most external probes in the segment. The **Max Size** is the length encompassed by the closest probes on either end of the segment that are not part of the segment. The **Length** column then displays the value that is midway between the **Min Size** and **Max Size**. See figure below.



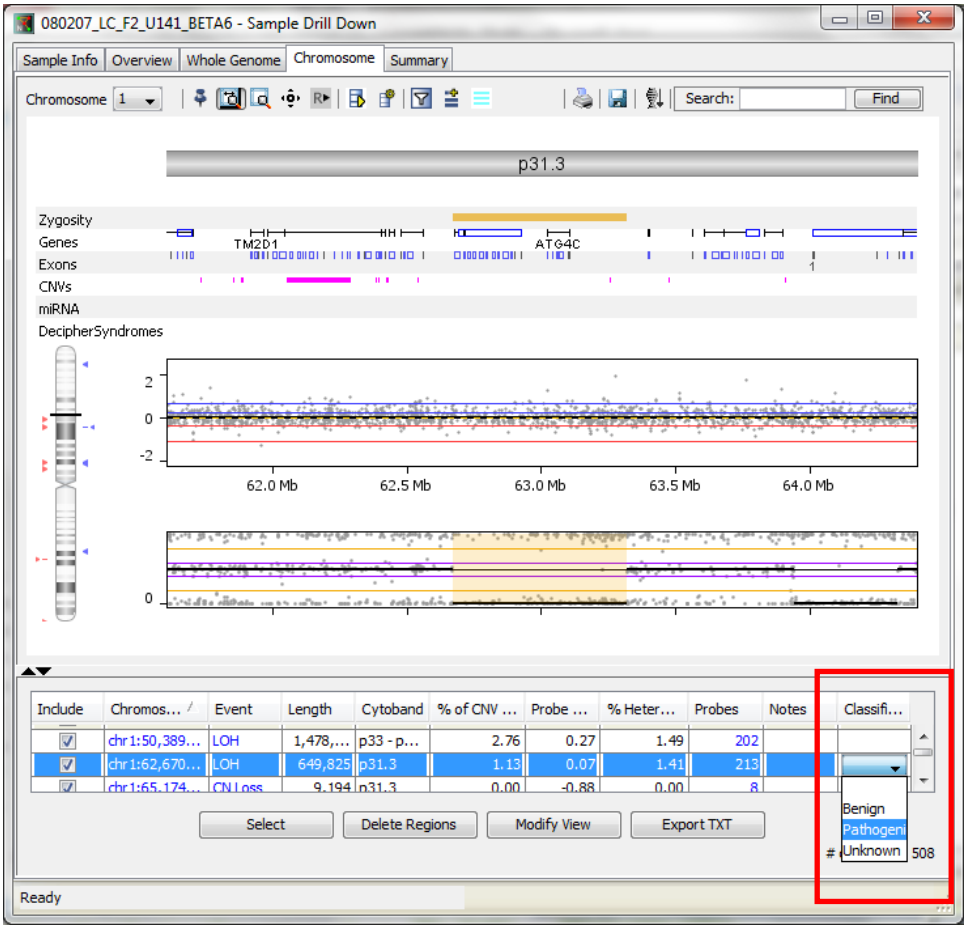
For any probe columns with “ID” selected in the track selection window, the value in the cell will display the name of the first two probes in the region followed by ellipses to indicate there are more (figure on right, below). When the table data is exported, the report will list all probes in the region. If “flanking” is selected in the track selection window, then the first and last probes in the region will be listed with ellipses in the middle (figure on left below).

<..>Affy SNP6
CN_330265...CN_1220026
CN_943457...CN_943512
SNP_A-8382037...CN_1283750
CN_912751...CN_925943
CN_411107...CN_949144
SNP_A-8478800...SNP_A-8544005
SNP_A-1835634...CN_937957
SNP_A-8360526...CN_583216
CN_700237...CN_713386

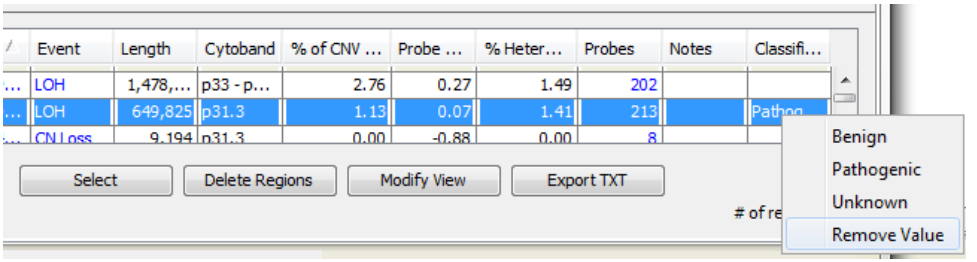
Affy SNP6
CN_330265, SNP_A-8448030,...
CN_943457, CN_943458,...
SNP_A-8382037, CN_1283649,...
CN_912751, SNP_A-8489420,...
CN_411107, SNP_A-2027982,...
SNP_A-8478800, SNP_A-8375049,...
SNP_A-1835634, SNP_A-8371362,...
SNP_A-8360526, SNP_A-4193740,...
CN_700237, CN_700238,...

CALL CLASSIFICATION

Users can classify calls as pathogenic, non-pathogenic or unknown in the classification column of the report table. If the **Classification** column is not visible in the report table, click on the **Modify View** button and select to display this column. Clicking in the cell in this column brings up a drop down menu with the selections. In the example below, we can choose to classify this copy neutral LOH as pathogenic.

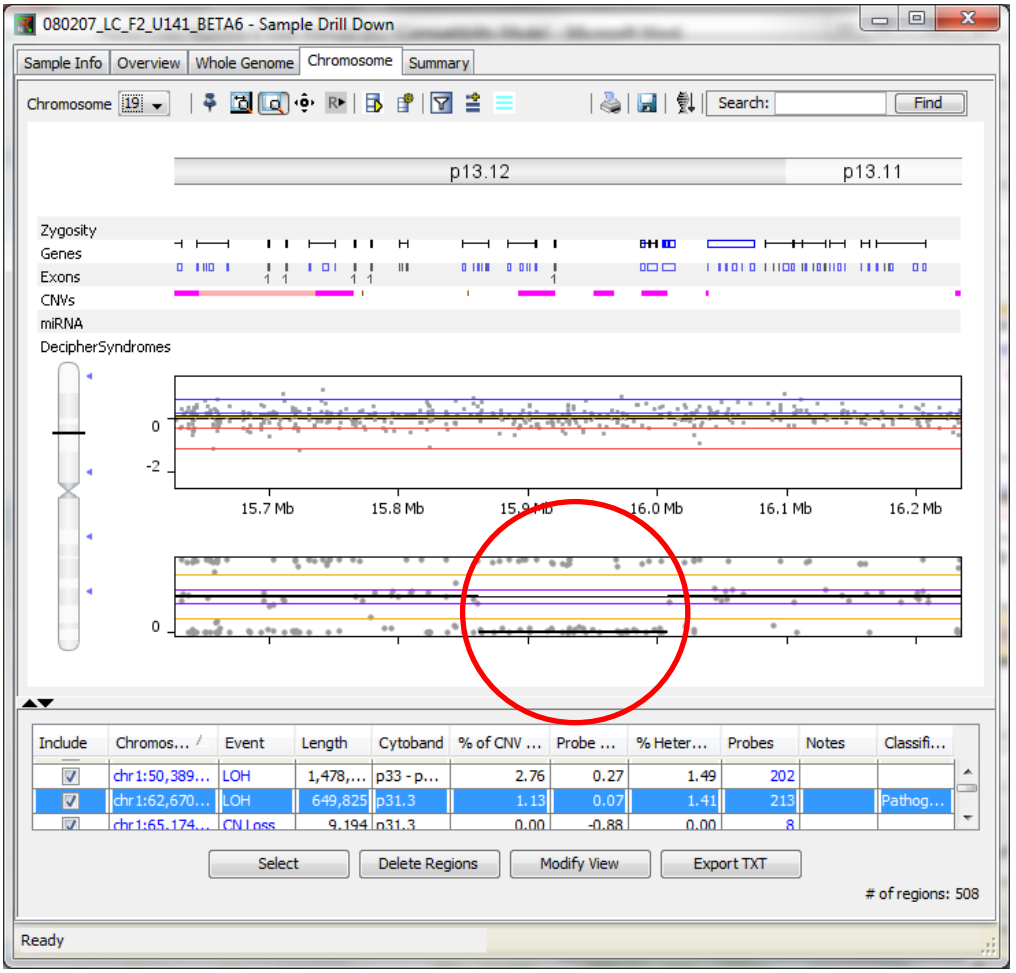


Right clicking in the cell will bring up a context menu with the same selections in addition to the option to remove the classification value.

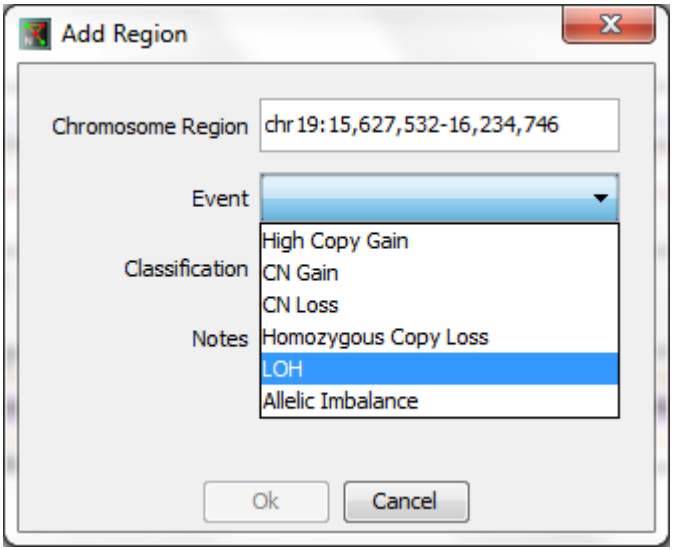


MANUAL CALLING

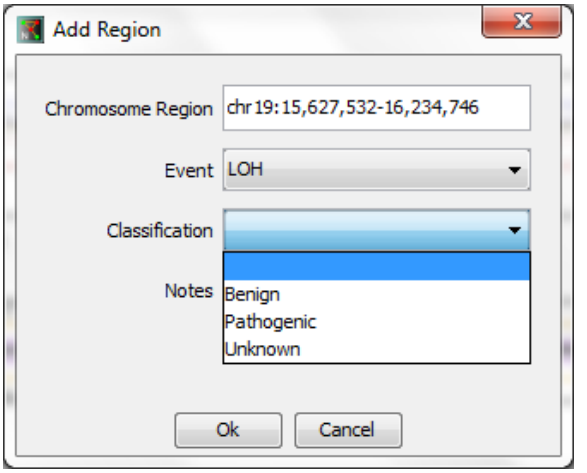
The Nexus Express SNP-FASST2 algorithm provides for automatic calling of regions. Occasionally, a user may want to make call in a region that was not called by the algorithm. Nexus Express provides for this with the **Add | Join Region** tool. In the figure below, an area that appears to be a copy neutral LOH was not called automatically. We want to define this as an LOH region.



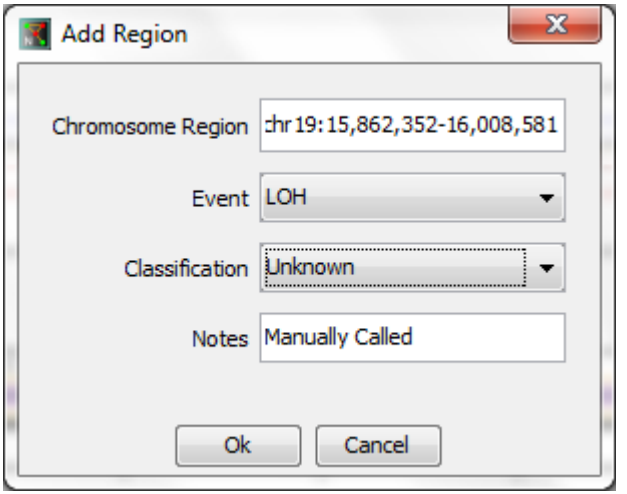
Clicking on the **Add | Join Region** tool  brings up the **Add | Join Region** dialog with the chromosome region field filled in with the base pair range visible in the current display. The coordinates here can be edited. Next, an Event needs to be selected:



Then a classification:

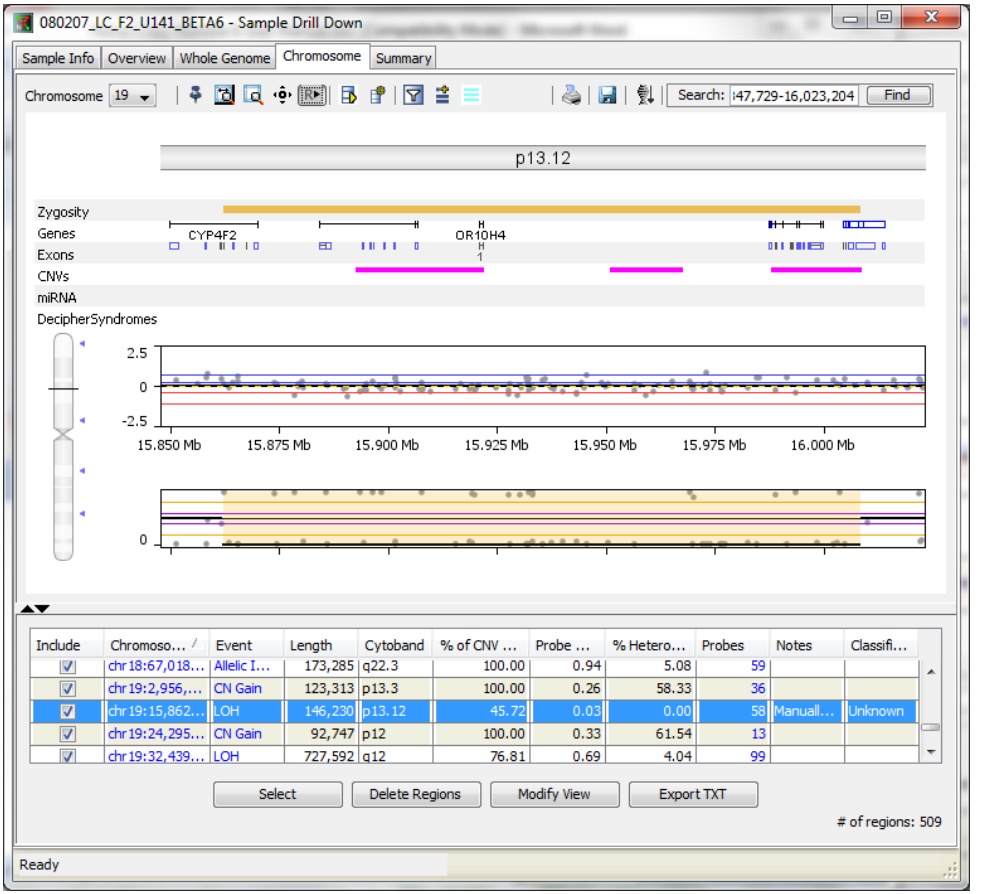


And finally, the **Notes** field can be edited. By default when adding a region, **Manually Called** is the value in this field but it can be changed to anything the user desires. After adjusting the region field to correctly span the LOH region, following is what we selected:

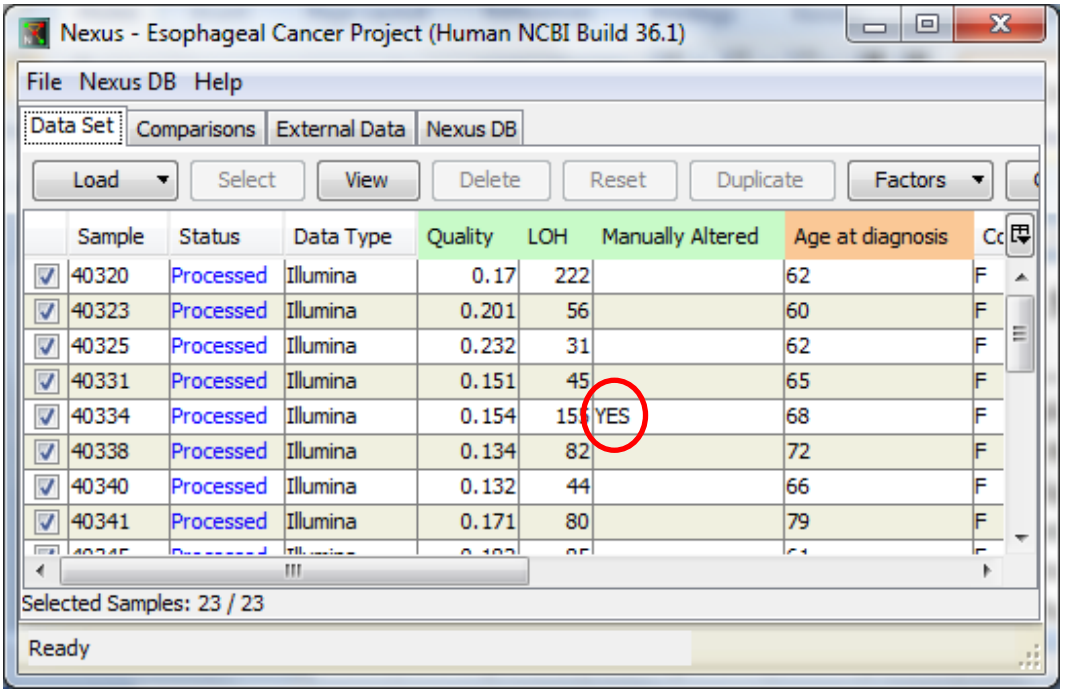


Now after clicking OK, we can see this region is now defined as an LOH region as indicated by the brown bar under the ideogram. Using the Locate Region tool, we see the corresponding entry in the report table (row highlighted in the figure below).

If automatically-called regions overlap the manually-called region, Nexus Express asks whether to delete the automatically-called regions and replace with the manually called region.

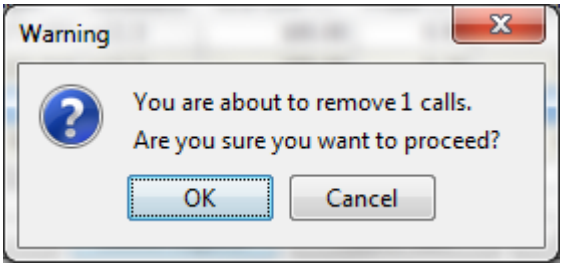


In addition to the notation “Manually Called” in the Notes column, the sample containing any manual alterations in calls is marked by **YES** in the Manually Altered column in the Data Set tab.



DELETING REGIONS

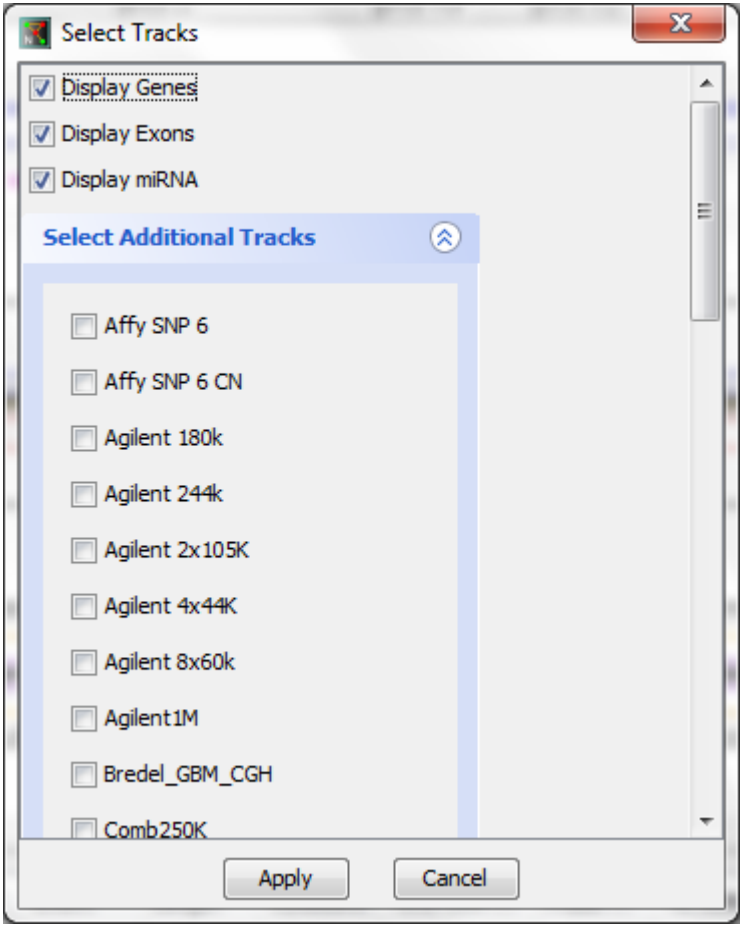
If you want to delete a region, either manually called or automatically called, use the **Delete Regions** button to do so. You can select one or multiple rows by highlighting. Then click on the Delete Regions button to delete the entry or entries. An alert appears to confirm that you do in fact want to delete regions.



After clicking OK, the region is removed from the report table and from the ideogram.

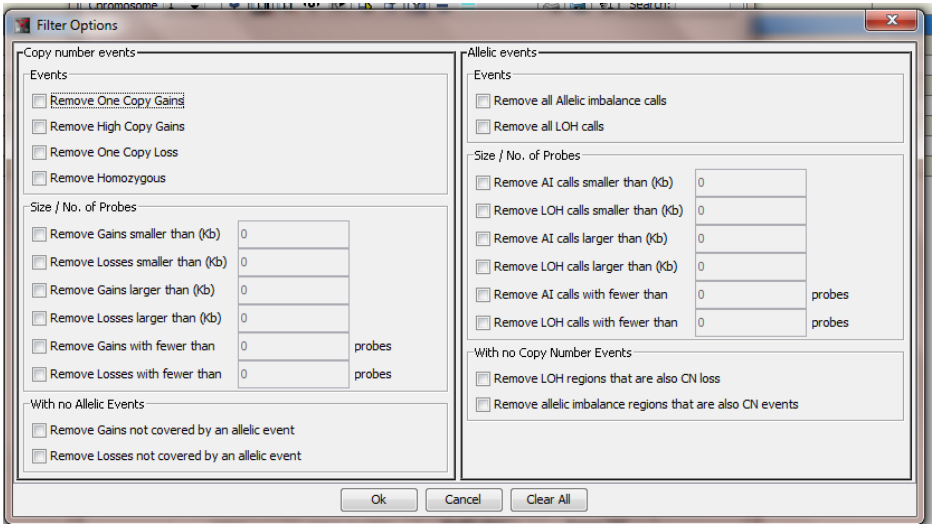
MODIFYING TRACKS

To quickly modify the visible tracks in the current view, click on the **Modify Tracks** tool  to bring up the track modification window. Here you can check or uncheck additional tracks to display them or hide from displaying them in the annotation tracks panel.



FILTERING THE RESULTS

To limit what is displayed in the results, the **Filter** button  provides quick access to the **Filter Data** options to filter out calls from the display. Clicking the button brings up a window with the Filter Data options accessed via the **File->Options** menu.

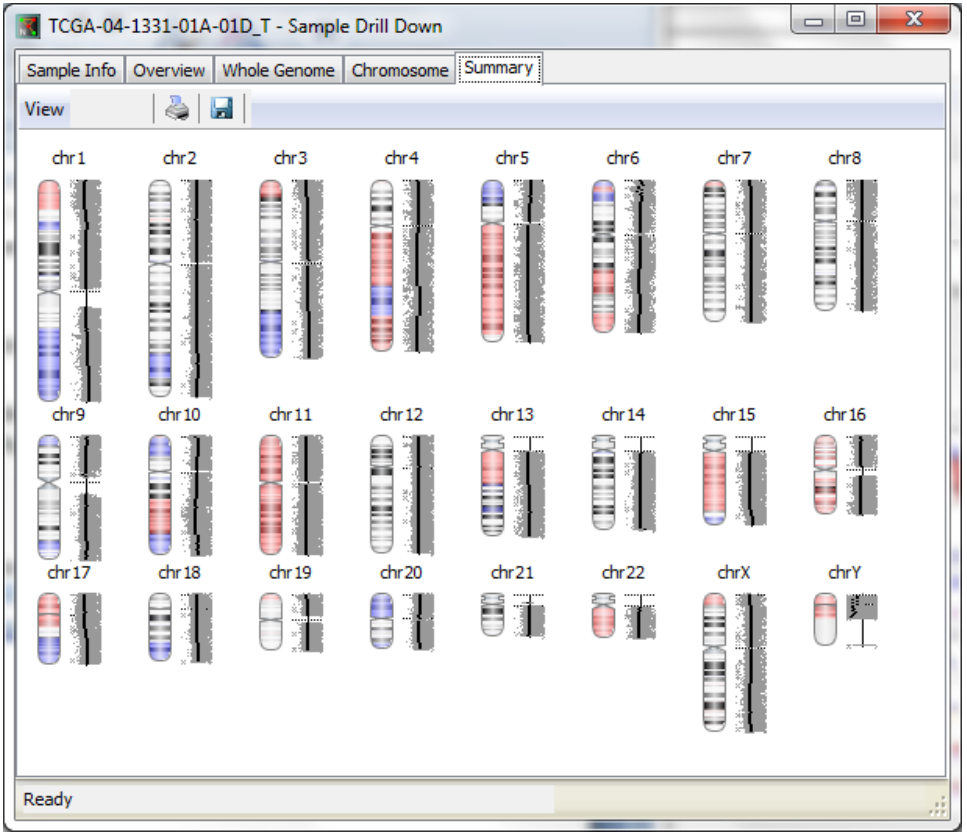


If any one of the checkboxes is marked, the button turns red indicating that the filter is active: 

SUMMARY TAB

The **Summary** tab shows a view of all the chromosomes indicating area aberration with colored shading on the ideogram and individual probes plotted as gray dots. With the **View** menu, either CN Aberrations or Allelic Events can be selected to be viewed. The figure below shows CN Aberrations. With CN Aberrations selected, gain is represented by blue and loss, red. With Allelic Events selected, LOH is indicated in light brown and allelic imbalance in purple. Clicking on a chromosome ideogram, label, or plot will take you to that chromosome's view in the **Chromosome** tab.

CN aberrations displayed:



ALLELE SPECIFIC ANALYSIS AND VISUALIZATION

Based on B allele frequency data, Nexus Express provides allele specific results such as LOH, imbalance, or no change. Please see the section on *Allele-Specific Computation* page 24 for more information on this computation and B-Allele frequency plots.

The B-allele frequency values provide information on LOH and allelic imbalance events. The section on Genome and Chromosome Views in the Results, page 54, talks about the graphical display of these results. The section on *Additional Thresholds For B-Allele Frequency Data*, page 211, discusses parameters specific to SNP-FASST2 Segmentation algorithms. All computation, such as comparisons, is performed for allelic events as well and is covered in their respective sections. *Appendix A: Data Type Format and Sample Descriptor Templates*, page 261, details input requirements for data types.

FREQUENCY SIGNIFICANCE TESTS

When evaluating aberrations across a set of samples, it is important to determine parts of the genome where there is a statistically significant high frequency of aberration over the "background" level of aberration. There are different approaches to identifying this and Nexus Express provides the STAC algorithm for frequency significance testing.

STAC

Nexus Express implements the global frequency statistic approach of the Significance Testing for Aberrant Copy number (STAC) method developed at the Penn Center for Bioinformatics at the University of Pennsylvania (Diskin SJ, Eck T, Greshock J, Mosse YP, Naylor T, Stoeckert CJ Jr, Weber BL, Maris JM, Grant GR. STAC: A method for testing the significance of DNA copy number aberrations across multiple array- experiments. Genome Res. 2006 Sep; 16(9):1149-58). The **Significant Peaks** button in the **Aggregate** page performs this function.

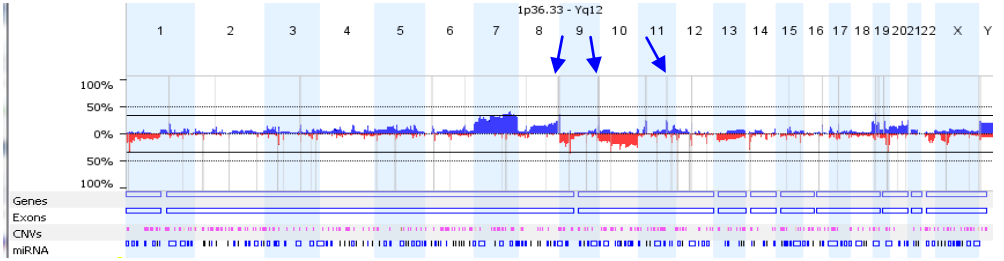
A maximum p-value cut-off needs to be specified in the **P-Value cut-off** text box at the bottom of the **Aggregate** page under the **Results** tab. Once significance testing is performed, the results in the **Aggregate** page are regions that meet both the p-value cut-off as well as the aggregate % cut-off. For example, if .05 is the **P-Value cut-off** and 35% is the **Aggregate % cut-off**, the regions in the aggregate page are ones for which the frequency is at least 35% and this frequency is significant at p=.05.

The algorithm tries to identify a set of aberrations that are stacked on top of each other such that it would not occur randomly. To find these events, we permute the aberrations in each arm of each chromosome and see how likely it is for an event (e.g. a gain) to occur at any location at a particular frequency. Then we use a p-value cut-off and highlight those areas that meet the given p-value. For example, let's say there are 10 samples and only 3 small aberrations in the p-arm of chromosome 1. If all three of these aberrations are at the same location, it is very unlikely to happen by chance alone so it will be a significant peak. As another example let's say most of 5q is gained in 8 of

the 10 samples; so these areas will have a frequency of 80%. However this will not be significant as obtaining that frequency with so many large events can happen randomly. It is important to note this fact since one should not dismiss the regions that are not marked as “significant” but should use this tool to examine some narrower events. A good example is in the figure below where the large deletion on chromosome 10 is not highlighted in gray. This does not mean that this aberration is not important.

Please note that since significance testing is done using random permutations of segments many times, it is very likely to see small fluctuations in the p-values in different tests. The magnitude of these changes should be rather minor.

Gray vertical areas indicating significant regions at a specified p-value:



Aggregate table before significance testing:

Region	Region Length	Event	Genes	Frequency %	P-Value
chr1:11,778,627-12,004,488	225,861	Loss	8	35.849	
chr10:40,300,000-44,480,418	4,180,418	Loss	11	35.849	
chr10:71,848,002-73,441,462	1,593,460	Loss	14	35.849	
chr10:93,736,047-95,151,287	1,415,240	Loss	10	37.736	
chr10:101,579,695-103,216,222	1,636,527	Loss	26	35.849	
chr10:104,069,813-104,443,883	374,070	Loss	11	35.849	
chr10:116,765,527-121,438,235	4,672,708	Loss	29	35.849	
chr10:134,991,656-135,374,737	383,081	Loss	11	35.849	
chr12:55,402,061-55,676,753	274,692	Loss	6	35.849	
chr19:54,233,641-54,580,069	346,428	Loss	15	35.849	
chr21:0-12,300,000	12,300,000	Gain	6	49.057	

Aggregate table after significance testing:

Region	Region Length	Event	Genes	Frequency %	P-Value
chr1:11,778,62...	225,861	Loss	8	35.849	0
chr12:55,402,0...	274,692	Loss	6	35.849	0
chr19:54,233,6...	346,428	Loss	15	35.849	0
chr21:0-12,300...	12,300,000	Gain	6	49.057	0

If you want to use a different p-value setting, just edit the **P-Value cut-off** input box and click anywhere on the table to update the results. The gray bars indicating significance will be updated on the frequency plot in the **Genome** and **Chromosome** pages.

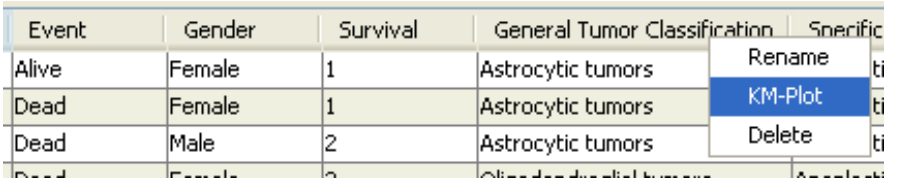
KAPLAN-MEIER CURVES

The Kaplan-Meier curve is a common method used to describe survival characteristics of a group of patients and is typically used to compare survival times between different groups. In an ideal situation a patient's data would be recorded up till an endpoint (e.g. the time of death, cancer free,…) but in real world situations a patient may not have been followed after a certain point in time (e.g. due to need to drop out of study, death, etc.) so different patients will have varying lengths of follow-up which can cause problems in calculations of survival. One would still like to allow a patient to contribute to survival up till the time he was followed and then removed from the curve after that time (censoring the patient). The Kaplan-Meier method adjusts for the varying follow up times to allow censored patients to contribute to the estimation of survival curves.

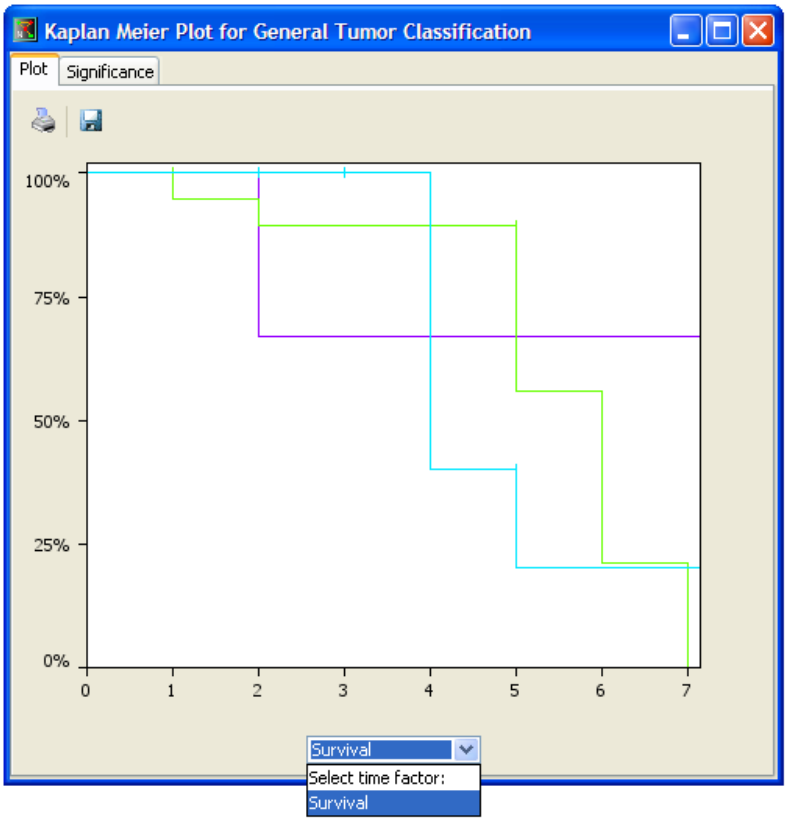
Generation of a K-M (Kaplan-Meier) plot requires the presence of a column called **Event** and another column with continuous data in the table in the **Data Set** tab. In the figure below is a column called **Event** containing values dead or alive and a column called **Survival** with continuous data.

Event	Gender	Survival	General Tumor Classification	Specific T
Alive	Female	1	Astrocytic tumors	Anaplastic
Dead	Female	1	Astrocytic tumors	Anaplastic
Dead	Male	2	Astrocytic tumors	Anaplastic
Dead	Female	2	Oligodendroglial tumors	Anaplastic
Alive	Male	2	Oligodendroglial tumors	Anaplastic
Alive	Male	2	Oligodendroglial tumors	Anaplastic
Alive	Female	2	Mixed gliomas	Anaplastic
Alive	Female	3	Mixed gliomas	Anaplastic
Dead	Male	4	Mixed gliomas	Anaplastic

To generate the K-M plot, just right click on the column header of the Factor you want to use to view the survival rate. In the figure below, General Tumor Classification is selected for the K-M plot.



Once **KM-Plot** is selected from the drop-down, a window opens up with the plot (in the **Plot** tab) as seen below.



The **Select Time Factor** drop-down menu at the bottom of the window allows you to select the time factor for the x-axis. The y-axis displays the percent survival from 0 to 100. The survival curves are drawn in different colors for each factor and moving the mouse over the line displays the Factor value for the curve.

The image can be printed or saved to disk by clicking on the icons in the top left of the window.

The **Significance** tab in the KM curve window displays p-values to help determine if the survival differences between factor values are significant:

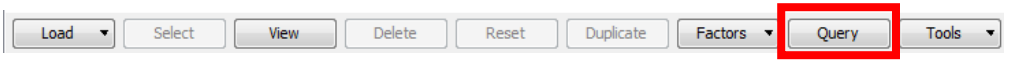
Plot | Significance

	Astrocytic tumors	Mixed gliomas	Oligodendroglial tumors
Astrocytic tumors	1.0	0.13555578759...	0.356001176845151
Mixed gliomas	0.1355557875938745	1.0	0.1266304579476171
Oligodendrogli...	0.356001176845151	0.12663045794...	1.0

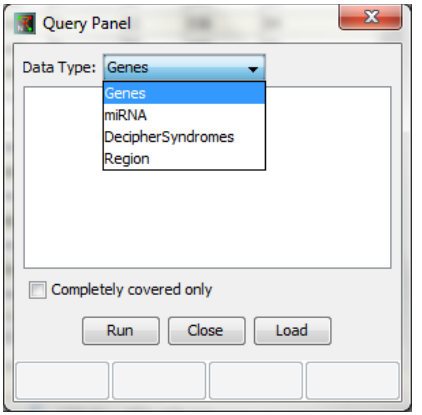
QUERYING DATA

The query tool displays copy number events for user defined genes or regions within the project. This tool is available in the **Data Set** tab as well as from a region drill down window (see end of this section). Another way to query data is by searching the Nexus DB repository. If your software has the Nexus DB license, then a **Search Repository** button is available in the **Data Set** tab (please see the section on *Querying the Repository*, page 249, for more information).

Query button in the **Data Set** tab:

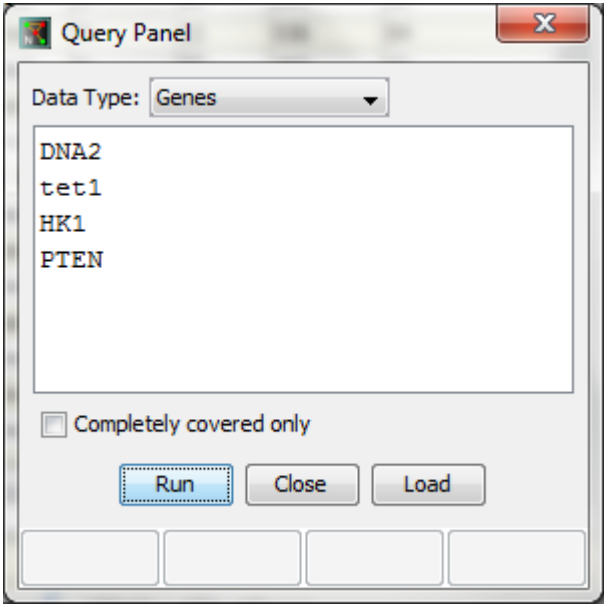


Clicking on the **Query** button brings up a window where you can either type in a list of genes, regions, etc. (selectable from the dropdown menu) or load a file containing these values (using the **Load** button). This search is case insensitive so querying for either **pten** or **PTEN** finds the PTEN gene. The default values in the Data Type dropdown are **Genes**, **miRNA**, and **Region**. Other types of items that can be queried (and shown in the dropdown) are dependent on the tracks you have selected under **File->Options->Track Selection**. In the figure below, **DecipherSyndromes** and **Gene Association Database** are shown in the drop down because these tracks were chosen to be displayed in the project.

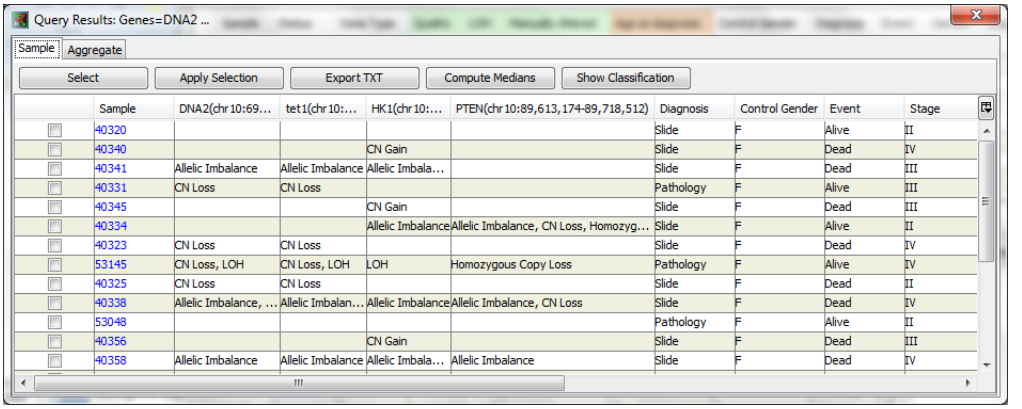


The **Completely covered only** checkbox will show only regions that are completely covered by the aberration. For example, if a search for PTEN is performed and a sample only shows part of the gene as gained, then this sample will not be included in the results.

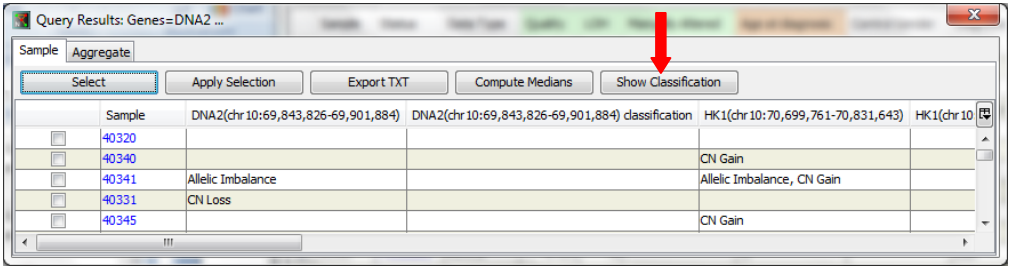
Typing in gene names:



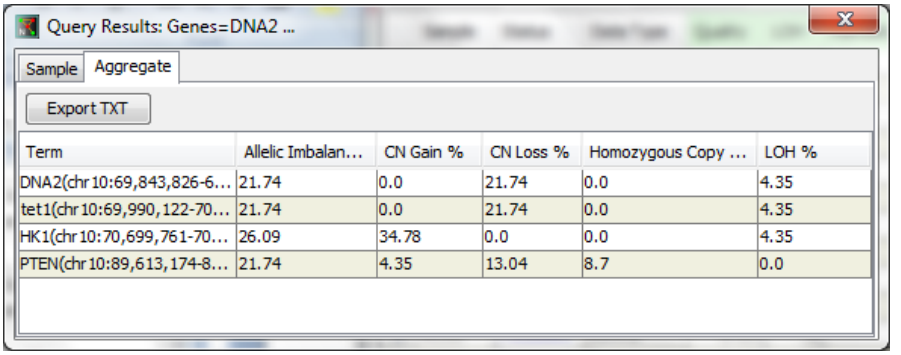
Results are displayed in a new window containing two tabs: **Sample** and **Aggregate**. The **Sample** tab lists all the samples in rows and contains one column for each queried item listing the type of events in each sample for the queried items and lists the factors columns as well (see figure below). Clicking on the sample name opens up the sample drill down window for that sample.



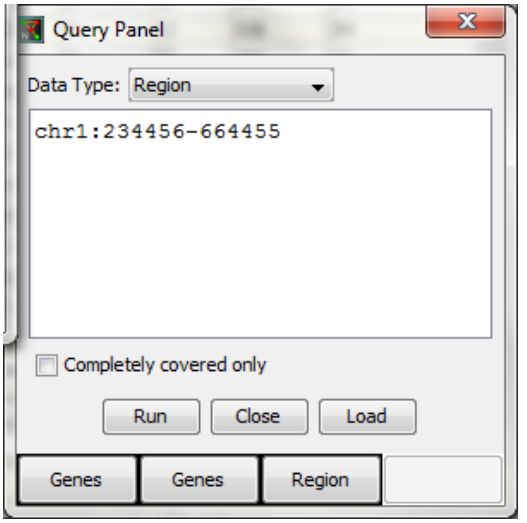
Clicking on Show Classification adds another column (one per queried gene), showing the classification values (benign, pathogenic, and unknown) if any were entered by the user (see section on *Call Classification* page 129).



The **Aggregate** tab shows the percentage of samples containing each type of event for the queried terms.

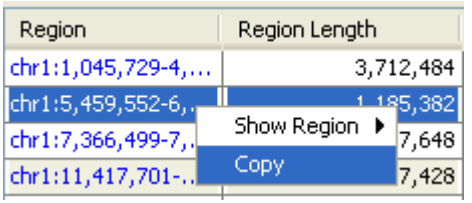


A maximum of four query results windows can be open at one time. Each time a query is run within the same query panel, the rectangles on the bottom will display the type of data that was queried (e.g. Genes, miRNA, etc.):



Once 4 queries have been run and their result windows are open, the **Run** button is inactive and no additional queries can be run until a query window is closed.

Rather than typing in a chromosomal region to query, a region in the Results->Aggregate table can be copied by right clicking on the cell containing the region and selecting **Copy**:



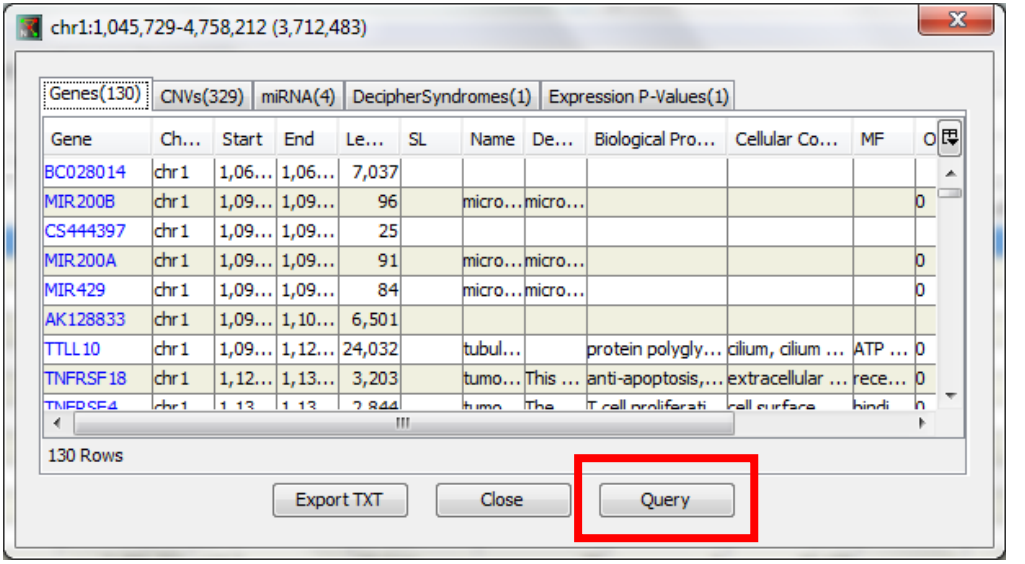
Now paste the region into the Query window, select **Region** from the dropdown and click **Run**.

From the Query Results window, you can also select just some samples, e.g. ones with CN gain as in the figure below, and click on **Apply Selection**. This will then select only

these samples in the **Data Set** tab and you can click on the **View** button to see the results.

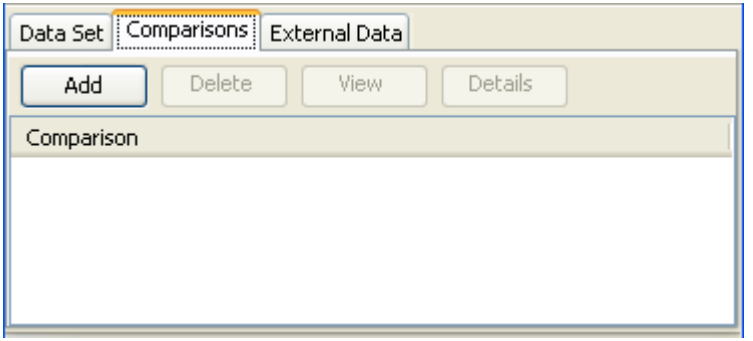
The Query tool is also available from the region drill down windows. Clicking on the tool here automatically carries out the operation on all terms listed in the window.

Query results button:



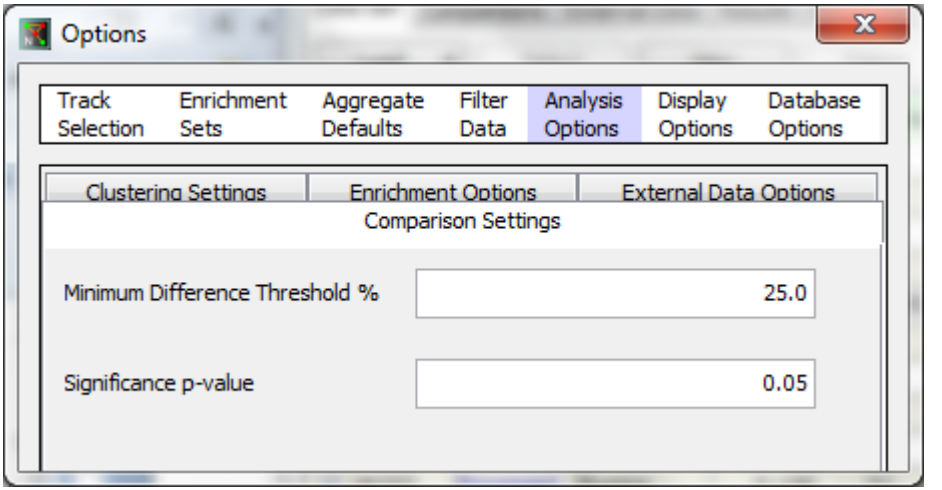
COMPARING TWO GROUPS OF SAMPLES

Comparisons are created in Nexus Express to compare two groups of profiles to look for differences between them. This is a very significant and unique feature of Nexus Express that allows comparison between biologically important groups with sample data obtained from diverse experiments facilitating hypothesis-driven statistical analysis. This feature can be accessed via the **Comparisons** tab in the main Nexus Express window. This tab contains four buttons, **Add**, **Delete**, **View**, and **Details**. If Comparisons have never been created for the project, the comparisons tab will be empty and the only active button is the **Add** button (as seen in the figure below).

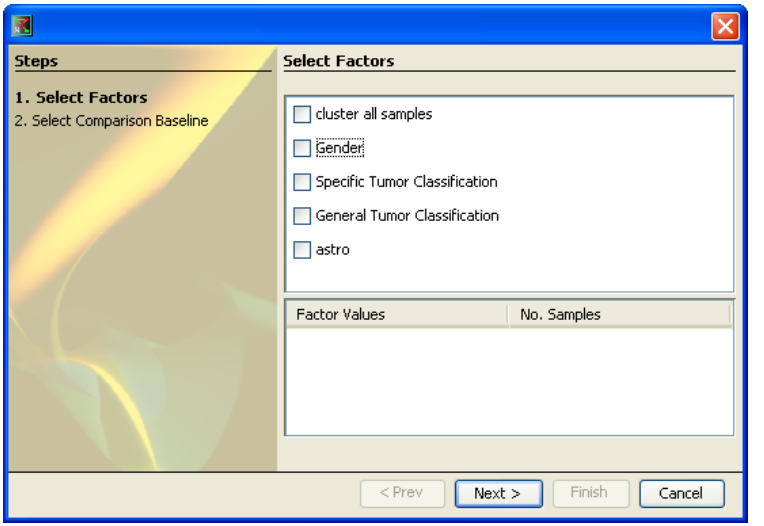


BUILDING COMPARISONS

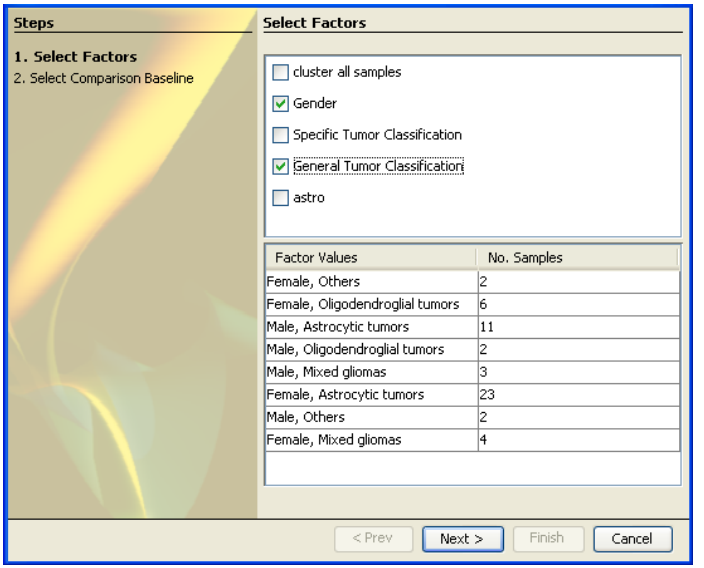
Settings for comparisons can be edited via **File->Options->Analysis Options->Comparison Settings**.



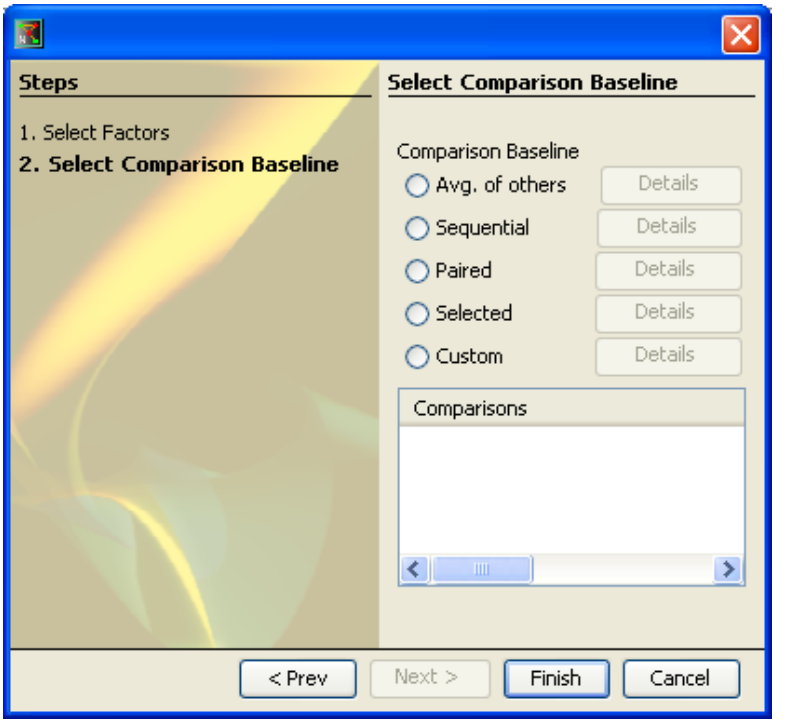
Please see the section on *Analysis Options*, page 228, for more information on these settings. Clicking on the **Add** button in the **Comparisons** tab opens up a dialog allowing you to select factors using the checkboxes.



As factors are selected, the **Factor Values** and the number of samples for each Factor Value group appear in the bottom panel. The columns in this panel can be sorted in ascending or descending order by clicking on the column header successively.



Next, in the **Select Comparison Baseline** window, a comparison type and comparison groups will need to be selected.



COMPARISON TYPES

AVG. OF OTHERS

One factor set is compared to the average of all the other factor sets. E.g. if a factor is **Tumor Grade** and the values are **Grade 1**, **Grade 2**, **Grade 3** and **Grade 4** then selecting **Avg. of others** would create the following comparisons:

Grade 1 vs. Average of {Grade 2, Grade 3, Grade 4}

Grade 2 vs. Average of {Grade 1, Grade 3, Grade 4}

Grade 3 vs. Average of {Grade 1, Grade 2, Grade 4}

Grade 4 vs. Average of {Grade 1, Grade 2, Grade 3}

SEQUENTIAL

Compares each factor set in a list to the subsequent factor set in the list. E.g. if the factor is **Tumor Grade** and the values are **Grade 1**, **Grade 2**, **Grade 3** and **Grade 4**, in order, then selecting **Sequential** would create the following comparisons:

Grade 1 vs. Grade 2

Grade 2 vs. Grade 3

Grade 3 vs. Grade 4

Grade 4 vs. Grade 1

PAIRED

The user pairs up factor sets for comparison. E.g. if the factor is **Tumor Grade** and the values are **Grade 1**, **Grade 2**, **Grade 3** and **Grade 4**, then selecting **Paired** would allow creation of the following comparisons:

Grade 1 vs. Grade 2

Grade 3 vs. Grade 4

or

Grade 2 vs. Grade 3

Grade 1 vs. Grade 4

or

Grade 1 vs. Grade 3

Grade 2 vs. Grade 4

or

Grade 2 vs. Grade 1

Grade 4 vs. Grade 3

or

Grade 4 vs. Grade 2

Grade 3 vs. Grade 1

or

Grade 4 vs. Grade 1

Grade 3 vs. Grade 2

SELECTED

One factor set is selected as the baseline and all other factor sets are compared to this in pairs. E.g.

If **Grade 2** is defined as the selected baseline, then the following comparisons are made:

Grade 1 vs. Grade 2

Grade 3 vs. Grade 2

Grade 4 vs. Grade 2

CUSTOM

Allows the user to choose Value sets to use as the selected baseline and compare each remaining value set to the average of the baseline value sets. E.g.

Value Sets: Baseline	**Value Sets: Experiment**
Grade 1	Grade 3
Grade 2	Grade 4

Comparisons are

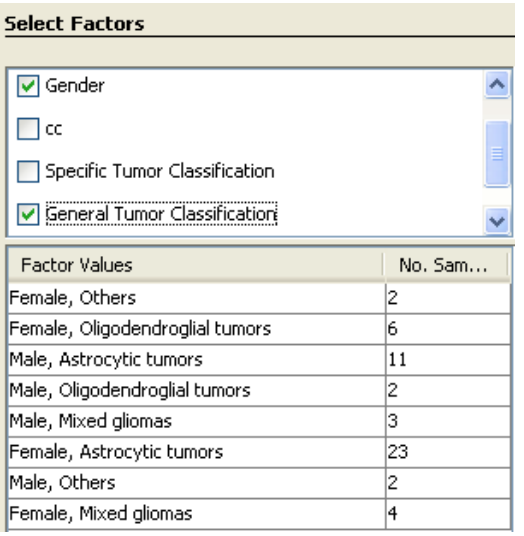
Grade 1 vs. Average of {Grade 3, Grade 4}

Grade 2 vs. Average of {Grade 3, Grade 4}

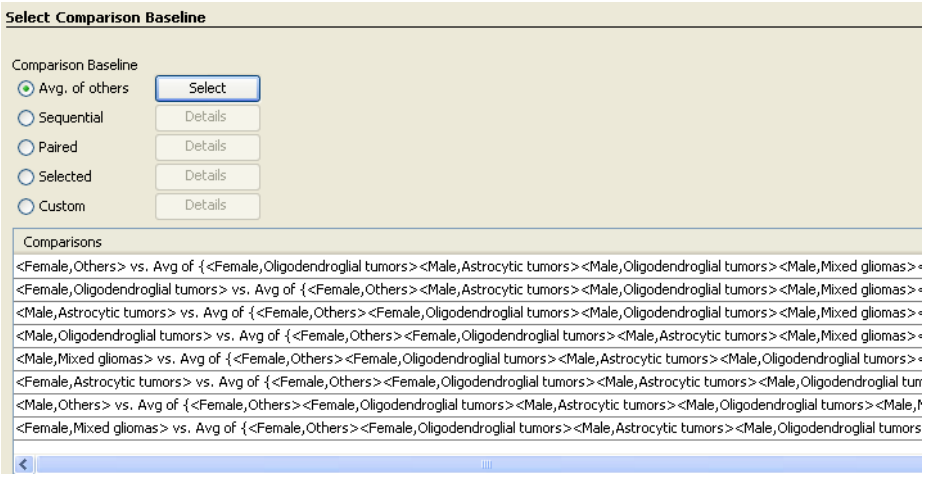
Upon selecting one of the comparison types, the associated inactive **Details** button will turn into an active **Select** button. If **Avg. of others** is chosen as the comparison type, then clicking on **Select** will list all the possible comparisons in the bottom panel. If any other comparison type is selected, then clicking on the **Select** button will open up a dialog where you will create the comparison groups and select the baseline.

SELECTING THE COMPARISON BASELINE

We'll use an example where **Gender** and **General Tumor Classification** are the Factors selected (depicted in figure below) and show how to select the baseline using some comparison types.

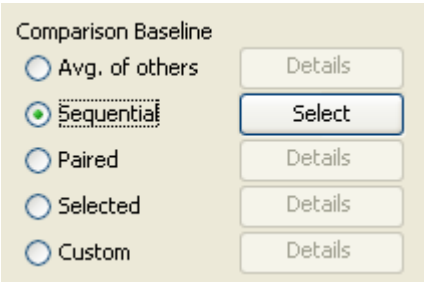


If the Comparison Baseline is **Avg. of others**, pressing **Select** lists all the possible comparisons in the bottom panel.

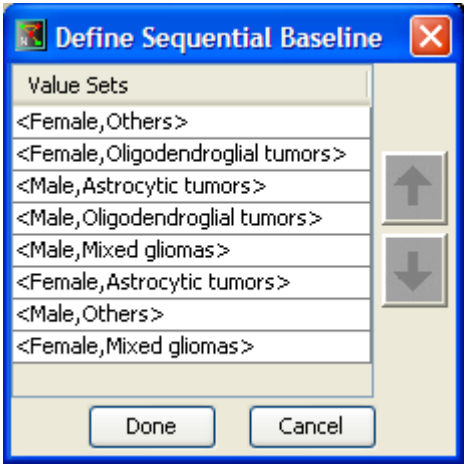


Clicking **Finish** will add all these comparisons to the list in the **Comparisons** tab.

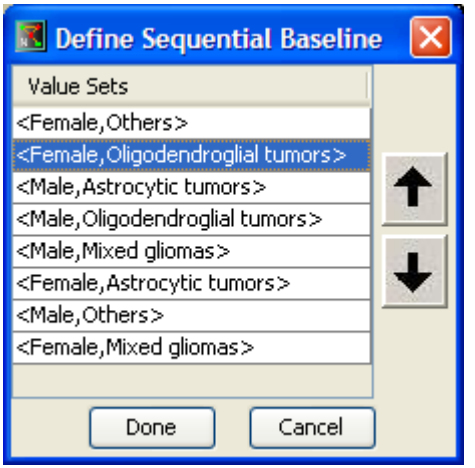
In the example below, **Sequential** is the comparison type selected:



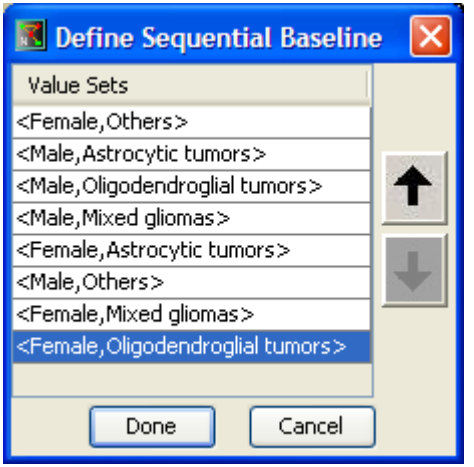
Clicking on **Select** brings up the list of value sets.



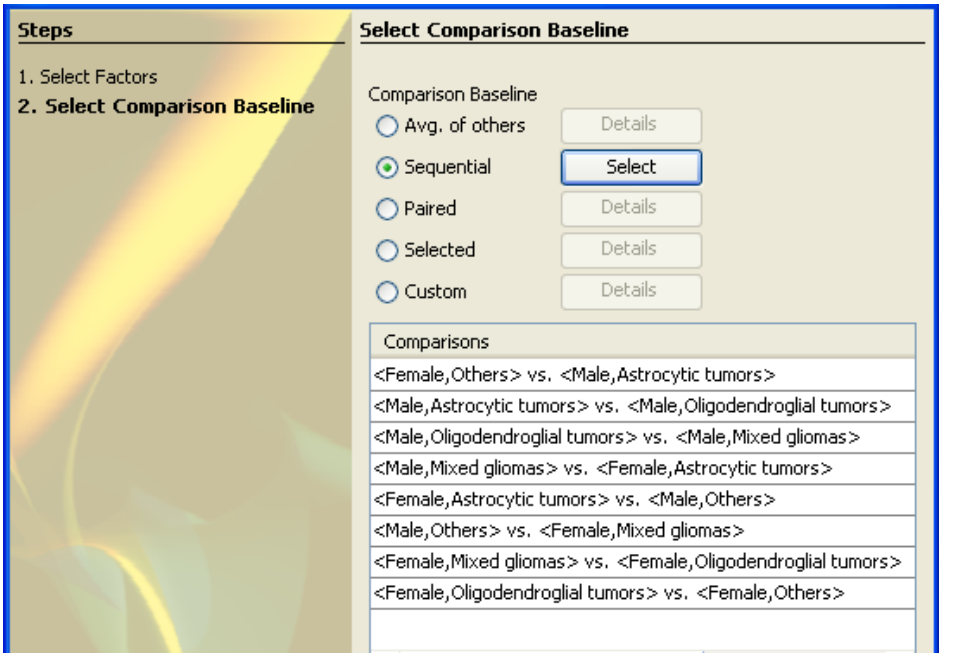
Highlight one or more values and use the arrow keys to move them up or down to change the sequential ordering.



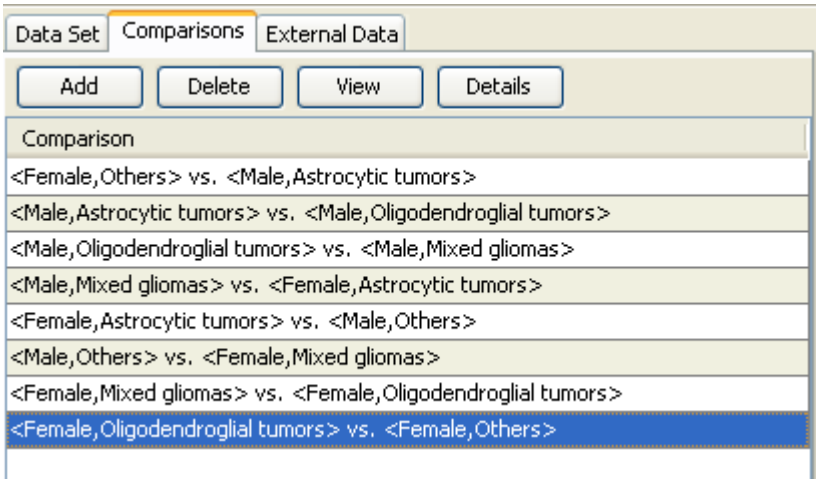
Female Oligodendroglial tumors was moved to the bottom in this case.



The next screen shows the list of comparisons that will be made in the **Comparisons** panel.

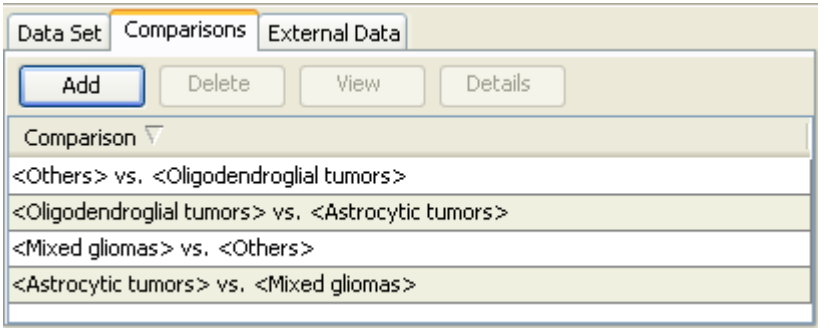


Once the dialog is closed, all the comparisons that were chosen are now listed in the **Comparisons** tab. Highlight a comparison and click **View** to see the results.

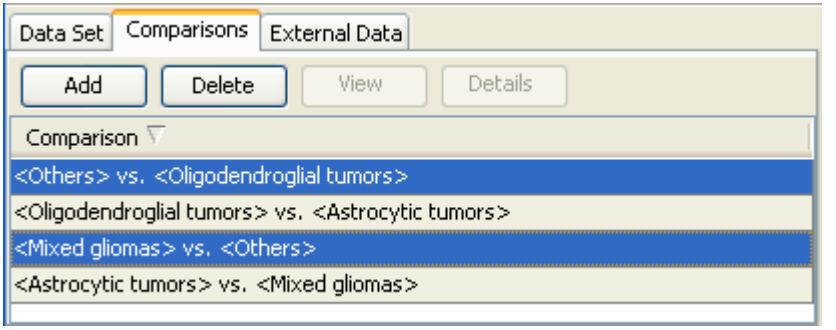


DELETING COMPARISONS

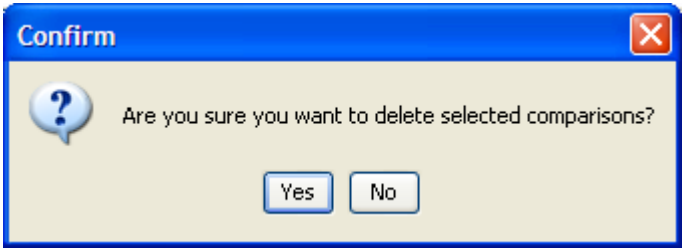
Comparisons can be deleted from the list in the **Comparisons** tab by selecting one or multiple comparisons and clicking on the **Delete** button. The **Delete** button remains inactive until at least one **Comparison** is selected:



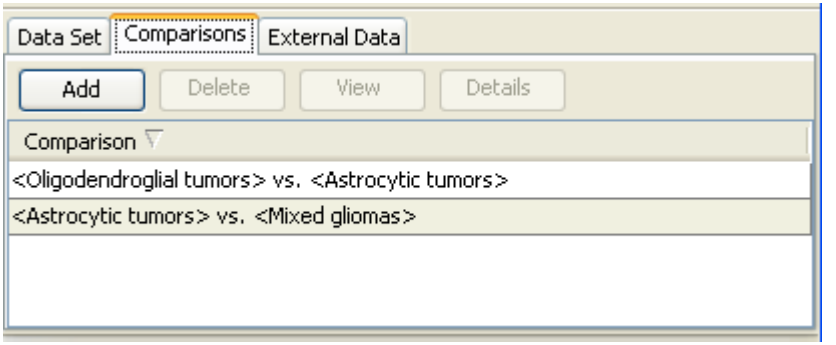
In the figure below, all **Comparisons** containing the factor **Others** have been selected.



Once the **Delete** button is pressed, a **Confirm** dialog opens up asking you to make sure you want to delete the selected comparisons.



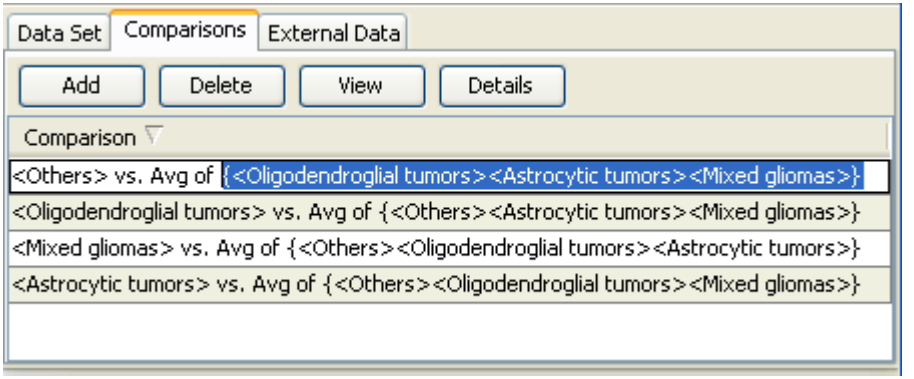
The **Comparisons** table now contains only two comparisons.



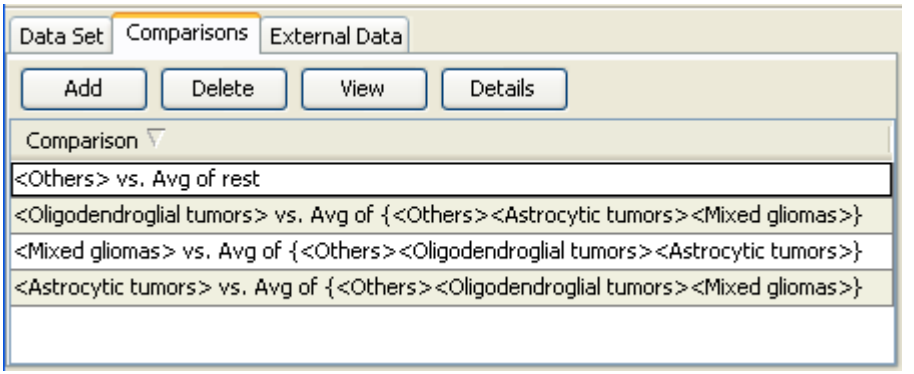
EDITING THE COMPARISON NAME

Once a comparison has been created, the default name given to the comparison can be changed by double clicking on the comparison name to make the field editable.

Here, we want to change the name for the first comparison in the list.



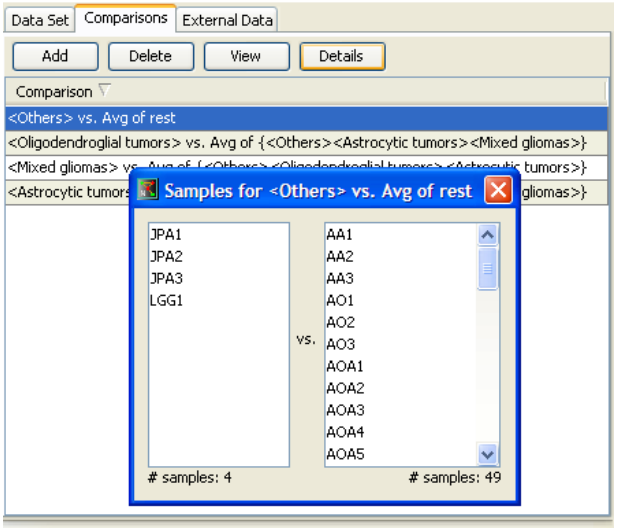
We chose to shorten the comparison name to **Others vs. Ave. of rest** and below is the **Comparisons** tab with the updated comparison name.



ACCESSING DETAILS OF A COMPARISON

Once comparisons have been created and you would like to know which samples are being used in a particular comparison, the **Details** button can be used for this information. Since comparisons can be created using a sub-set of the samples contained in a project, this feature is especially useful in determining the specific samples that are being used in a comparison if only a sub-set of all data in a project is used to create the particular comparison.

The **Details** button is inactive until a comparison is selected. Selecting a comparison and clicking **Details** opens up a dialog displaying the samples in each Value set that are used in the comparison. In the example below, details are viewed for the **Others vs. Avg. of rest** comparison. Each panel in the dialog lists the sample names that are contained in the respective Value set. In this example, all samples belonging to the Value set **Others** are listed in the left panel and all samples belonging to the Value set **Avg. of rest** are listed in the right panel. The total number of samples in each value set is listed below each panel.



VIEWING COMPARISON RESULTS

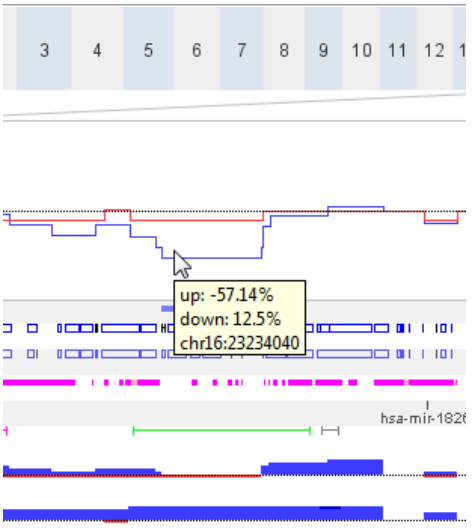
Comparison results are organized into three tabs: **Results**, **Genome**, and **Chromosome**. Please see the **Options->Comparisons** section for more detail on how the comparisons are performed. Below we will cover the results available in each tab in the Comparisons window. Computation is performed for both copy number and allelic events . The display can be toggled to show either event at one time. In the examples below, we will refer to copy number changes but all analysis is applicable to allelic events as well.

GENOME AND CHROMOSOME TABS

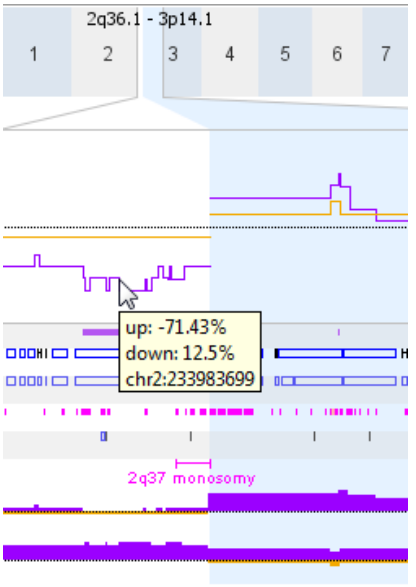
The **Genome** and **Chromosome** plots graph the frequency difference between the two groups in a Comparison. The exact calculation for the difference can be selected using the dropdown at the bottom of the window; please see the section below on

Calculations, page 172, for further details. Using the **View** menu, the display can be changed to show either copy number or allelic events (see figure below). Red indicates regions of loss and blue indicates regions of gain. For the Allelic Events, purple indicates allelic imbalance and brown indicates LOH. The difference is obtained by subtracting the second group (comparison baseline) from the first group. Moving the mouse over the plot in the top part of the window displays the difference in percent for the gain and loss for each location (gain and loss is shown as **up** and **down** respectively).

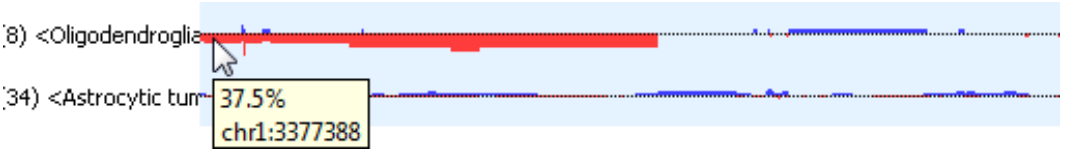
Copy Number display:



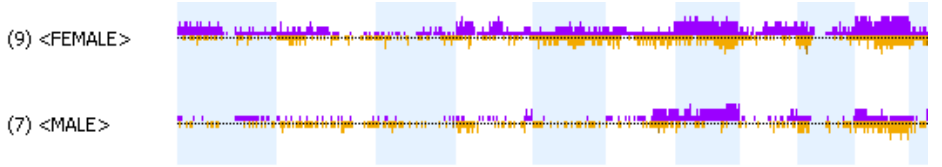
Allelic Event display:



Moving the mouse over the plots of each comparison group in the bottom part of the window shows the percent gain or loss for the individual group. In the example below, we can see that 37.5% of the samples in the **Astrocytic tumors** group have a loss at chr1:3377388.



An example of allelic events plots of comparison groups:



In the Comparison list below, the comparison **Astrocytic tumors vs. Mixed gliomas** is selected to be viewed.

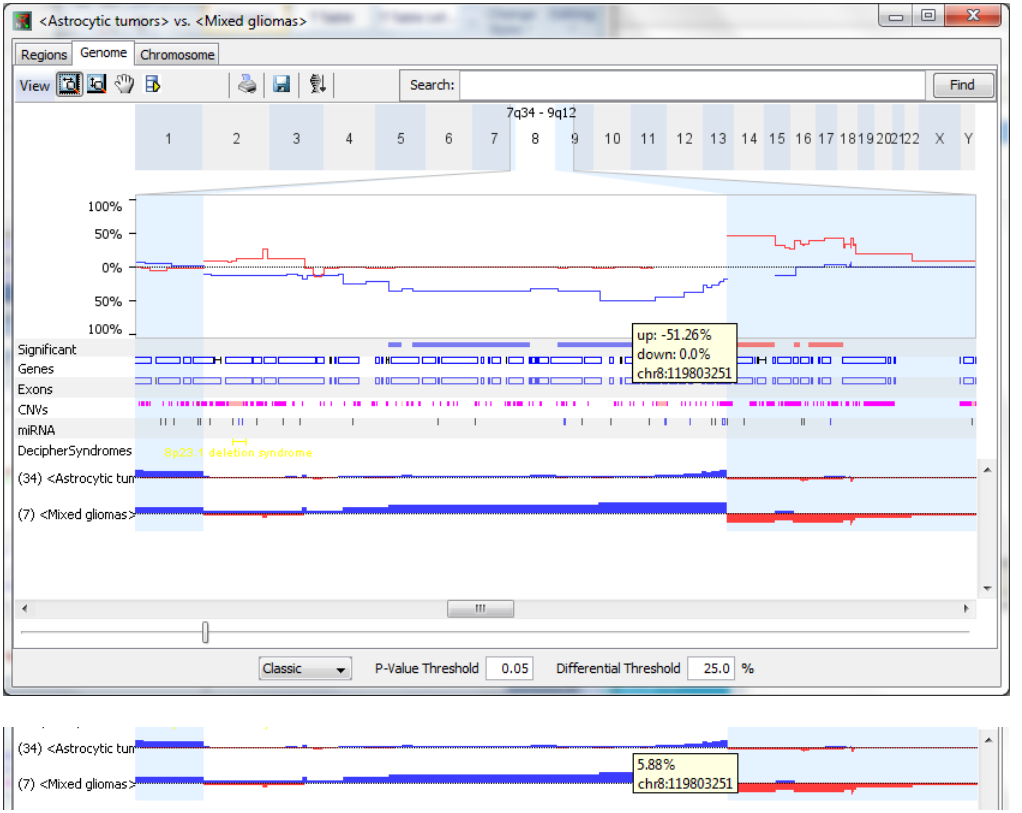
Comparison
Others vs. Ave. of rest
<Oligodendroglial tumors> vs. Avg of {<Others><Astrocytic tumors><Mixed gliomas>}
<Astrocytic tumors> vs. Avg of {<Others><Oligodendroglial tumors><Mixed gliomas>}
<Mixed gliomas> vs. Avg of {<Others><Oligodendroglial tumors><Astrocytic tumors>}
<Astrocytic tumors> vs. <Mixed gliomas>
<Oligodendroglial tumors> vs. <Others>

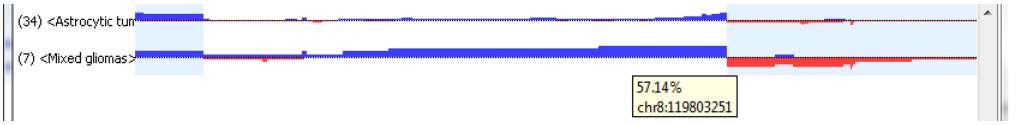
The difference taken in this comparison would be **<Astrocytic tumors>** minus **<Mixed gliomas>**. For copy number events, if the magnitude of a gain or loss is greater in the first group (**<Astrocytic tumors>** in this example), then the gain is plotted (in blue) above the 0 baseline and the loss (in red) is plotted below the 0 baseline. If the magnitude of a gain or loss is lower in the first group, then the gain is plotted in blue below the 0 baseline and the loss is plotted in red above the 0 baseline. For allelic events, if the magnitude of an allelic imbalance or LOH is greater in the first group, then

the allelic imbalance is plotted (in purple) above the 0 baseline and the LOH (in brown) is plotted below the 0 baseline.

The figure below shows the difference in frequency for gains to be -51.26% at chr8:119803251. The gain frequency is 0% for **Astrocytic tumors** and 57.14% for **Mixed gliomas** at this position. So the difference (-51.26%) is obtained by subtracting **Mixed gliomas** from **Astrocytic tumors** (5.88-57.14 = -51.26). And this percent difference is depicted in blue below the 0% line.

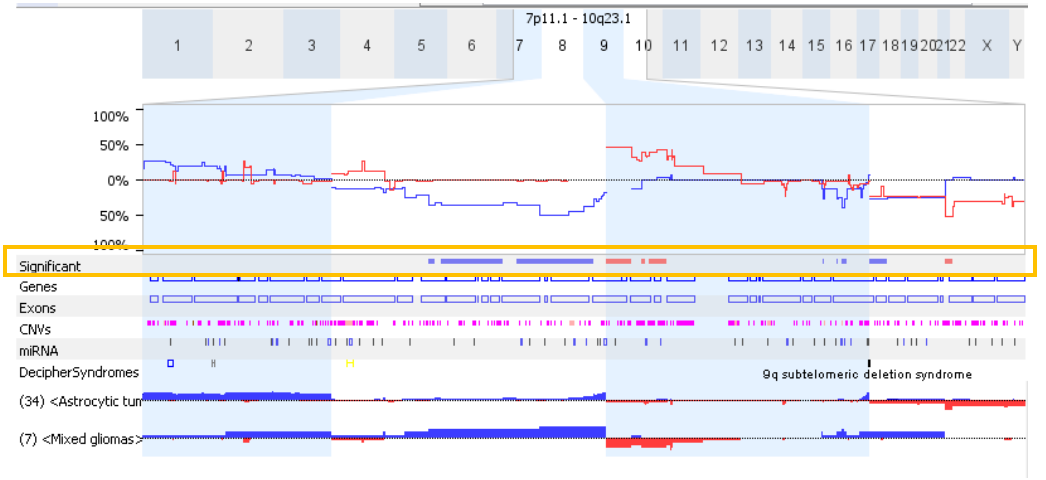
Figure showing gain and loss frequency differences:





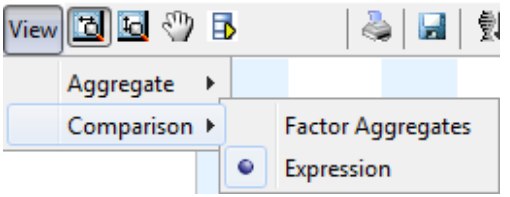
A track called **Significant** is present just below the plot (see figure below). Bars in this track indicate regions where there is a significant difference between the two groups in a comparison based on the p-value indicated at the bottom of the window as well as having a minimum difference in percentage as set in the **Differential Threshold**. Regions of loss are indicated by red bars and regions of gain, by blue bars for the copy number display. For allelic event display, LOH is indicated with brown bars and allelic imbalance with purple bars.

The **Significant** track is displayed just below the frequency difference plot (indicated below with the gold rectangle):

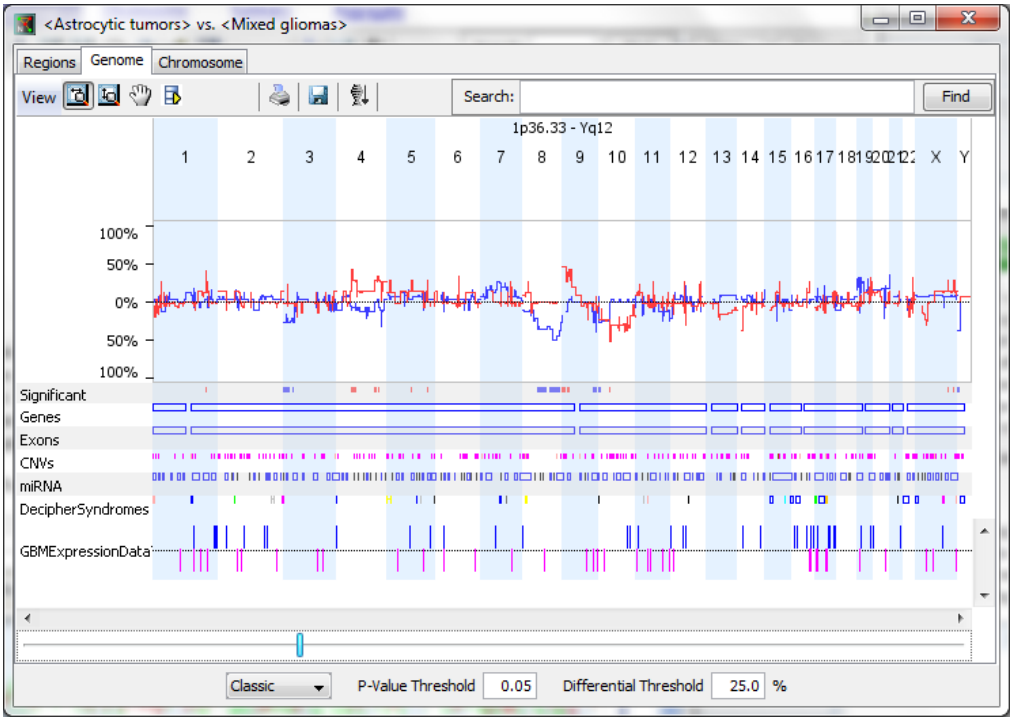


External Data Tracks

If the project contains any external data, it can be viewed in the Genome and Chromosome tabs as an additional track for each external data file loaded. It can be selected under **View->Comparison->Expression**.



For each external data file in the project, a track will be added below the annotation tracks. This allows you to immediately see which genes are up or down regulated for example in the regions of significant difference in the comparison.

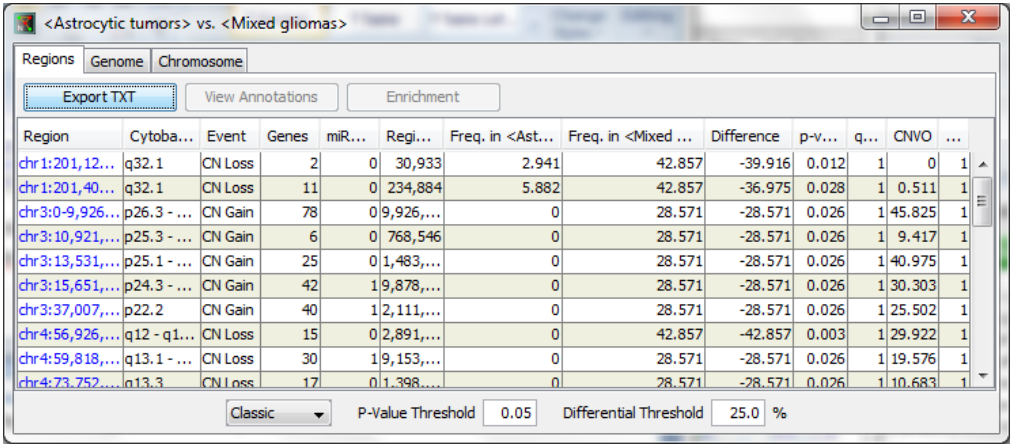


REGIONS TAB

The table in the **Regions** tab contains the regions which meet both the **Differential Threshold** and the **P-Value Threshold**. Nexus Express performs a Fishers Exact test statistical comparison to determine the likelihood of having so many events, say losses, in one group vs. having a different number of the same events in the other group based

on random chance. The p-value indicates how likely such an event is based on pure chance. Since there are multiple such tests, we also offer the Q-bound value which corrects for multiple testing by performing False Discovery Rate (FDR) correction.

Results in the **Regions** tab from **<Astrocytic tumors> vs. <Mixed gliomas>** comparison are shown in the figure below.



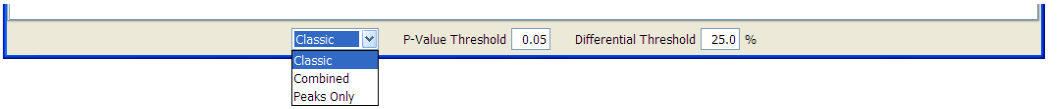
Names and description of the columns in the data table in the **Regions** tab:

Column	Description
Region	The chromosome number and base pair range of the aberrant region.
Cytoband Location	The cytoband on which this region resides
Event	**CN Gain**, **CN Loss, Allelic Imbalance, Allelic Loss**
Genes	Number of genes in this region.
miRNAs	Number of miRNAs in this region.
Region Length	Length of the region in base pairs.

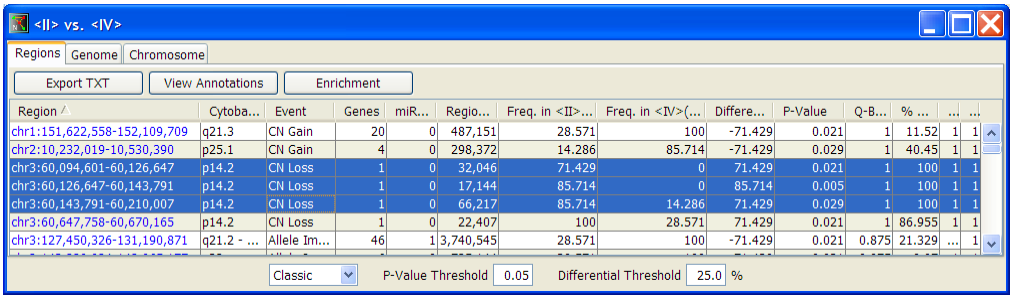
Freq. in <*name of comparison group 1***>**	Frequency of copy number of allelic event in this region for comparison group 1
Freq. in <*name of comparison group 2***>**	Frequency of copy number of allelic event in this region for comparison group 2 (the baseline)
Difference	Absolute difference between the two comparison groups. Difference is taken as **comparison group 1** minus **comparison group 2**
P-value	P-value cut off for the difference
Q-bound	Bound on the false discovery rate (FDR)
% of CNV Overlap	Percentage of this region covered with CNVs
Expression P-value	If the project contains external expression data, then a p-value is displayed for that
Methylation P-value	If the project contains methylation external data, then a p-value is displayed for that

CALCULATIONS

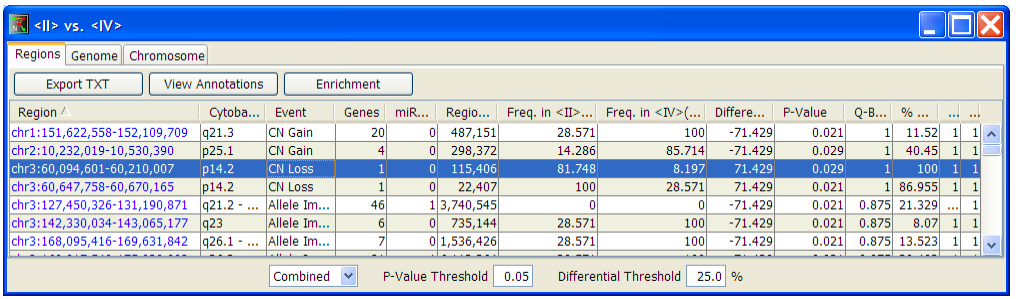
There are three choices in the type of calculation to use for Comparisons. This can be selected using a drop down menu at the bottom of the window. The **Regions**, **Genome**, and **Chromosome** tabs are updated immediately after making the calculation selection.



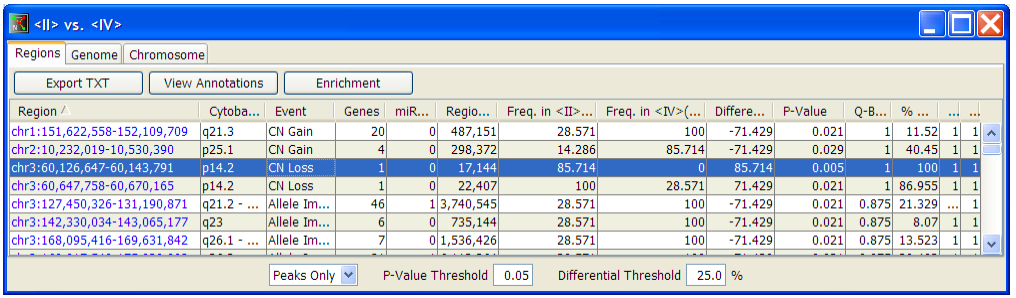
Classic - Shows each region meeting a max p-value and difference between adjacent groups. It defines a region reported in the table as one having a constant frequency. If a contiguous area for a given event has different frequencies, the region is split into multiple regions (multiple rows in the table). See the three regions adjacent to each other on chromosome 3 but with differing frequencies.



Combined – Merges all contiguous regions that meet the p-value threshold as specified and displays the largest. It defines a region as a contiguous event (e.g. gain, loss, etc.) even if areas within this event have differing frequencies. The three regions on chromosome 3 in the figure above have been combined into one region in the figure below.



Peaks only – Selects a single most significant region within a set of contiguous regions. Among all the regions displayed in the classic view for adjacent regions, selects the region with lowest p-value. Of the three adjacent regions in the classic view figure above, the region with the lowest p-value is displayed in the figure below.



The **Query Database** tool  in the Genome and Chromosome tabs of the Comparisons window allows region queries against Nexus DB to identify projects which share aberrations at the selected chromosomal location. After selecting the tool, click above or below the 0% line in the plot to search for gain and loss aberrations, respectively. See the section on *Region Searches with the Query Database Tool*, page 249, for more information.

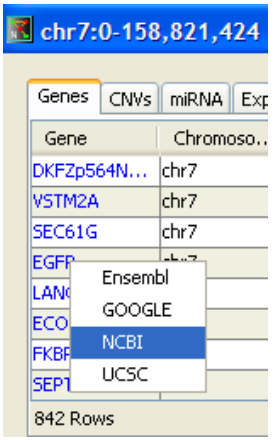
ANNOTATION DRILL DOWN

The annotation drill down feature is available in Nexus Express within several different contexts. It is available via the **Drill Down** tool on the **Genome** and **Chromosome** pages in the **Results** tab of the main Nexus Express window as well as in the **Comparisons** results window. It is also available via the **Annotation** button in the **Aggregate** page of the **Results** tab, the **Drill Down** button on the **Chromosome** page of a sample drill down window, and in the **Regions** tab in the **Comparisons** window.

Clicking on either the **Drill Down** tool or the **Annotations** button brings up a pop up window with various tabs depending on the data available for the selected regions and a table of data within each tab. If there is data for genes or CNVs, then a corresponding tab for each will appear in the new window. If there are any probe tracks displayed in the browser, then a tab for each probe type will appear if there are any probes present in the selected regions.

When using the **Drill Down** button, the information in the annotation tables represents the current viewable area in the browser. For example, when the **Genome** tab is active and the view is completely zoomed out, the frequency plot visible represents the whole genome. Clicking on drill down now will bring up a window listing all the genes and CNVs in the whole genome. If you now zoom in on a small region of chromosome 3 and then click on the drill down button, the genes and CNVs in this small region will be listed in the pop up window. This behavior applies to the **Chromosome** tab as well. When zoomed out completely within the **Chromosome** tab, clicking on the **Drill Down** tool supplies you with genes and CNVs in the entire currently selected chromosome. Upon zooming in on a region, clicking on the tool will result in listing only those genes and CNVs that are present in the base pair range displayed in the frequency plot. If any probe tracks are displayed in the browser then the pop up window will have additional tabs, one for each probe track. The probe name and its location on the genome (chromosome, start bp, and end bp) will be listed.

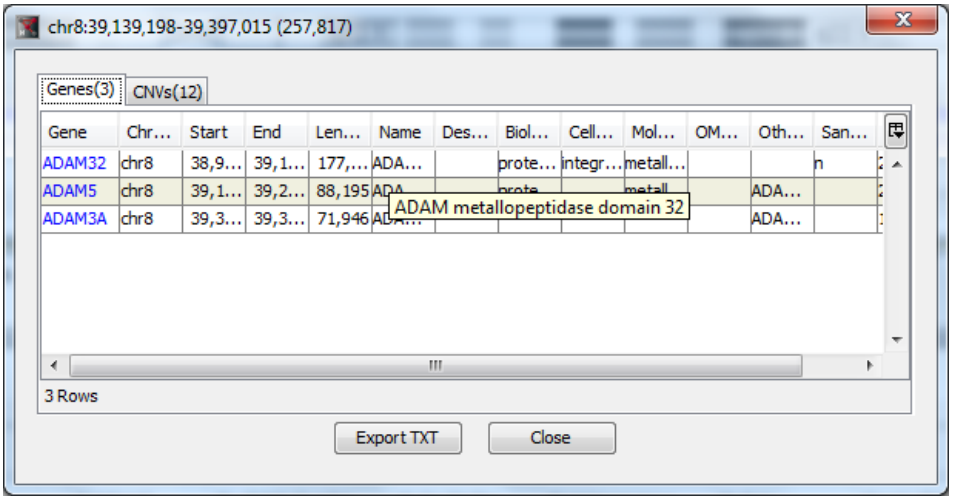
Right clicking on the Gene brings up a drop down menu from which you can choose the data base to search for more information on the gene. Selecting one of these menu options opens up the appropriate page in your default browser.



An **Export TXT** button allows you to save the data in a tab delimited text file.

The **Annotations** window contains several tabs and this depends on what is in your project as well as the options you have chosen. For example, if there are no CNVs or miRNAs present in the region you have selected to drill down upon, then the CNV and miRNA tabs will not be present. If you have not checked off any probe tracks to view, then there will be no probe tabs present.

Since some of the values in the table are quite lengthy, they may not be fully visible in the window. Moving the mouse over the **Name**, **Description**, **Biological process**, **Cellular component**, or **Molecular function** values displays the full contents of the table cell. In the figure below, positioning the mouse over the **Description** for the CPN1 gene brings up the full name of ADAM32.

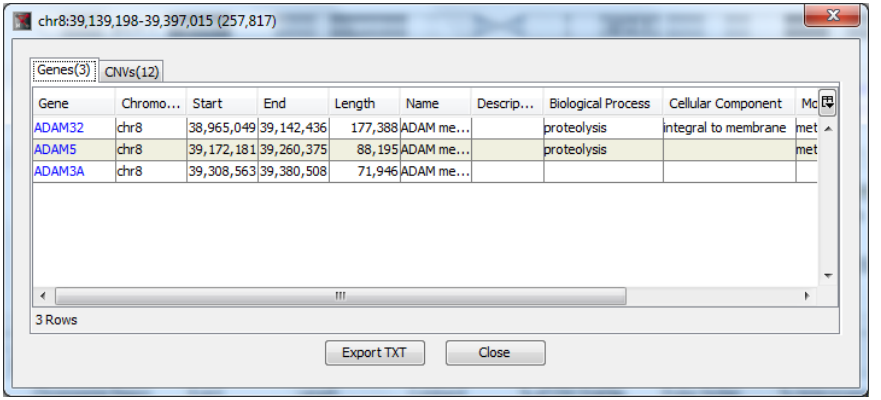


Columns in the **Genes** tab:

Column	Description
Gene	Gene symbol (UniGene ID)
Chromosome	Chromosome
Start	Base pair coordinate for the start position of the gene on the chromosome
End	Base pair coordinate for the end position of the gene on the chromosome
Biological process	A classification group defined by the Gene Ontology Consortium. This column contains the GO term annotations for this gene that belong to the biological process classification group.
Cellular component	A classification group defined by the Gene Ontology Consortium. This column contains the GO term annotations for this gene that belong to the cellular component classification group.

Column	Description
Description	A description of the gene from UCSC genome browser
LocusLink ID	The LocusLink ID associated with this gene
Molecular function	A classification group defined by the Gene Ontology Consortium. This column contains the GO term annotations for this gene that belong to the molecular function classification group.
OMIM MorbidMap Count	Number of hits for OMIM diseases for this region
Other Aliases	Other names which may be used to refer to this gene
Name	Gene name
External Data Columns	If external data is available, a column for each external data set selected to be viewed is present. The header will be the name of the external data set as given in the **External Data** tab. If regulation information is available for the gene in the corresponding external data set, then the value (**up**, **down**, **no change**) will be present.

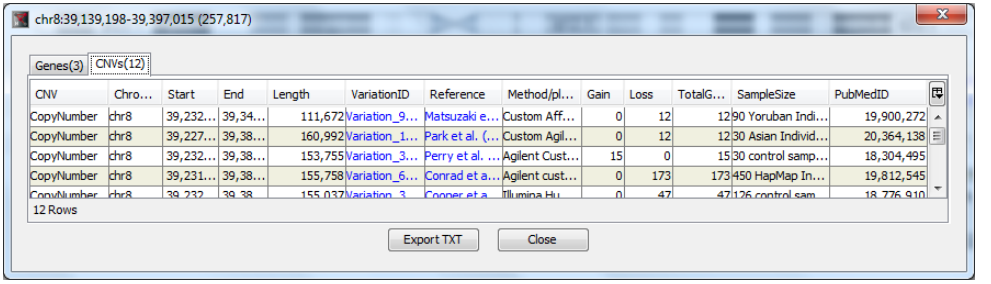
Data table in the **Genes** tab:



Columns in the **CNVs** tab:

Column	Description
CNV	The type of variation. Possible values are **CopyNumber**, **Inversion, InversionBreakpoint,** and **Indel**
Chromosome	Chromosome number
Length	Variation length in bp
End	Base pair coordinate for the ending position of the CNV on the chromosome
Variation ID	The ID from the Database of Genomic Variants which is hyperlinked to the data page on this CNV
Reference	Link to the PubMed abstract for this reference
Method/Platform	Technology used to determine the CNV
Gain	Number of gains found
Loss	Number of losses found
Total GainLossInv	Total number of aberrations found
SampleSize	Total number of samples that were examined in this reference
PubMedID	PubMed ID of the reference paper

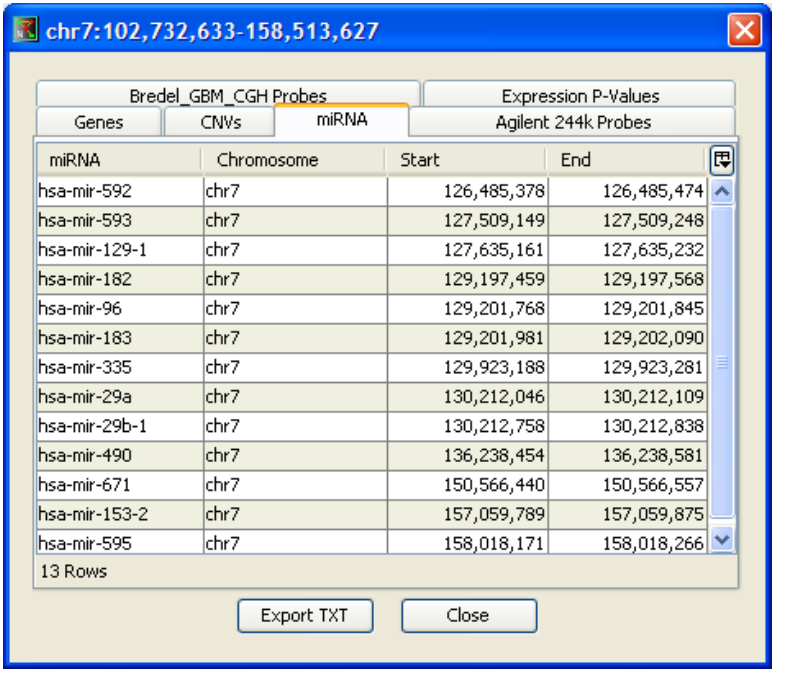
Data table in the **CNVs** tab:



Columns in the **miRNA tab**:

Column	Description
miRNA	miRNA ID
Chromosome	Chromosome number
Length	Region length in bp
Start	Base pair coordinate for the starting position of the miRNA on the chromosome
End	Base pair coordinate for the ending position of the miRNA on the chromosome

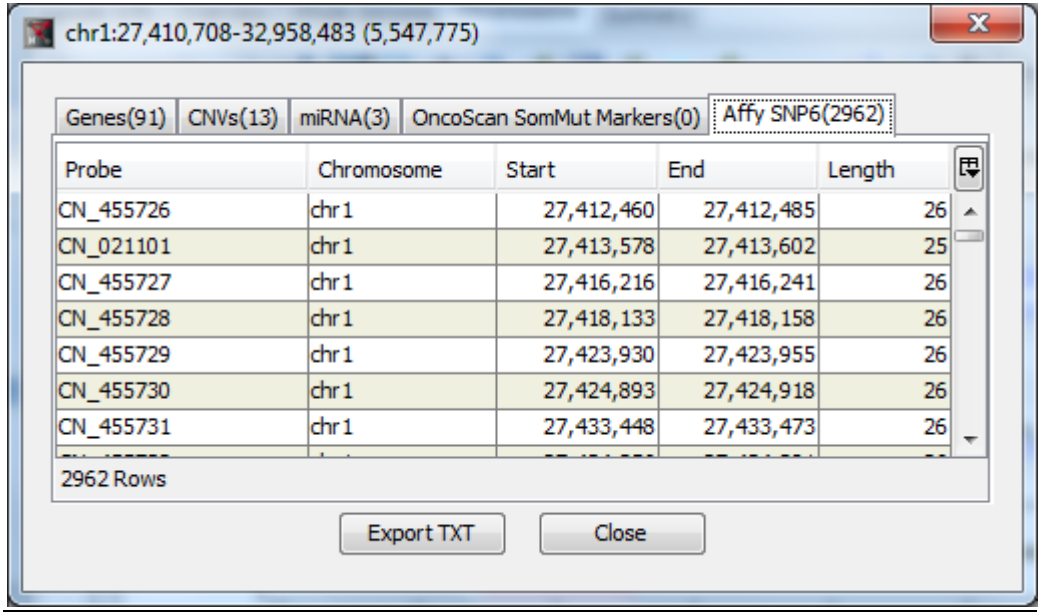
Data table in the **miRNA** tab:



Columns in the

Probe tabs:

Column	Description
Probe	Name of probe
Chromosome	Chromosome number
Start	Base pair coordinate for the starting position of the probe on the chromosome
End	Base pair coordinate for the ending position of the probe on the chromosome
Length	Length of probe in bp



Columns in the **Expression P-Values tab**:

Column	Description
Comparison	Name of expression data set
Total Genes	Total number of genes in this aberrant region
Diff. Reg. Genes	Number of genes in this region that that are differentially regulated in this data set
P-Value	The likelihood of having this many differentially regulated genes in this region by chance

Another powerful feature of Nexus Express is its ability to incorporate data from other types of experiments alongside the copy number data. This allows you to view your copy number results in concert with other data such as array expression results, miRNA, methylation, etc. and be able to generate additional hypotheses and be able to answer these with more confidence.

External data can be loaded into Nexus Express via the **External Data** tab found in the main Nexus Express window. Within this tab are additional tabs, one for each type of external data. When there is no prior external data in a Nexus Express project, the only active button in this tab is the **Add** button which is used to add external data to a project.

EXTERNAL DATA QUICK START

The external data should be in a tab delimited text file in the following format for Expression and miRNA data:

gene symbol	regulation
AKAP1	up
AMFR	up
BCL2L2	up
BEX1	up
CASP7	up
CDH13	up
CHN1	up
CRYM	up
DAP3	up
DKK3	up
EFNB3	up
EGFR	up
EHD3	up
ETS2	up
FGFR3	up
GABRG2	up

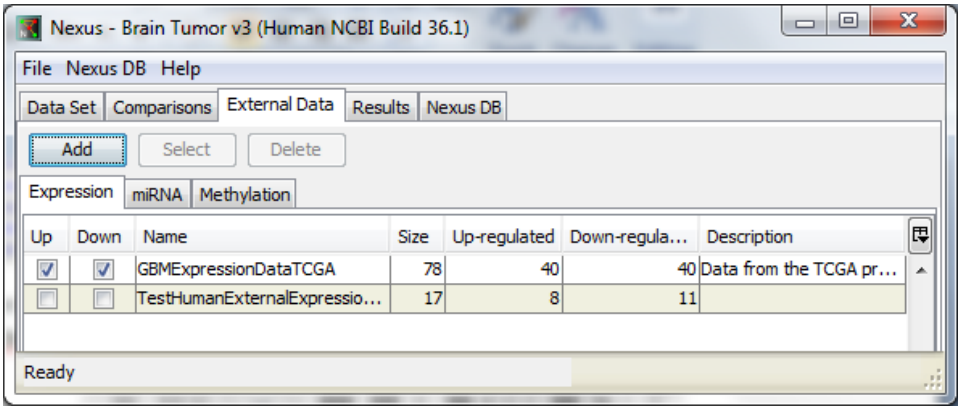
Gene Symbol and **Regulation** are required columns. Additional columns in the file will be ignored. Values for regulation are **up**, **down**, and **not regulated**.

The following format is for Methylation data which only requires the **Gene Symbol** column:

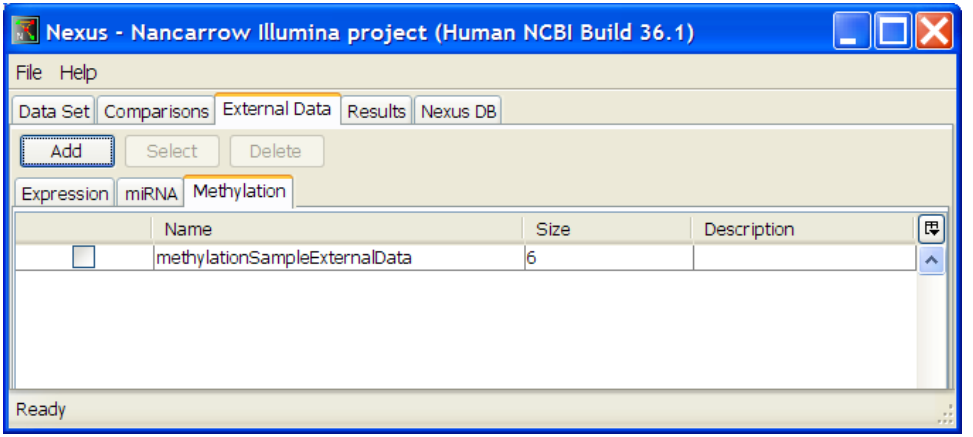
Gene Symbol
AMY1A
AMY1C
AMY1B
AMY1B
AMY1C
AMY1A
RABGA
LGR6
UBE2
PPP1
OR14C36
OR2T4

Click on the appropriate **External Data Tab** and click on the **Add** button to load the data file.

Expression external data:



Methylation external data:



Use the checkbox columns to select which type of data to display (up-regulated genes and/or down-regulated genes). The first time you load an external data file, you will need to click on the View button in the **Data Set** tab to display this data in the Genome/Chromosome displays. Subsequently, you can just hide/show the external data by marking off the checkboxes.

The columns show the number of genes in the list. It also shows how many are up-regulated and how many are down-regulated for Expression and miRNA data only. You can double-click on the **Description** column to enter any notes. The checkbox column allows you to select which data you want to display. You can choose to only show up-regulated gene or only down-regulated genes or both.

External Data Types

There are different external data types supported by Nexus Express. Below are the details for each data type and the data format that is needed to load into Nexus Express.

Expression Data

If you have expression data for your project, you can integrate these results with the copy number aberrations. The input file for this data needs to be a tab delimited text file containing regulation information as well as an identifier (either a gene name or probe name). If a gene symbol is used in the file, then this is the only file necessary. If the probe name is used then additional files need to be created which map the probes to gene symbols. This will be covered later.

Descriptions of columns that may be contained in the expression data input file:

Column	Description
Gene Symbol	The unique symbol identifying the gene (UniGene ID)
Probe	Name or symbol identifying the probe on an array. This is only used in place of the **Gene Symbol** column if you are going to use a Probe to Gene mapping file.
Regulation	Indicates whether the gene was up or down regulated or if there was no change. Possible values are **up**, and **down**, to indicate up or down regulation respectively. Any other value in this column is taken as meaning 'not regulated'
p-value	*Optional*
log-ratio	*Optional*

The input file can be in one of two formats:

FORMAT 1: GENE IDENTIFIER

For this format, the gene is identified by its gene symbol. The input file must contain a column called **Gene Symbol** and a column called **regulation**. The file must be a tab delimited text file.

FORMAT 2: PROBE IDENTIFIER

For this format, the gene is identified by the probe on the array. The first line of the input file must contain **Probe Type** followed by the name of the probe type. The next line must contain a column called **Probe** and a column called **regulation**.

Example input file with probe identifier:

Probe type	HTAv2
Probe	regulation
TC07000328.hg.1	up
TC13000345.hg.1	down
TC06002043.hg.1	up
TC11001721.hg.1	down
TC01000678.hg.1	down

In addition, two additional files need to be present. One is a file which will map the probes to the gene symbols. This file must contain a column identifying the probes used in the input file and a column containing the gene symbols to which the probes map. The column headers can be any values you choose and this file needs to be placed into the **ProbeSymbolMappings** folder in the Nexus Express installation directory. In this same directory is a file called **mappingdescriptors.txt**. A one line entry needs to be made into this file which describes the Probe type and tells Nexus Express where to look for the probe to gene mapping information. Please see the section below for the format of the **mappingdescriptors.txt** file.

MIRNA DATA

If you have miRNA data for your project, you can integrate these results with the copy number aberrations. The miRNA tab is only visible if the genome build you are using has

supporting miRNA files in the Nexus Express core directories. A **mirnas.txt** file needs to be present in the appropriate organism subfolder in the **Organisms** folder to have access to the miRNA tab.

The input file for miRNA data needs to be a tab delimited text file containing regulation information as well as an identifier (ID from Sanger Institute's miRNA database). The table below lists descriptions of columns that may be contained in the miRNA data input file.

Column	Description
miRNA	The miRNA ID from Sanger Institute's miRNA database
Regulation	Indicates whether the miRNA was up or down regulated or if there was no change. Possible values are **up**, **down**, and **not regulated**
p-value	*Optional*
log-ratio	*Optional*

METHYLATION DATA

If you have methylation data for your project, you can integrate these results with the copy number aberrations. Click on the **Methylation** tab in the External Data tab to load the methylation data. This input text file requires only a list of gene symbols (one per row).

EXTERNAL DATA ASSOCIATED FILES

Depending on the type of input external data file used, the user may need to create or edit other files. If a probe identifier rather than a gene identifier is used in the external data input file (see section above), the user needs to make changes to the existing **mappingdescriptors.txt** file in the Nexus Express installation directory.

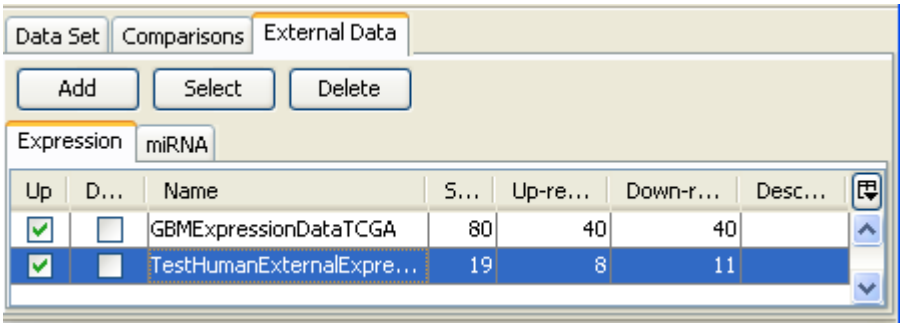
MAPPING DESCRIPTOR FILE

A file called **mappingdescriptors.txt** exists in the **ProbeGeneMappings** folder in the Nexus Express installation directory.

Column	Description
Name	A name given to this probe type.
File Name	Name of the file containing the probe to gene mapping information. This file should be placed in the **ProbeGeneMappings** folder.
Probe column	Column in the probe to gene mapping file indicating the probes
Gene Column	Column in the probe to gene mapping file indicating the gene symbols

EXTERNAL DATA TAB

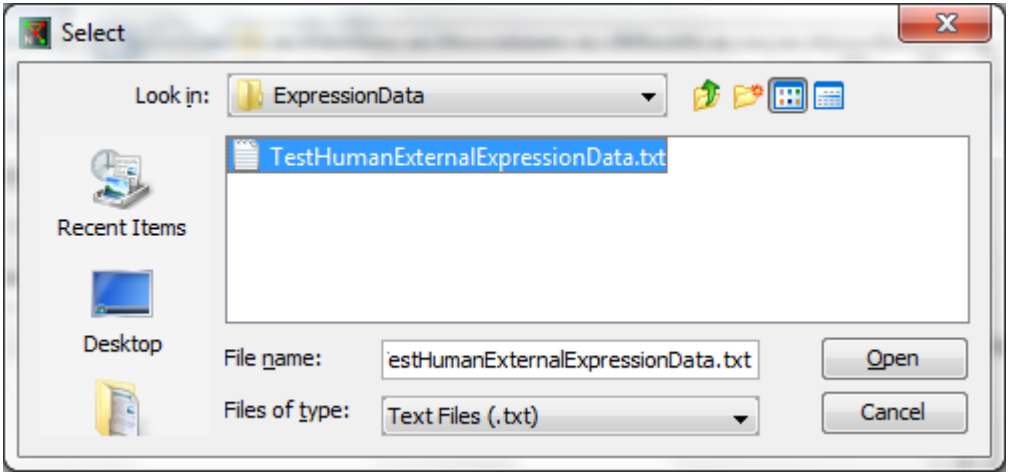
The External Data tab contains three buttons (**Add**, **Select**, **Delete**). Under these buttons are sub tabs, one for each type of external data. Each tab contains the same buttons and columns in the table. The functionality is the same for both and the details below apply to any external data type.



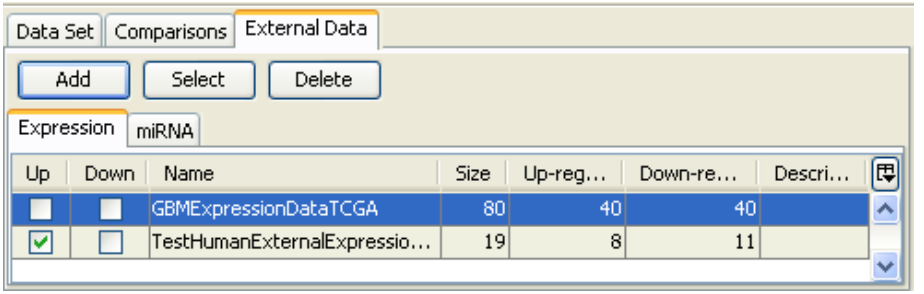
We are going to use the **Expression** tab to demonstrate functions in the **External Data** sub tabs in the sections below.

Loading External Data

Clicking on the Add button opens up the **Open** dialog where you can navigate to the location of your external data file and select the file.



If the file you selected is not in the correct form required by Nexus Express, an alert box will tell you as such. Please read the section above to format your data file correctly and try opening again. Once your data is loaded, it will show up in the appropriate tab. In the figure above, we loaded in an expression data file as seen in the **Expression** tab below.



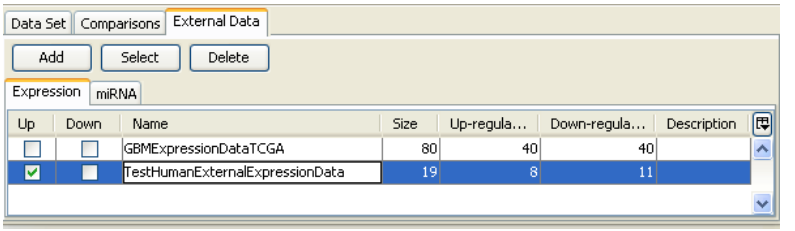
Along with the name of the data file, additional columns are present in the table. These are described in the table below.

Description of columns in the **Expression** tab data table:

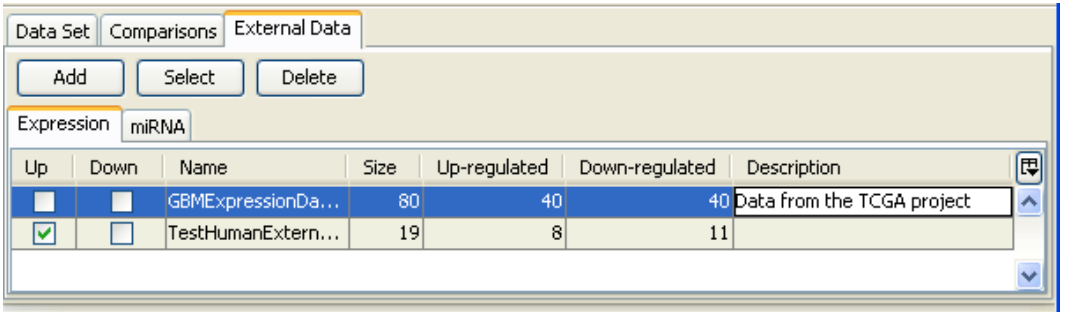
Column	Description
Up	Checkbox allows user to select viewing up-regulated genes
Down	Checkbox allows user to select viewing down-regulated genes
Name	Name of the external data file. Double click in the field to edit the name.
Size	Number of genes in the data file that matched the gene symbols in the organism's genes.txt file which is loaded into Nexus Express
Up-regulated	Number of up-regulated genes in the data file
Down-regulated	Number of down-regulated genes in the data file
Description	A description given to this data set by the user. Double click in the field to type in a description.

EDITING/DELETING EXTERNAL DATA

The name of an external data file can be edited by double clicking in the name field and editing the text.



A description for the external data set can be added by double clicking in the **Description** field and typing in some text:

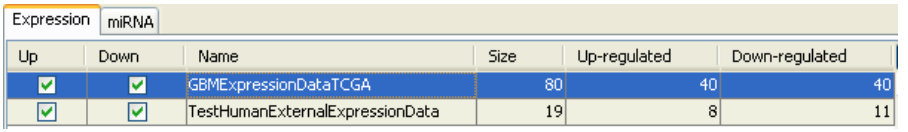


An external data file can be deleted by highlighting the row and clicking the **Delete** button. An alert box will ask you to confirm the deletion before actually deleting the data.

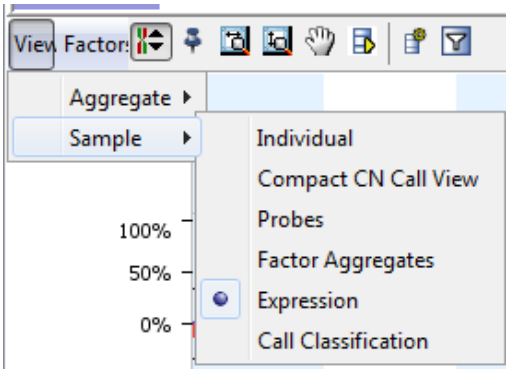
VIEWING EXTERNAL DATA

Once an external data file is loaded, the user chooses whether to view the up-regulated genes, down-regulated genes, or both. The checkboxes in the **Up** and **Down** columns

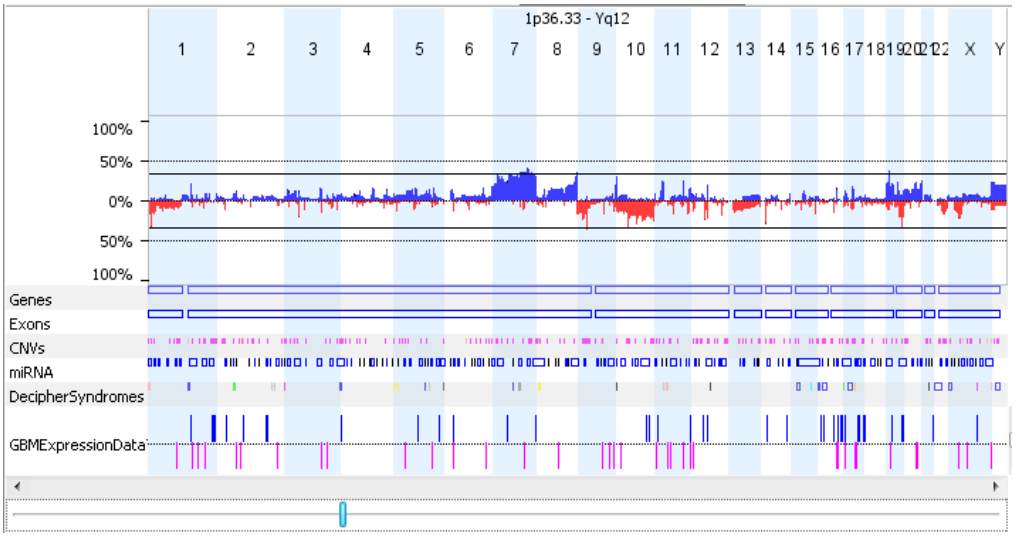
can be marked by checking off each box individually or by highlighting multiple rows and then clicking the **Select** button to mark all the checkboxes.



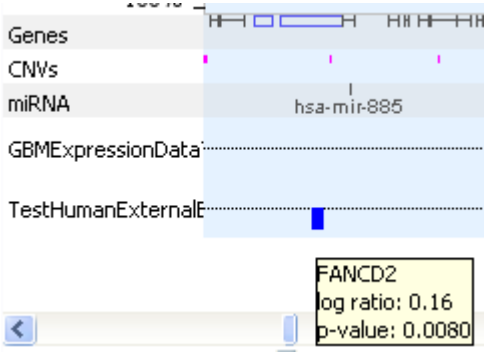
The data can then be viewed in the **Results** tab by selecting **Expression** (or any other external data item listed) in the **View->Sample** drop down menu.



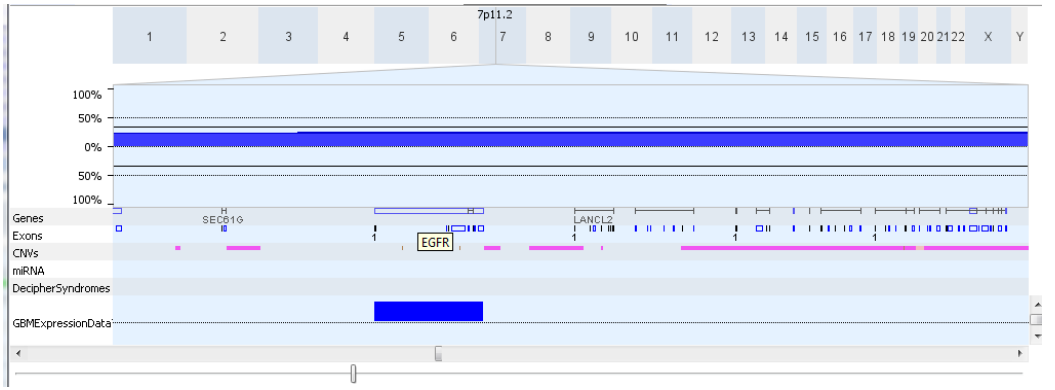
The bottom panel will now display all Expression data tracks that were checked off in the **Expression** tab. In the zoomed out view, up-regulated genes are shown as vertical ticks in magenta above the baseline and down-regulated genes are shown as vertical ticks in blue below the baseline.



Moving the mouse over the expression data bars displays the gene symbol and additionally, the log ratio and p-value if they are available in the input file.



The ability to view external data such as this makes it convenient to make correlations easily with one look. For example, in the figure below, it can be seen that an increase in copy number of EGFR correlates with an up-regulation of EGFR (as seen in the expression data track).


7p11.2
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
Y
100%
50%
0%
50%
100%
Genes
Exons
CNVs
miRNA
DecipherSyndromes
GBMExpressionData
SEC61G
LANCL2
EGFR

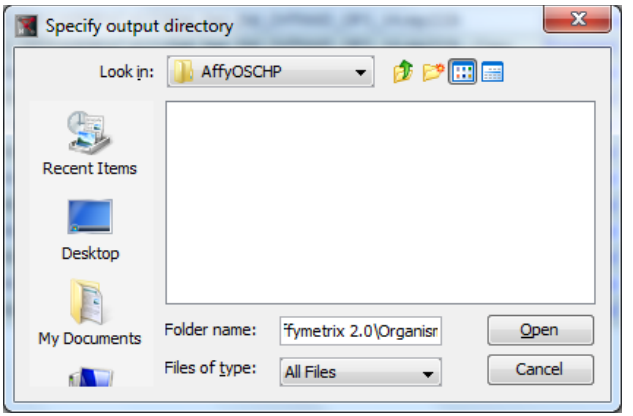
UTILITIES

The **File** menu in the main window has an option called **Utilities**. Certain utilities that users can use are housed here. Currently, there is a single utility here called **Convert .ivg to .txt**.

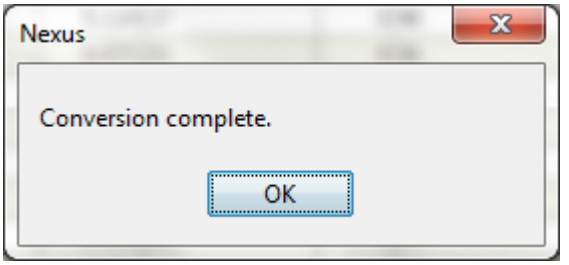
CONVERT .IVG TO .TXT

This utility allows conversion of binary .ivg files into tab-delimited text files. When Nexus Express creates a project, the processed data is stored in a number of files. Some of these files are binary (.ivg) which saves a lot of drive space.

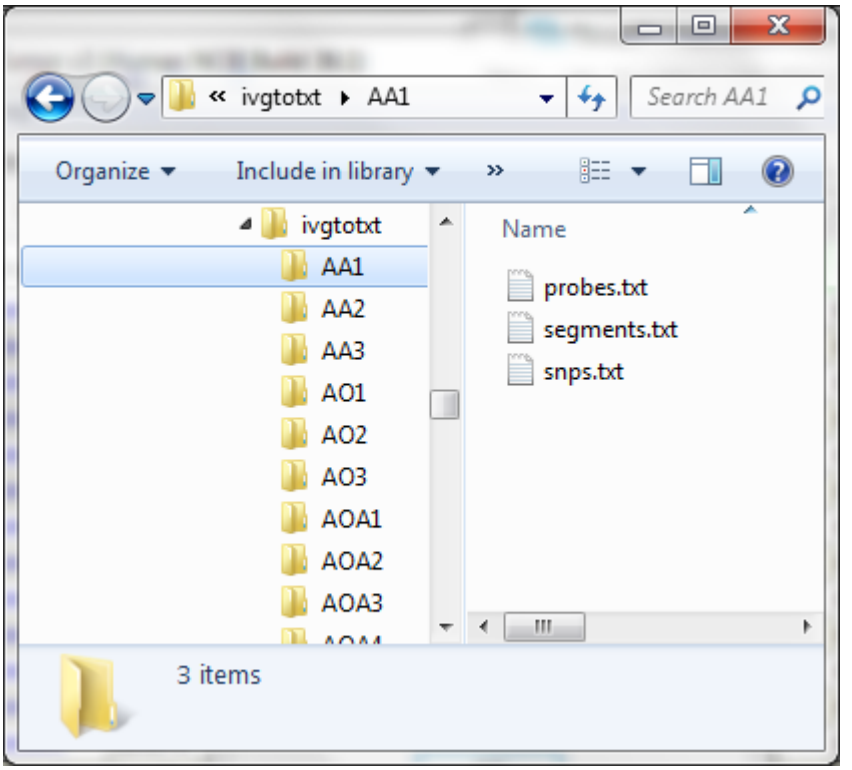
Files that are now saved in binary format are probes.ivg, segments.ivg, and snps.ivg. When using this utility, a window opens asking for a location to save the files:



If the folder does not exist, it will be created. When the conversion is finished and alert indicates this:



All ivg files will be converted to txt format and saved in sample subfolders. Each sample will have a folder bearing the sample name and the txt files will be contained within each sample folder:



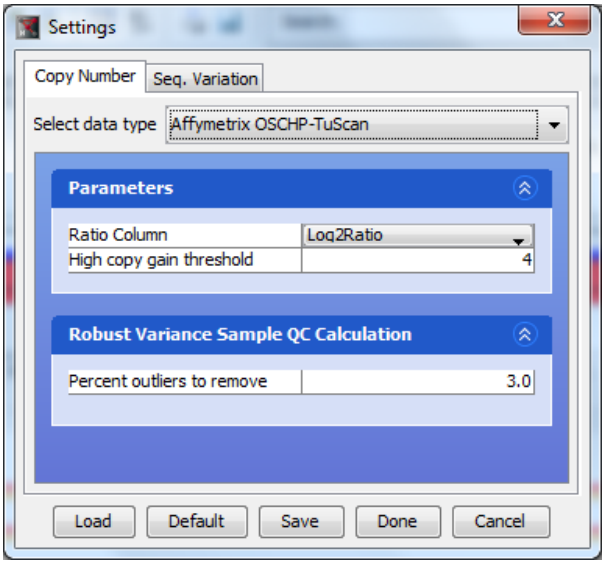
CREATE SNP BINARY FILTER FILE

This utility allows conversion of a BED file (.bed) to a binary file (.bin) for filtering based on regions specified in BED files. This is only needed with large sequence variant files where conversion to binary format increases the efficiency of loading and displaying this data. Once the BED file is converted, it is located in the **customtracks-bin** folder of the install directory.

SETTINGS

The **Settings** window allows the user to set various parameters for pre-processing and for the calling algorithm. If the input data to Nexus Express are calls (i.e. where regions are already classified into Gain/Loss), there are no settings to specify. The various settings that are available depend on the input data type. At the top of the window is the **Input Data Type** drop down menu. The data type of the samples that are going to be processed needs to be selected first and the individual settings for this Data Type can be edited. If multiple Data Types are being processed within a single Nexus Express Project, the settings for each Data Type being used in the Project needs to be set.

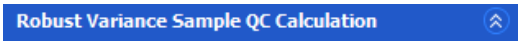
Settings window:



Input Data Types in the drop down menu include the following common types:

- Affymetrix OSCHP-TuScan
- Affymetrix OSCHP-SNP-FASST2
- Affymetrix OncoScan CBZIP

Below this drop down menu are several panels where different settings can be edited. The number and contents of the panels vary based on the Data Type selected in the drop down. Each panel has a double arrow icon in the top right which indicates whether the contents of the panel are visible or hidden. Arrows pointing up indicate visible, arrows pointing down indicate hidden. Clicking anywhere on the panel bar



will hide or display the contents. A vertical scroll bar may be present on the right to allow access to all panels within the window.

Each data type has specific groups of settings that can be specified. Some settings are available to various different data types whereas others are only specific for an individual data type.

SNP-FASST2 SETTINGS RECOMMENDATIONS

Many factors such as array data quality as well as type of array affect what settings to use. As a starting point we recommend using a significance threshold in the range of 5.0E-4 to 1.0E-5. If array quality scores are above 0.05, the **Significance Threshold** used should be at the higher end of the range. We recommend **Gain/Loss** cut-offs in the range of .10 to .20 with the highest quality arrays. For the **Min number of probes per segment** setting, Affymetrix recommends twenty probes per segment for OncoScan.

Default Values for Affymetrix OncoScan CBZIP and Affymetrix OncoScan OSCHP-SNP-FASST2 Data Type

Parameter	Value
Significance Threshold	5.0E-7
Max Contiguous Probe Spacing (Kbp)	1000
Min number of probes per segment	3
High Gain	0.7
Gain	0.1
Loss	-0.15
Big Loss	-1.1
Male Sex Chromosome Big Loss	-1.1

3:1 Sex chromosome gain	1.2
4:1 Sex chromosome gain	1.7
Homozygous Frequency Threshold	0.85
Homozygous Value Threshold	0.8
Heterozygous Imbalance Threshold	0.4
Minimum LOH Length (KB)	500
Minimum SNP Probe Density	0.0

GENERAL SETTINGS

SYSTEMATIC CORRECTION FOR ONCOSCAN DATA

Systematic correction is only applicable to the CBZIP Data Type not the OSCHP-TuScan or the OSCHP-SNP-FASST2 data types. Often samples will have artifacts that form a wave-like pattern in the probe distribution across a chromosome. This waviness can affect the accuracy of calls. Many different methods have been proposed to correct for such artifacts including those based on GC content and fragment length. Nexus Express employs a more general approach that can use any number of user-defined parameters to correct for wavy data. A drop down menu allows for a choice of methods. After selecting one of the methods, a **File** field appears below the drop down where you will need to specify the correction file. Click on the field value ("not selected" or the file name if a file is already specified here) to open the dialog for file selection. Select the correction file and click **Done**. Now when samples are processed the correction is applied to the data.

CORRECTION METHODS

Linear Correction– The bias values in the columns of the correction file (e.g. GC%, PCR fragment GC%, fragment length) are used to create a linear model whose parameters are estimated using the least squares method. The estimate is then subtracted from the probe Log(2)Ratio to obtain the corrected probe values.

Quadratic Correction– The bias values in the columns of the correction file (e.g. GC%, PCR fragment GC%, fragment length) are used to create a quadratic model whose parameters are estimated using the least squares method. The estimate is then subtracted from the probe Log2Ratio to obtain the corrected probe values.

Sequential Lowess Correction – A lowess curve is fitted, using the smoothing parameter (default=0.2) specified in the settings, to the Log2Ratio data versus each of the bias values (such as GC% of the 2kb neighborhood on either side, probe GC%, PCR fragment GC%, fragment length, etc.) in the columns of the correction file. The Log2Ratio of each probe is then corrected for this variation between the lowess curve and the mean Log2Ratio. This is performed (left to right) in the order in which the columns appear. This correction method is based on the WACA paper (Lepretre et al 2009): http://nar.oxfordjournals.org/content/38/7/e94.full. The smoothing parameter is the fraction of the data to use when estimating the curve. So for example, with 1 million points, a .2 smoothing factor would mean that the nearest 200000 points are used to estimate the value.

CORRECTION FILE FORMAT

The correction is a tab-delimited text file specifying either probe ids or genomic locations in the first column followed by any number of additional parameter columns. The first column header must be either

Probe: [probe type] or **Location**

If **Location** is used, then values would be specified using the following format:

chr1:28909776-28909827

Here is a correction file using chromosomal locations:

Location	gc4kb	gc100kb	gc1mb	gcprobe
chr1:100000811-100000843	0.391	0.3763	0.3768	0.4545
chr1:100012778-100012810	0.338	0.3791	0.3768	0.3333
chr1:100017685-100017717	0.3801	0.3799	0.3769	0.3939
chr1:100021727-100021759	0.3633	0.3798	0.377	0.6667
chr1:10003304-10003336	0.4927	0.4644	0.471	0.4848
chr1:100039503-100039535	0.3737	0.3736	0.3777	0.3636
chr1:10007388-10007420	0.46	0.4652	0.4711	0.5455
chr1:100081101-100081133	0.3828	0.3775	0.3771	0.303
chr1:10008205-10008237	0.4483	0.467	0.4711	0.4545
chr1:100083142-100083174	0.4104	0.3777	0.3771	0.5152
chr1:100116378-100116410	0.3757	0.3805	0.3769	0.2727

The correction file above provides the GC% of the 4000bp neighborhood of a probe, 100kb neighborhood, 1mb neighborhood and GC% of the probes.

RECENTER PROBES FOR THE SNP-FASST2 PROCESSED DATA

Sometimes an entire set of probes will be seen shifted up or down from the baseline and this processing step aids in moving the center of probe distribution to zero. The default is set to recenter using the **Median** but can be changed so that no re-centering is performed or user-specified regions are used to re-center (**Diploid Regions**). This option allows re-centering of data based on the median probe log ratio of user-defined region(s). This is especially useful for cancer samples with different ploidy so that one can re-center to a known diploid region.

To use this re-centering method, **Diploid Regions** needs to be selected in the Type dropdown of the **Recenter Probes** parameter. In addition, a column called **Diploid Regions** needs to be present in the Data Set table and the regions to use for re-centering need to be specified as values of this column. The regions can be specified in the following formats:

- chr2
- chr4, chr5
- chr4:0-10000
- chr4:10-10000, chr8:0-10000

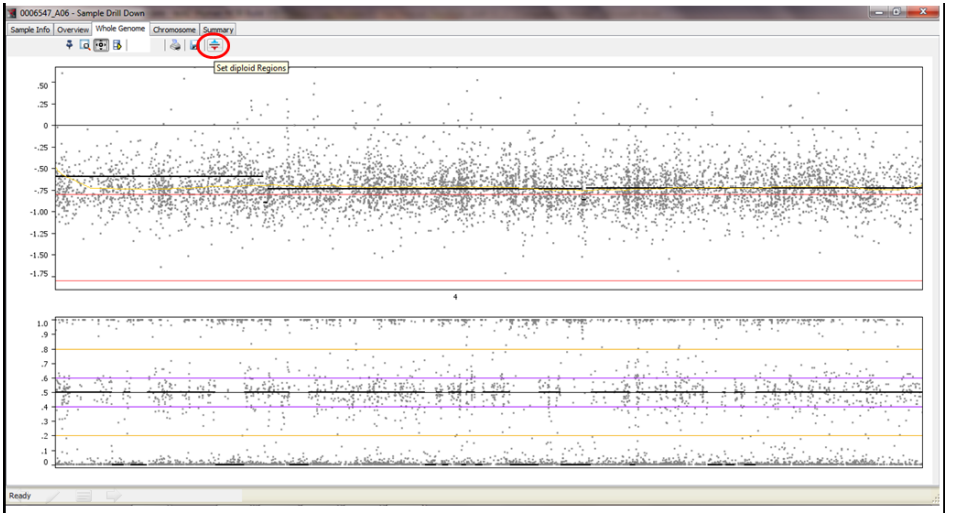
The Data Set tab with Diploid Regions column is shown below. When a cell in this column is edited, the sample is reset automatically if it was already processed and the **Status** changes from “Processed” to “Unprocessed”.

Sample	Data Type	Status	Diploid Regions	Quality	One copy gain	Two or more co
40320	Illumina	Processed		0.17	63	
40320 copy	Illumina	Processed	chr1	0.17	238	
40323	Illumina	Unproces...	chr4:100000-230799	0.201		
40325	Illumina	Processed		0.232	26	
40328	Illumina	Processed		0.176	77	
40331	Illumina	Processed		0.151	67	
40334	Illumina	Processed		0.154	242	
40338	Illumina	Processed		0.134	87	
40340	Illumina	Processed		0.132	142	
40341	Illumina	Processed		0.171	169	
40345	Illumina	Processed		0.182	175	

Clicking on **View** will then re-process the samples using the regions specified to re-center the data.

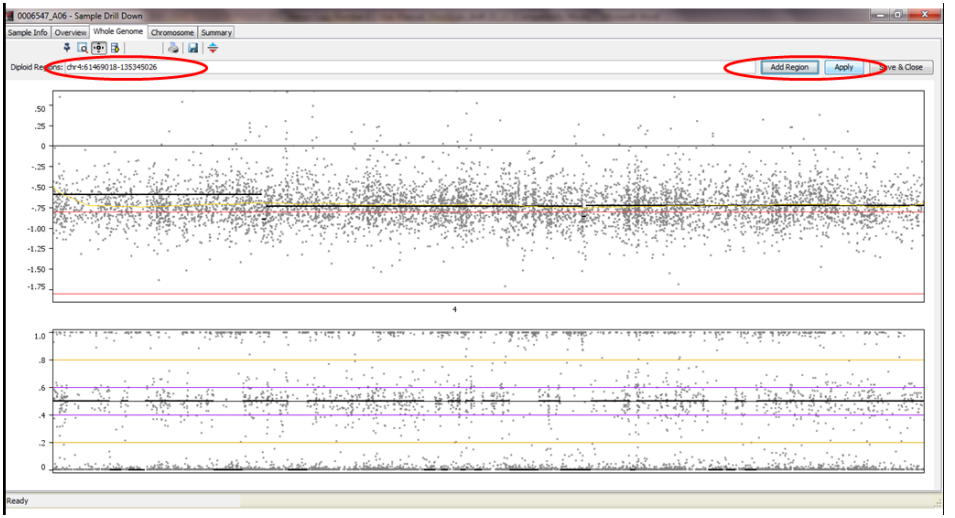
Alternative Method

An alternative method for re-centering based on diploid regions is to use the Genome tab. Zoom in so that the region of interest occupies the window, and click the “Set Diploid Regions” button.



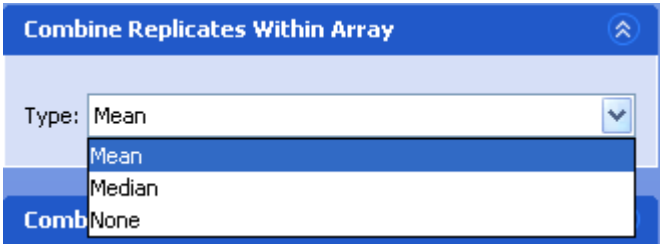
A box appears where you can enter the region. Clicking “**Add Region**” will enter the region you are viewing (you can adjust, or append additional regions manually here),

and then click "**Apply**" to reprocess the sample with the new region. Nexus Express will automatically add the region(s) identified as a factor (and create the factor if it does not already exist). Likewise, the settings for the sample will be modified to reflect the probe centering option as 'diploid regions'.



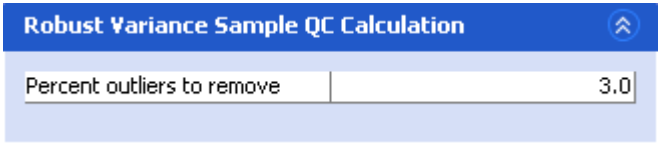
COMBINE REPLICATES WITHIN ARRAYS

A drop down menu allows the user to select how to treat replicate probes on an array. Options are **Mean**, **Median**, and **None**. The user can choose to combine these probes using the mean or median value or keep them separate (by selecting **None**).



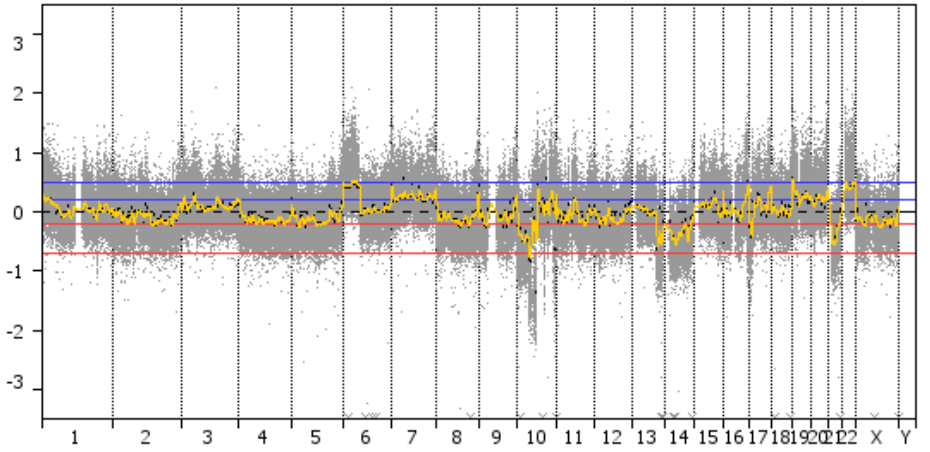
ROBUST VARIANCE SAMPLE QC CALCULATION – QUALITY SCORE

This QC step calculates the probe to probe variance and this single parameter is used to remove from calculation of the variance, the extreme outliers that one would expect to be due to copy number breakpoints. It is meant to measure how much successive probes differ from each other on average. The score is displayed in the **Quality** column in the Data Set tab. It is computed by first ordering by magnitude the difference between adjacent probes and then removing a percentage of the probes that fall at the top and bottom of the list. For example, if probes ordered along the genome have values [1, 2, 1.5, 2.1] then the differences would be [1, -.5, .6]. Then a percentage of the probes would be removed (from calculation of the mean variance) from the top and bottom of the variance spectrum. The percentage to remove is set by the user via the **Settings** window (Percent outliers to remove). If the setting is 3.0 (for 3%), then half of this percentage of probes (1.5%) are removed from the top of the list and the other half, from the bottom. The default value is 3% but can be changed individually for each data type for which QC calculation is available. Double click in the field to change the value.

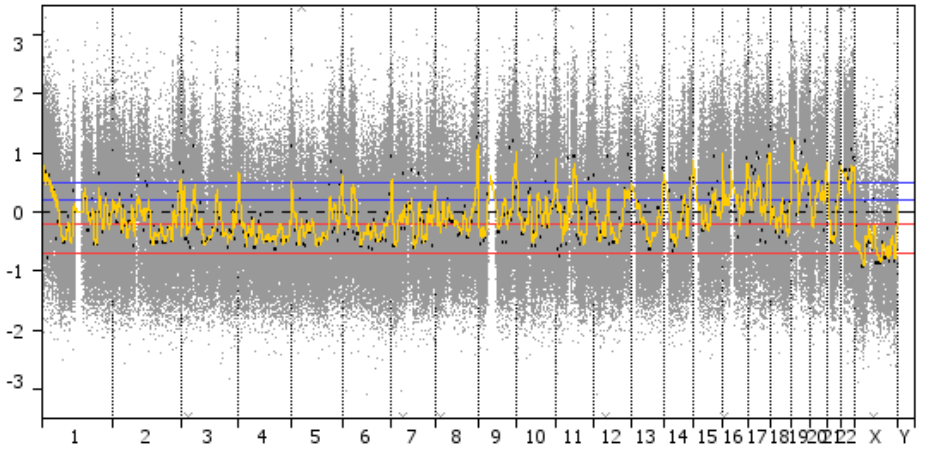


Please note that the QC scores are only used to compare relative QC scores between arrays and the outliers are not removed from processing. It allows you to determine how good a sample is. The lower the QC score, the better quality the sample is. . You can visually see the difference between samples with high and low QC scores in the figures below. The sample with a higher QC score (second figure) shows a lot more noise as compared to the sample with the lower QC score.

Sample with lower QC score: 0.09:



Sample with higher QC score: 0.67:

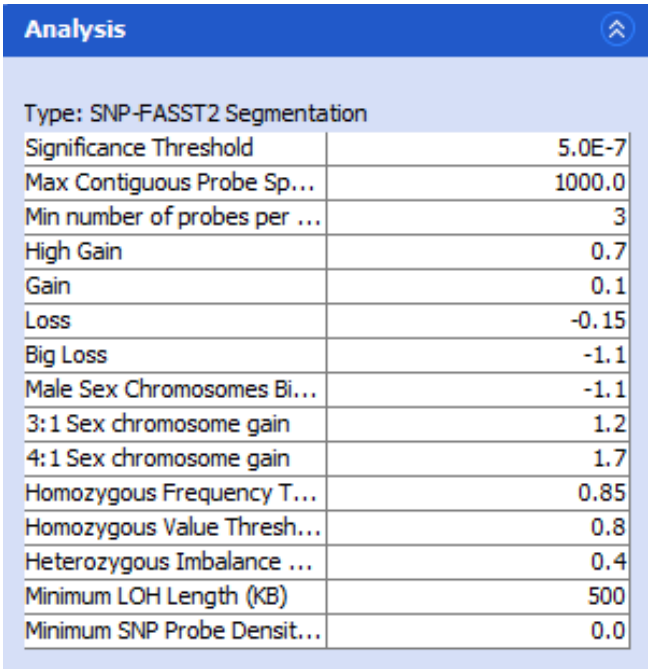


Having the QC value gives you a better idea of how much confidence you have in an individual sample's contribution to the aberration calls. If you feel that a sample's scores are just too low, you can sub select only the "good" samples for further analysis or you can completely remove the 'bad' samples from your project by deleting those samples.

A quality score of 0.15 – 0.2 is generally considered the cut-off range for “good” samples.

ANALYSIS

The **Analysis** panel for the SNP-FASST2 algorithm allows the user to select the call algorithm and specify the calling thresholds and cut-offs.



OncoScan v2 data in CBZIP files needs to be processed with the SNP-FASST2 segmentation algorithm. OncoScan FFPE Assay kit data provided in OSCHP files has already been processed by the TuScan algorithm in the OncoScan Console Software. In this case, the only parameter to adjust is the threshold for defining high gains. It is possible to reprocess the OncoScan FFPE Assay kit data using the SNP-FASST2 algorithm (described in *SNP-FASST2 Segmentation Algorithm,* page 24 - it combines data from both copy number and B-allele frequencies for segmentation).

THRESHOLDS

THRESHOLDS FOR MAKING CALLS

The following thresholds are applicable to all chromosomes but different values can be specified for the autosomal and the sex chromosomes.

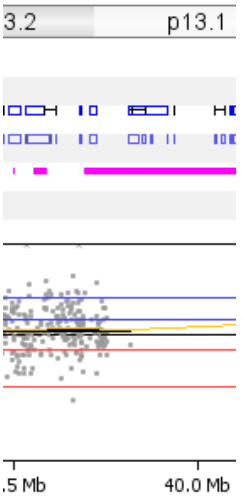
Please note that the recommendation here will not be appropriate for every array and the selection of parameters should be done by the user based on his own arrays and knowledge of the underlying biology. The following recommendations should be used as a starting point.

Big Loss: This will indicate the threshold for a homozygous loss (both copies deleted). Theoretically, this would be log(0/x) or a very large negative number. A threshold of -1 offers a good performance.

Loss: This will indicate the threshold for a hemizygous loss (one copy of the pair is deleted). Theoretically, this would be log(1/2) or -1. Due to the inherent noise in the experiment and for detection of mosaic loss, a threshold of -0.2 for non-SNP arrays and -0.18 for SNP arrays offers a good performance.

Gain: This will indicate the threshold for a single copy gain. Theoretically, this would be log(3/2) or 0.585. Due to the inherent noise and possible mosaicism of the sample, a threshold of 0.18 for SNP arrays and 0.2 for non-SNP arrays offers a good performance.

High Gain: This will indicate the threshold for two or more copy gain. Theoretically, this will be equal to or greater than log(2) or 1. Due to the inherent noise in the experiment a threshold of 0.6 offers a good performance.



High Gain	0.41
Gain	0.13
Loss	-0.23
Big Loss	-1.1

Male Sex Chromosomes Big Loss

Threshold used for loss calls on male sex chromosomes.

3:1 Sex Chromosome Gain

Log ratios between this value and the value in 4:1 Sex Chromosome gain parameter are considered as 3:1 gain in certain cases. See sectio

Thresholding for Sex Chromosomes, page 214, for explanation.

4:1 Sex Chromosome Gain

Log ratios above this value are considered as 4:1 gain in certain cases. See section

Thresholding for Sex Chromosomes, page 214, for explanation.

Significance Threshold

A significance threshold needs to be set in the **Analysis** panel. The **Significance Threshold** is used to adjust the sensitivity of the segmentation algorithm. The smaller the number, the less sensitive the algorithm is in creating a new segment. So, if some known aberrations are not being called because they are too small, then increase this value. To obtain sound results, we have found that ideally this setting is inversely proportional to the number of probes. The larger the number of probes, the smaller the value used for this setting. We have processed a large number of probes at a setting of 1E-6 or lower.

Max Contiguous Probe Spacing

Here, the user can specify the maximum spacing between adjacent probes before breaking a segment. This is especially useful for areas across the centromere and would allow the software not to make any calls where there are no probes and also for achieving focused arrays where there are probes only in parts of the chromosome. The default value is 1000 Kb (1 Megabase) but can be changed by double clicking in the field and typing in a new value.

Min. Number of Probes Per Segment

Here the user can set the minimum number of probes required in order to create a new segment. A value of three is sufficient for normal samples. For OncoScan tumor samples, this value should be increased to 20 to eliminate small CNVs.

Additional Thresholds For B-Allele Frequency Data for SNP-FASST2 processed data

The following thresholds work in concert and sequentially. Nexus Express first determines whether a region can be called as an LOH area by testing the criteria that a set percentage of probes (defined by the homozygous frequency threshold) have B-allele frequency values falling in the range defined by the homozygous value threshold. If this criterion holds then the region is an LOH. For example, if this threshold is set to 95%, Nexus Express will see if more than 95% of the probes in that region are

homozygous probes (e.g. AA or BB). If this is the case then the region is called Homozygous. If the region is not found to be homozygous, Nexus Express then checks to see if this region can be considered an allelic imbalance (where there are more copies of one allele over another). If those same probes have B-allele frequency values falling in the range defined by the heterozygous imbalance threshold, then it is labeled as an area of allelic imbalance. If the probes do not fall in that range, then the region is said to contain equal copies of each allele and therefore has no allelic aberration.

Homozygous Frequency Threshold

States what percent of the probes have to be homozygous (based on the homozygous value threshold) for the segment to be called an LOH region.

Homozygous Value Threshold

Using x as the value in this field, if in a region, the percentage of probes with B-allele frequency values between x and 1 and values between 0 and 1 – x equals or exceeds the homozygous frequency threshold, then this region will be considered an LOH.

E.g. with the homozygous frequency threshold set at .95 and the homozygous value threshold set to .8; if the B-allele freq. of 95% of the probes in a segment is greater than .8 or less than .2 then that area will be considered an LOH area.

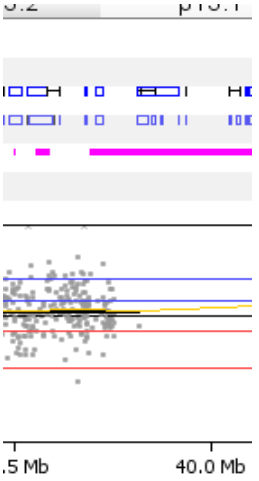
Heterozygous Imbalance Threshold

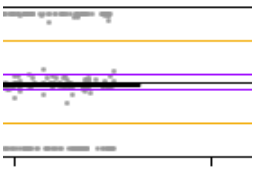
Using x as the value in this field, if in a region that is found to be not homozygous, the segment value is found to be less than x or greater than 1-x, then this region will be considered as an allelic imbalance.

E.g. with the homozygous frequency threshold set at .95 and the heterozygous value threshold set to .45; if the segment is not called an LOH, then if the B-allele freq. of 95% of the probes in the segment is between 0.2 and 0.45 or is between 0.55 and 0.8 then that area will be considered an allelic imbalance.

B-allele frequency plot thresholds

The figure below shows the threshold lines on the B-allele frequency plot. The homozygous frequency threshold is in yellow and the heterozygous imbalance threshold line is in purple. These thresholds are mirrored around the .5 (50%) line along the Y axis.





Homozygous Frequency Thr...	0.95
Homozygous Value Threshold	0.8
Heterozygous Imbalance Th...	0.4
Minimum LOH Length (KB)	500
Minimum SNP Probe Density...	0.0

Minimum LOH Length (MB)

This allows you to set a minimum length for a region to be called an LOH. If the length of the region equals or exceeds this value, then it will be called an LOH (if it also meets the other criteria for an LOH).

Minimum SNP Probe Density (Probes/MB)

Specifies the minimum number of SNP probes needed in an area in order to make calls.

THRESHOLDING FOR SEX CHROMOSOMES

This feature provides automatic threshold adjustment for sex chromosomes based on the sample gender.

AFFYMETRIX ARRAYS

For the **Affymetrix OSCHP**, there is no need for **Control Gender.** Only the **Gender** column is used because of the way the OSCHP output was normalized by the Affymetrix OncoScan Console software -- the X chromosome is normalized against a pool of female samples and the Y chromosome is normalized using a pool of male samples as the reference. Nexus Express takes this into account and adjusts thresholds accordingly. As long as the **Gender** column is present for these **Affymetrix** data types, thresholds are adjusted automatically. If a **Control Gender** column is present in the Data Set table for these arrays, the values here are ignored.

REPROCESSING

If after data has been processed, the gender values are changed in either gender column, the processing status is reset to **unprocessed** and these samples must be processed again to view results. This only requires changes to the thresholds and the sample is not re-segmented.

HOW AUTOMATIC THRESHOLDING WORKS

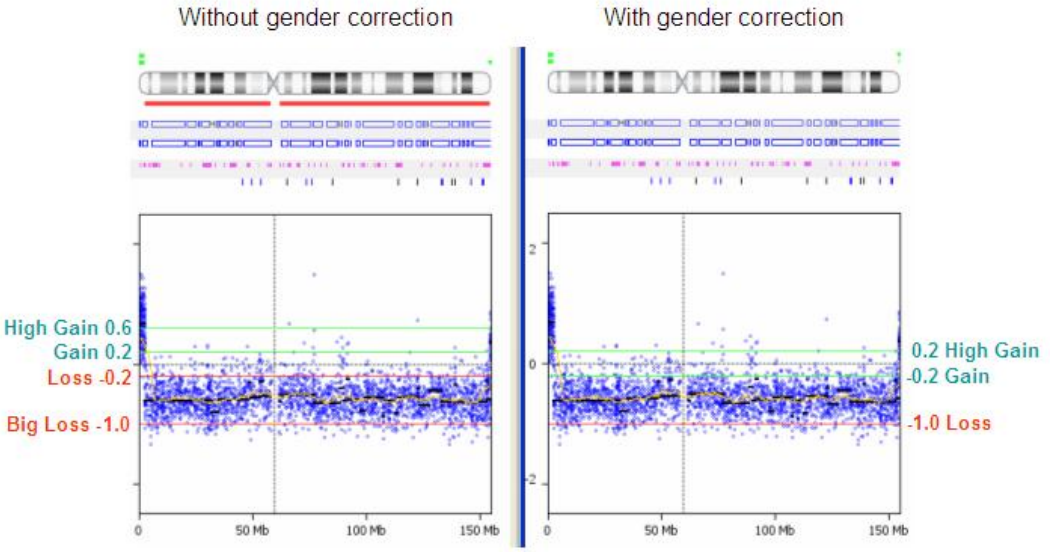
For determining gain or loss on autosomal chromosomes the gender of the experimental vs. control sample does not matter as normal copy numbers on both samples would be the same. But for the sex chromosomes, it does matter. If the experimental and control samples are not from the same individual they could differ in sex. Let's take an example where the experimental sample is Female and the control sample is Male, then for the X chromosome, normally the experimental sample would

have two copies and the control would have one copy. Now to determine if there was a one or two copy gain, one would need to use different thresholds than those for Gain and High Gain for autosomal chromosomes.

Signal Intensity Ratio for experimental (F) vs. control (M)	Call on X chromosome	Call on autosomal chromosome
2:1	No gain (Normal expectation of X chromosome for female sample is 2 and for a male sample it is 1)	High gain – Two or more copy gain since this is equivalent to a 4:2 gain.
3:1	Gain - Single copy gain	NA
4:1	High Gain – Two or more copy gain	NA

This table could be expanded to include all variations of gender in comparing experimental to control samples and in some cases we would need to define additional thresholds (for the 3:1 and 4:1 ratios) in addition to the **Gain** and **High Gain**.

The figure below shows the effect of using the **Factor:Gender** and **Factor:Control Gender** columns in a project. The plot on the left is not corrected for gender but the plot on the right is. The experimental sample is male and control sample is female. On the X chromosome if the intensity of experimental is ½ that of the control, the results show a one copy loss when not corrected for gender (left panel). But when corrected for gender, there is no loss as a normal male sample should have ½ the intensity of the female sample for chromosome X.



The thresholds for the plot on the right have been corrected for gender and are different than that for the plot on the left. Below are the settings that were used for analysis:

High Gain	0.6
Gain	0.2
Loss	-0.2
Big Loss	-1.0
3:1 Sex chromosome gain	1.0
4:1 Sex chromosome gain	1.5

Based on the control sample, the thresholds for Gain, High Gain, etc. shift for the sex chromosomes with gender correction. Also note that now there is only one threshold for loss (one red line on plot) and loss of a single copy on the X chromosome uses the threshold of -1 which is the setting for a homozygous loss (complete loss of that chromosome) for the autosomal chromosomes.

DATA TYPE SPECIFIC SETTINGS

The next section covers settings for each data type. All panels available for each particular Data Type are listed below. However, only the panels in the **Settings** window that are specific for each corresponding data type are described in detail below. Please refer to the information above for detailed descriptions of the remaining panels that are in common among data types.

AFFYMETRIX ONCOSCAN OSCHP-TUSCAN

Parameters:

Log2Ratio

High copy gain threshold

Robust Variance Sample QC Calculation

AFFYMETRIX ONCOSCAN OSCHP-SNP-FASST2

Parameters

Call Rate Greater than: A text box that allows you to enter a call rate.

Relative standard deviation less than: A text box that allows you to adjust the standard deviation

Re-center Probes

Analysis

Robust Variance Sample QC Calculation

AFFYMETRIX ONCOSCAN CBZIP

Parameters

Call Rate Greater than: A text box that allows you to enter a call rate.

Relative standard deviation less than: A text box that allows you to adjust the standard deviation

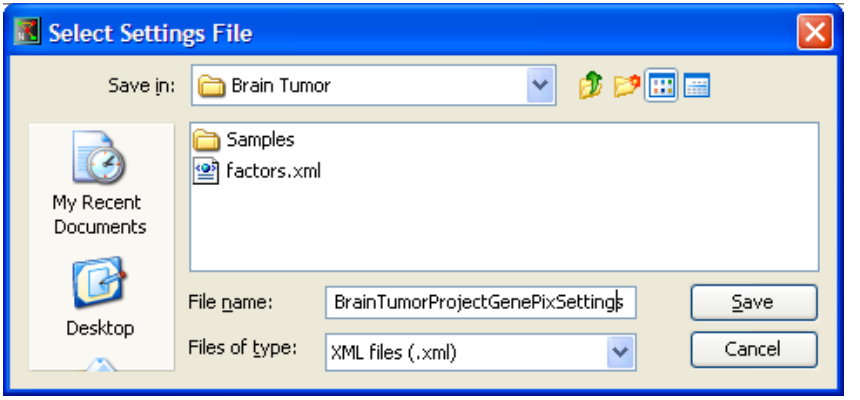
Combine Replicates Within Array

Systematic Correction

Re-center Probes

SAVING THE SETTINGS

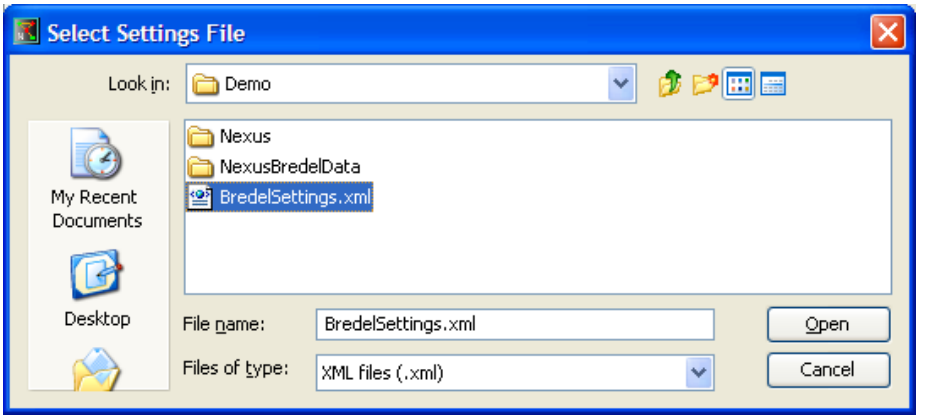
Once you have adjusted the settings for any one or several data types, you can save these settings and give it a name. Clicking on **Save** brings up the **Select Settings File** dialog where you can specify a name for this settings file.



The file is saved as an XML document. Please note that the settings file contains values for all data types and not just the one which is currently viewed in the **Settings** window when you click **Save**.

LOADING A SETTINGS FILE

Once you have previously saved a Settings profile, you can open it and use those settings to process samples. Clicking on **Load** in the **Settings** window brings up the **Select Settings File** dialog.



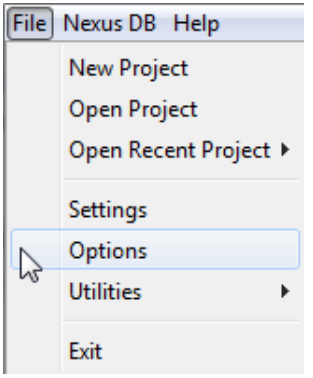
Navigate to your saved xml settings file and click **Open**. The saved settings profile will now be applied to each data type. Click on the **Done** button in the **Settings** window to finish.

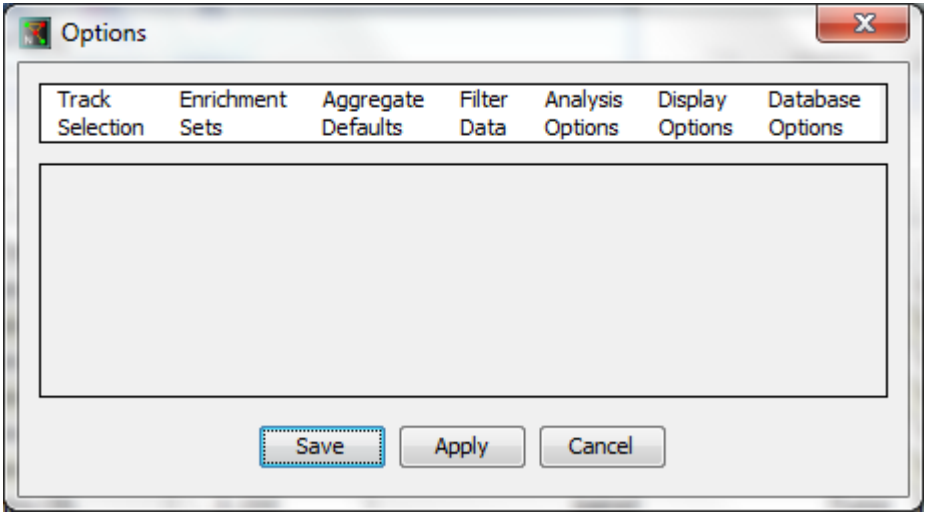
RESTORING DEFAULT VALUES

To restore the default settings, click on the **Default** button and then click **Done**. Please note that pressing **Default** will restore default values for all data types and not just for the data type that is selected in the **Input Data Type** drop down when the **Default** button is pressed.

OPTIONS

The **Options** window accessed from the **File** menu in the main menu bar contains many items for which the user can set defaults for his projects before analysis begins. These are meant to be set once for a user's typical project.



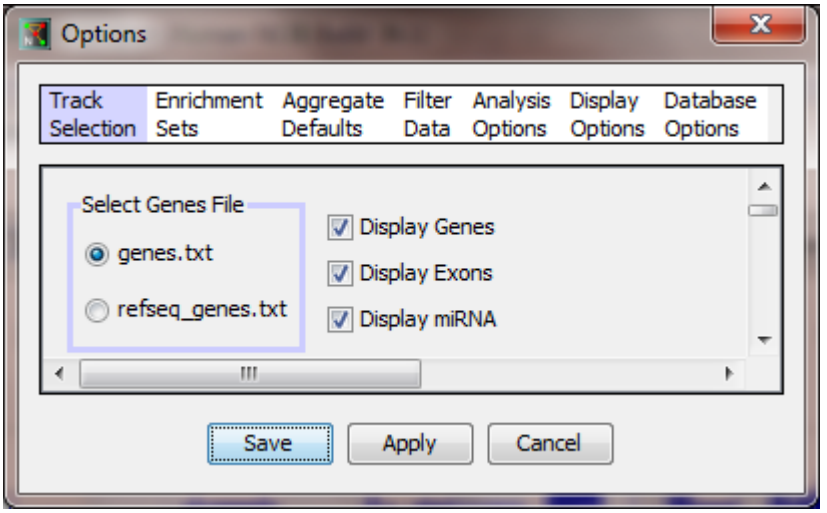


Clicking the **Apply** button applies your changes immediately. Clicking the **Save** button will save these settings so that they will persist each time you open Nexus Express. The sections below will cover each item.

TRACK SELECTION

This allows you to select which tracks will be displayed in the **Genome** and **Chromosome** genome browsers. **Display Genes, Display Exons,** and **Display miRNAs** are selected by default but you can un-check these boxes to hide these tracks on the browser.

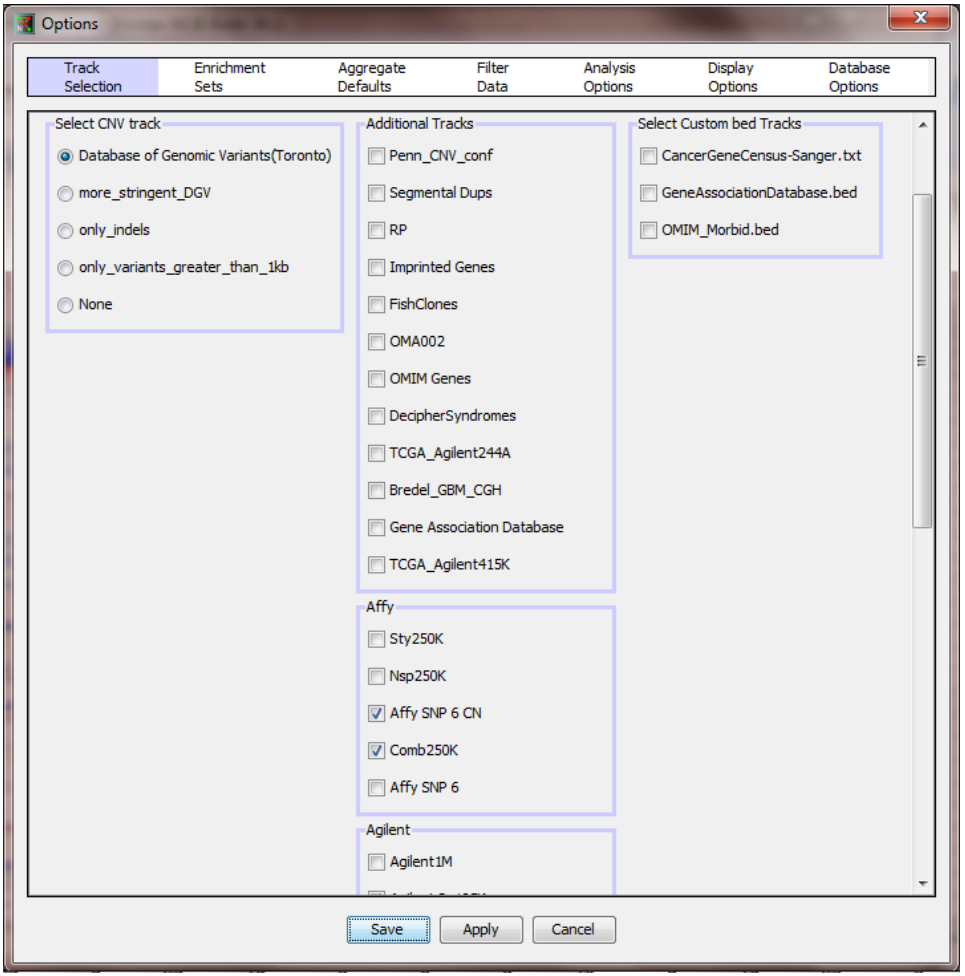
For the Genes track, there is an option for the type of genes to display. Selecting **genes.txt** will display the Ref Seq genes along with the UCSC predicted genes. Selecting **refseq_genes.txt** will display only the Ref Seq genes.



For the CNV track, the Toronto Database of Genomic Variants is selected by default but you can make other selections here. You can choose to display only those CNVs that are indels or only those that are >1Kb in size. If you have created your own custom CNV database and added it to Nexus Express, it will show up in the selection panel. You can also select **None** to remove any CNV tracks from the display. You can also add tracks for

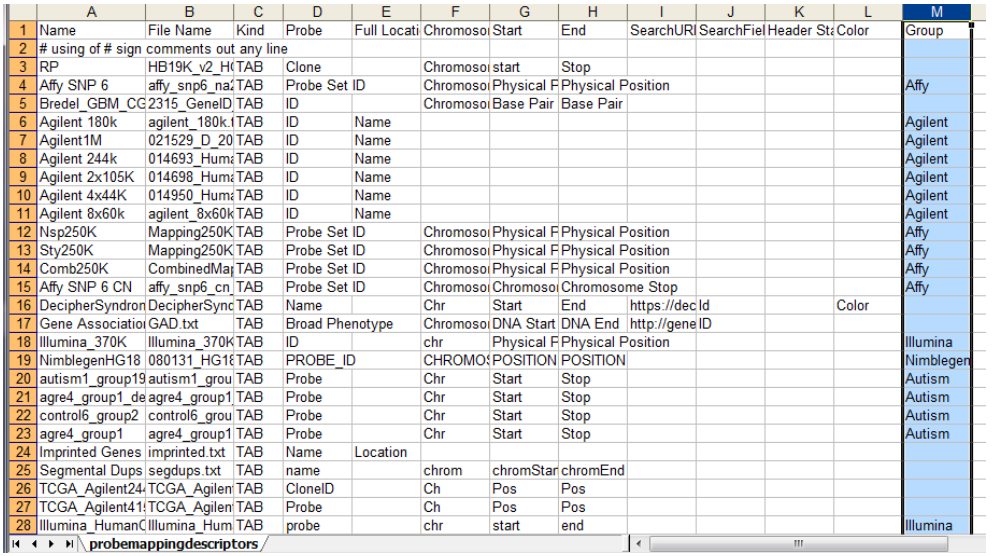
various probe types onto the browser by selecting them here. Included with Nexus Express are **DecipherSyndromes** from the Decipher database (https://decipher.sanger.ac.uk/) containing information about chromosomal microdeletions/duplications/insertions, translocations and inversions. Clicking on the aberration opens the browser on the data page for the particular aberration. Another option under probe tracks is **Gene Association Database**. This is from the Genetic Association Database at the NIH (http://geneticassociationdb.nih.gov/cgi-bin/index.cgi; Lin BK, et al. Tracking the epidemiology of genes in the literature: the HuGE published literature database. Am J Epidemiol 2006; 164:1-4) containing human genetic association studies of complex diseases and disorders. Clicking on items in this track opens the browser displaying the record from the Genetics Association Database.

The annotation track options are organized into groups by CNV, platform and bed file options. After the initial installment, you may see fewer options as all tracks are not provided in the installer to keep the installer size manageable. Once you update annotations, many additional tracks will be displayed (see figure below).



The groups are defined in the **probemappingdescriptor.txt** file in the **Probemappings** folder of the organism/genome build folders.

The probemappingdescriptors.txt file shown in Excel with the Group column highlighted:



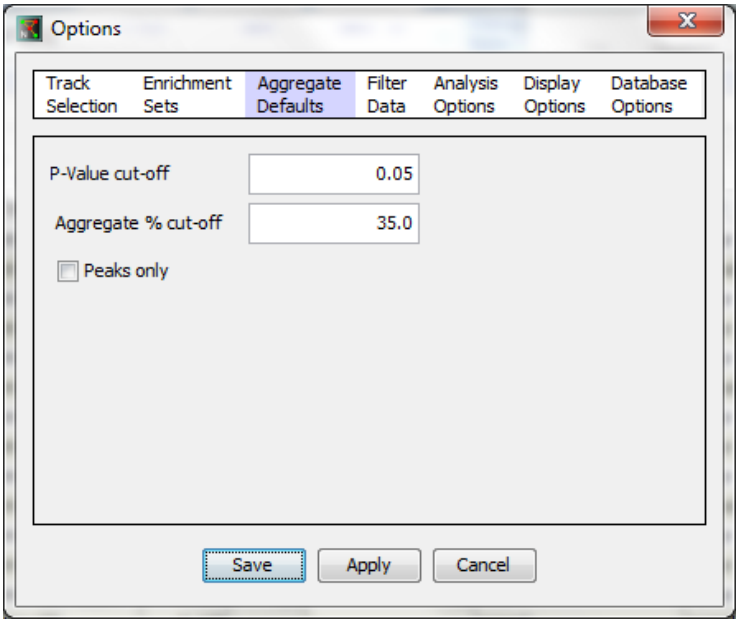
	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Name	File Name	Kind	Probe	Full Locati	Chromoso	Start	End	SearchURI	SearchFiel	Header St	Color	Group
2	# using of # sign	comments out	any line										
3	RP	HB19K_v2_H	TAB	Clone		Chromoso	start	Stop					
4	Affy SNP 6	affy_snp6_na	TAB	Probe Set	ID	Chromoso	Physical P	Physical Position					Affy
5	Bredel_GBM_CG	2315_GeneID	TAB	ID		Chromoso	Base Pair	Base Pair					
6	Agilent 180k	agilent_180k.t	TAB	ID	Name								Agilent
7	Agilent1M	021529_D_20	TAB	ID	Name								Agilent
8	Agilent 244k	014693_Huma	TAB	ID	Name								Agilent
9	Agilent 2x105K	014698_Huma	TAB	ID	Name								Agilent
10	Agilent 4x44K	014950_Huma	TAB	ID	Name								Agilent
11	Agilent 8x60k	agilent_8x60k	TAB	ID	Name								Agilent
12	Nsp250K	Mapping250K	TAB	Probe Set	ID	Chromoso	Physical P	Physical Position					Affy
13	Sty250K	Mapping250K	TAB	Probe Set	ID	Chromoso	Physical P	Physical Position					Affy
14	Comb250K	CombinedMa	TAB	Probe Set	ID	Chromoso	Physical P	Physical Position					Affy
15	Affy SNP 6 CN	affy_snp6_cn	TAB	Probe Set	ID	Chromoso	Chromoso	Chromosome	Stop				Affy
16	DecipherSyndron	DecipherSync	TAB	Name		Chr	Start	End	https://dec	Id		Color	
17	Gene Associatio	GAD.txt	TAB	Broad Phenotype		Chromoso	DNA Start	DNA End	http://gene	ID			
18	Illumina_370K	Illumina_370K	TAB	ID		chr	Physical P	Physical Position					Illumina
19	NimblegenHG18	080131_HG18	TAB	PROBE_ID		CHROMOS	POSITION	POSITION					Nimblegen
20	autism1_group19	autism1_grou	TAB	Probe		Chr	Start	Stop					Autism
21	agre4_group1_de	agre4_group1	TAB	Probe		Chr	Start	Stop					Autism
22	control6_group2	control6_grou	TAB	Probe		Chr	Start	Stop					Autism
23	agre4_group1	agre4_group1	TAB	Probe		Chr	Start	Stop					Autism
24	Imprinted Genes	imprinted.txt	TAB	Name	Location								
25	Segmental Dups	segdups.txt	TAB	name		chrom	chromStar	chromEnd					
26	TCGA_Agilent24	TCGA_Agilen	TAB	CloneID		Ch	Pos	Pos					
27	TCGA_Agilent41	TCGA_Agilen	TAB	Probe		Ch	Pos	Pos					
28	Illumina_HumanC	Illumina_Hum	TAB	probe		chr	start	end					Illumina

probemappingdescriptors

To change groupings just change the value in the **Group** column. You can use any group name you would like and these will be displayed in the **Options->Track Selection** window arranged in alphabetical order under the section called **Additional Tracks**.

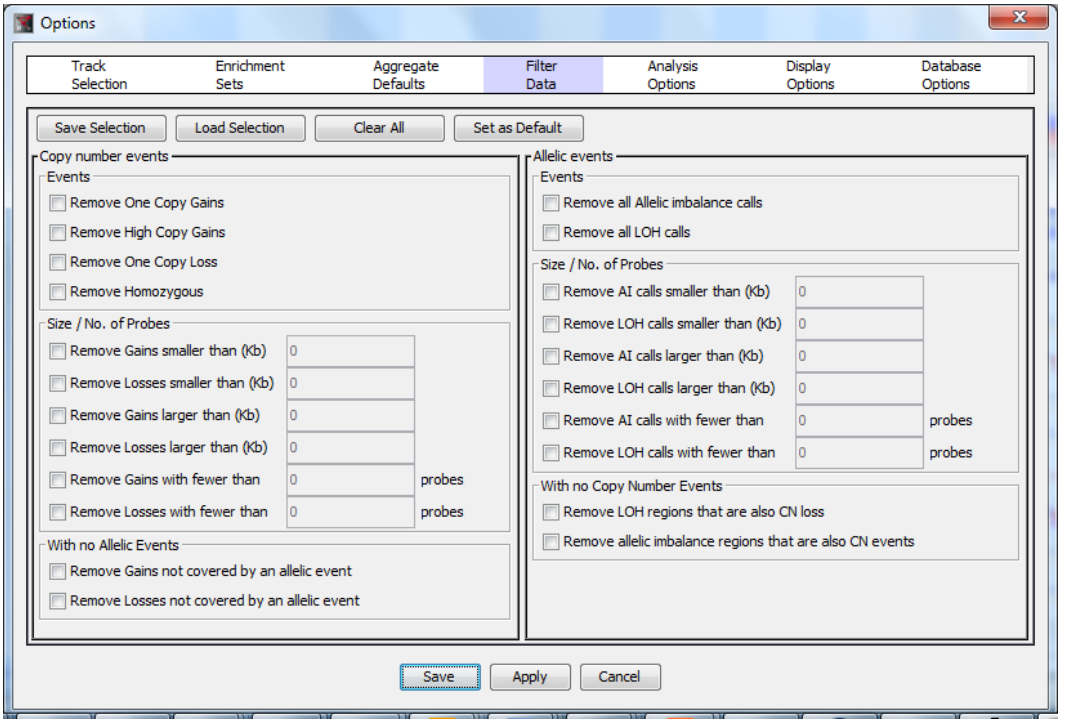
AGGREGATE DEFAULTS

Defaults for the **P-Value cut-off**, **Aggregate cut-off**, and **Peaks Only** can be set from this menu. The values here will be applied to each new project that is processed/viewed. These parameters are also available in the bottom of the **Aggregate** page in the **Results** tab so that they can be changed easily when reviewing results within a single project.



FILTER DATA

The **Filter Data** Options Dialog contains several checkboxes as shown in the figure below. When a filter is used, any affected data columns in the Data Set tab will have a * next to the header indicating that the column contains filtered data. These options do not persist between sessions. If the program is closed and opened again, all the checkboxes will be cleared.



REMOVE ONE COPY GAIN

Removes all single copy gain calls from the graphical display and report tables

REMOVE ONE COPY LOSS

Removes all single copy loss calls from the graphical display and report tables

REMOVE HIGH COPY GAIN

Removes all high copy gain calls from the graphical display and report tables

REMOVE HOMOZYGOUS LOSS

Removes all homozygous loss calls from the graphical display and report tables

REMOVE CALLS FROM SEX CHROMOSOMES

Removes all calls from the sex chromosomes

REMOVE CNVS FROM AGGREGATES

Selecting this checkbox affects the results displayed in the **Aggregate** tab. If an Aggregate region is covered entirely with CNVs then this region will be removed from the table.

Aggregate tab with the Filter option turned off. Note that the highlighted region:

Region	Region Length	Event	Genes	Freque...	P-Value	Expression...
chr1:11,778,627-12,097,833	319,207	Loss	9	35.849		1
chr3:49,816,666-50,210,877	394,211	Loss	10	35.849		1
chr9:34,665,730-35,274,387	608,657	Loss	11	35.849		1
chr10:40,510,772-44,480,418	3,969,646	Loss	11	37.736		1
chr10:71,848,002-73,959,499	2,111,497	Loss	20	35.849		1
chr10:93,736,047-95,281,405	1,545,358	Loss	11	37.736		1
chr10:101,579,695-102,390,503	810,808	Loss	13	37.736		1
chr10:104,069,813-104,311,666	241,853	Loss	9	35.849		1
chr10:104,932,406-105,488,033	555,627	Loss	10	35.849		1
chr10:134,957,770-135,217,395	259,626	Loss	10	35.849		1
chr12:47,523,503-47,960,153	436,651	Loss	16	39.623		1
chr12:55,368,067-55,684,468	316,401	Loss	8	41.509		1
chr14:20,001,631-20,416,412	414,782	Loss	12	35.849		1
chr19:54,294,853-55,034,815	739,963	Loss	32	35.849		1
chr22:36,554,608-36,740,736	186,129	Loss	6	35.849		1
chrX:37,835,812-44,155,346	6,319,535	Loss	20	35.849		1
chrY:17,109,480-23,684,896	6,575,417	Gain	23	39.623		1

Aggregate tab with the Filter option turned on. Note that the region on chr 10 (134,957,770-135,217,335) is no longer present:

Region	Region Le...	Event	Genes	Frequenc...	P-Value	Expression ...
chr1:11,778,627-12,097,833	319,207	Loss	9	35.849		1
chr3:49,816,666-50,210,877	394,211	Loss	10	35.849		1
chr9:34,665,730-35,274,387	608,657	Loss	11	35.849		1
chr10:40,510,772-44,480,418	3,969,646	Loss	11	37.736		1
chr10:71,848,002-73,959,499	2,111,497	Loss	20	35.849		1
chr10:93,736,047-95,281,405	1,545,358	Loss	11	37.736		1
chr10:101,579,695-102,390,503	810,808	Loss	13	37.736		1
chr10:104,069,813-104,311,666	241,853	Loss	9	35.849		1
chr10:104,932,406-105,488,033	555,627	Loss	10	35.849		1
chr12:47,523,503-47,960,153	436,651	Loss	16	39.623		1
chr12:55,368,067-55,684,468	316,401	Loss	8	41.509		1
chr14:20,001,631-20,416,412	414,782	Loss	12	35.849		1
chr19:54,294,853-55,034,815	739,963	Loss	32	35.849		1
chr22:36,554,608-36,740,736	186,129	Loss	6	35.849		1
chrX:37,835,812-44,155,346	6,319,535	Loss	20	35.849		1
chrY:17,109,480-23,684,896	6,575,417	Gain	23	39.623		1

REMOVE ALL ALLELIC IMBALANCE CALLS

Removes all allelic imbalance calls from the graphical display and report tables

REMOVE ALL LOH CALLS

Removes all LOH calls from the graphical display and report tables

REMOVE COPY NUMBER REGIONS SMALLER THAN (KB)

This allows you to specify a minimum length for regions to be displayed as copy number loss or gain.

REMOVE REGIONS OF ALLELIC EVENTS SMALLER THAN (KB)

This allows you to specify a minimum length for regions to be displayed as allelic loss or gain events.

REMOVE COPY NUMBER REGIONS GREATER THAN (KB)

Only displays copy number gain or loss regions smaller than the specified length.

REMOVE ALLELIC EVENT REGIONS GREATER THAN (KB)

Only displays allelic events smaller than the specified length.

REMOVE LOH REGIONS THAT ARE ALSO CN LOSS

This displays and reports only the LOH regions that have no corresponding copy number loss (copy neutral LOH).

REMOVE ALLELIC IMBALANCE REGIONS THAT ARE ALSO CN EVENTS

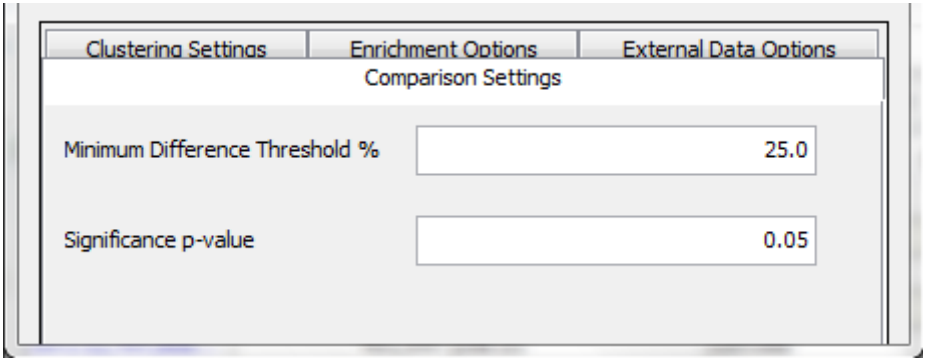
This displays and reports only the copy number events that have no allelic imbalances in the same region.

ANALYSIS OPTIONS

These are a collection of options for various analyses. Each has a sub tab: Comparison Settings and External Data Options.

COMPARISON SETTINGS

The **Comparison Settings** tab allows you to edit the default values used when creating comparisons between groups in the **Comparisons** tab.



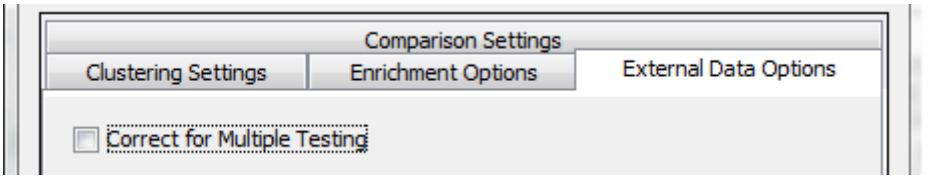
Minimum Difference Threshold %: This is the absolute difference in percent between the two groups being compared.

Significance p-value: Cut-off to use in the two-tailed Fisher's exact test to determine significance.

The **P-value threshold** and the **Differential threshold** are present to filter out regions such that only those regions that meet both criteria will be displayed in the **Regions** tab. Only aberrations of the same type are compared to each other for the differential threshold criteria. Gains to gains or losses to losses are compared. Gains to losses are not compared.

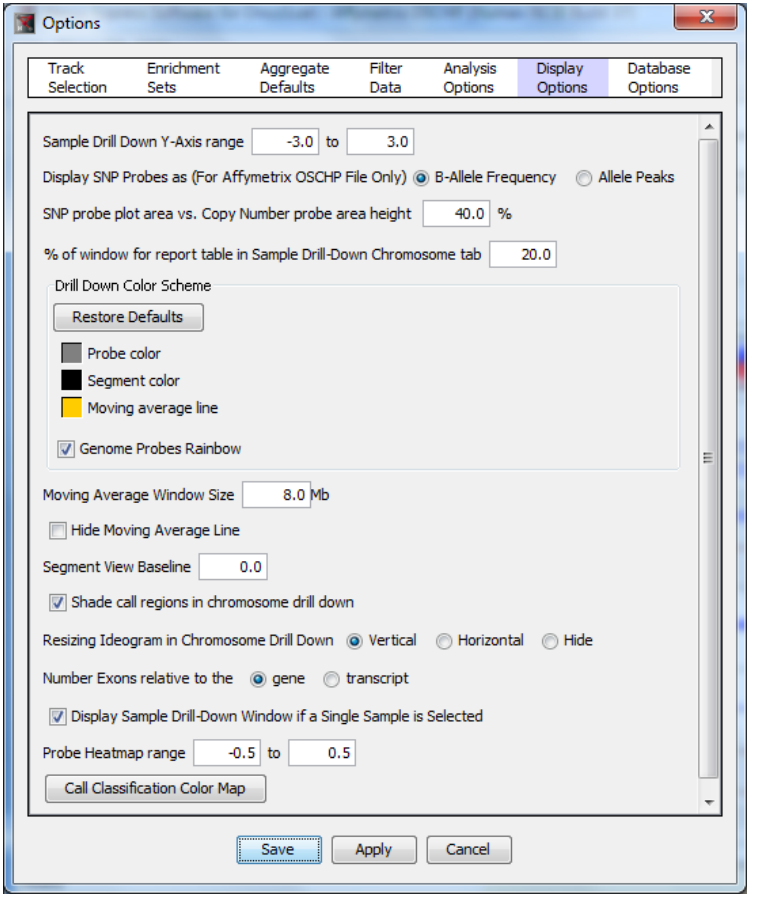
EXTERNAL DATA OPTIONS

For external data analysis, you can choose to display the Q-bounds which are the p-values corrected for multiple testing by marking the **Correct for Multiple Testing** checkbox.



DISPLAY OPTIONS

Allows you to set some parameters for the graphical displays.



The Y-axis range in the sample drill down plots can be specified by editing the values in the boxes for **Sample Drill Down Y-Axis Range**.

For OSCHP data type, SNP probes can be displayed either as B-allele frequency or as Allele Peaks, other data types are only displayed as B-allele frequencies.

For OncoScan™ OSCHP data, Nexus Express supports two modes of display for the SNP-Probe information: B-Alelle Frequency and Allele Peaks.

You can switch between the display options by setting the default in File->Options and the "Display Options" tab, or on a per-sample basis in either the Chromosome or Whole Genome tabs by clicking on the "Display Options" icon  to bring up the Display Options window.

B-Allele Frequency

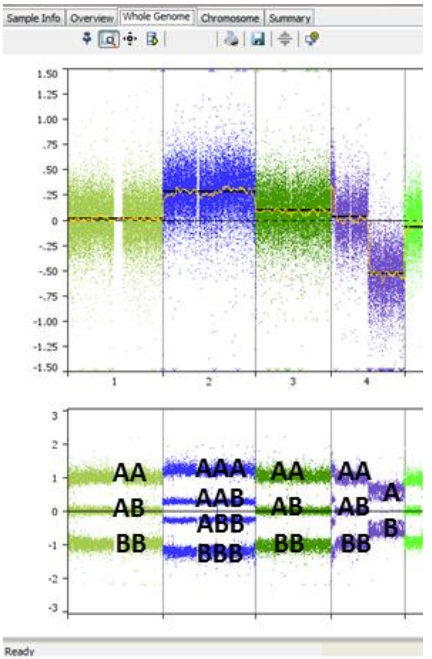
The B-Allele frequency is a measure of the normalized ratio of the "B" allele (typically the thymine or guanine allele) to the total of both alleles (i.e., the percentage of "B" alleles), given as "B/(A+B)". For example, genomic locations in the sample that are homozygous for the "B" allele are rendered as "1" (100% "B" allele), homozygous for the "A" allele rendered as "0", and heterozygous loci as 0.5.

Allele Peaks

Allele Peaks are filtered and smoothed values for individual markers. Nonparametric estimation is used to understand possible regional peak structure towards which the data is smoothed. The amount of filtration and smoothing is dynamically adapted based on sample quality. Allele difference is computed based on differencing A signal and B signal, then standardizing based on reference file information (source: Affymetrix).

For Allele Peaks, each allele (either "A" or "B") is displayed in its own track with its normalized intensity. The "A" alleles are displayed as positive values, and the "B" alleles as negative.

The below figure illustrates heterozygous, gains and deletions as represented by Alelle Peaks.

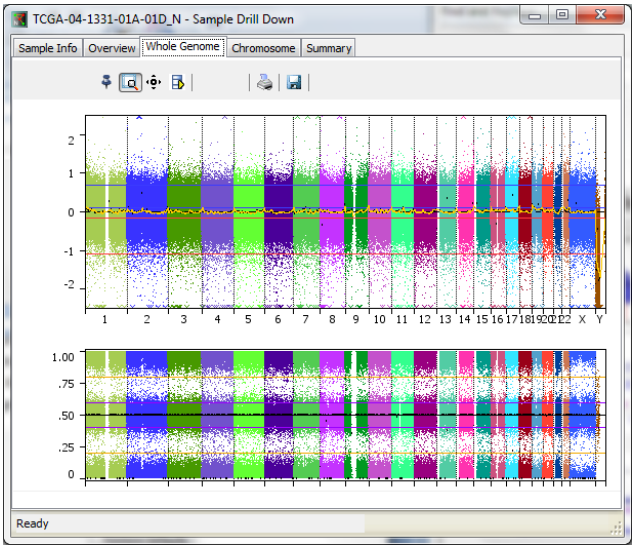


The **SNP probe plot area vs. Copy Number probe area height** setting sets the heights of the probe plots. The value specified here is the percent of vertical space that the SNP probe plot will occupy. The remaining space is reserved for the copy number plot area. E.g. a value of 50% in this field gives the same amount of space to both the SNP probes plot and the copy number probes plot. A value of 40% here gives the SNP probe plot 40% of the vertical space with the remaining 60% going to the copy number probes plot.

The **% of window for report table in Sample Drill-down Chromosome tab** specifies the amount of space the report table will occupy in the Chromosome tab whenever a new sample drill down window is opened. To show only the graphical display (top pane) when a drill down window is opened, use 0 in this field; to show only the report table (bottom pane), use 100 in this field. Once the drill down window is opened, the amount of space used by each pane can be adjusted using the horizontal handle.

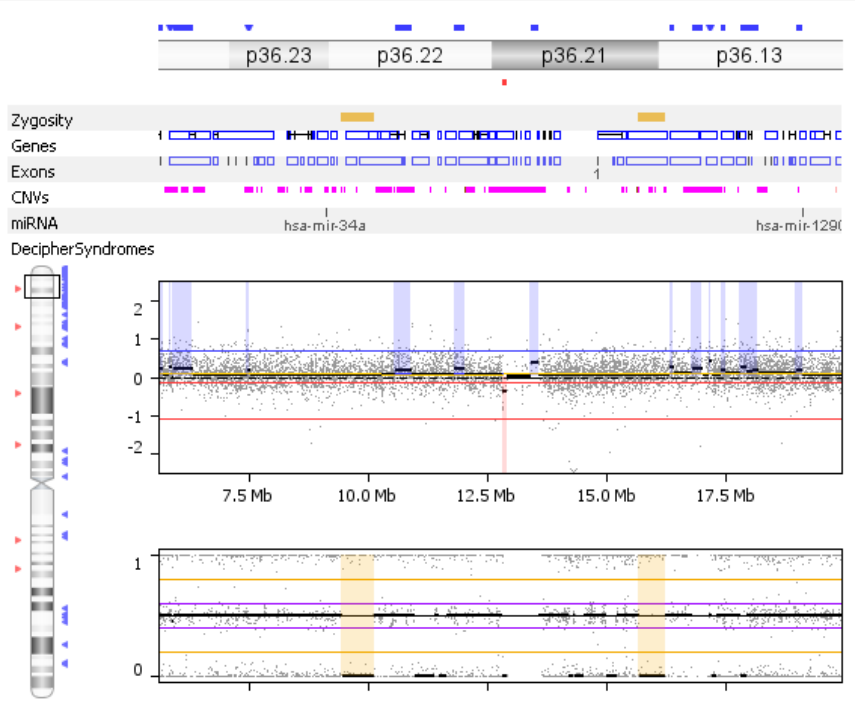
The **Drill Down Color Scheme** section allows you to specify colors used in the frequency plots in the individual sample drill down window. Click on the colored square to edit the color. Click on the **Restore Defaults** button to return the colors to the default values. Default colors are shown in the image below:

The **Genome Probes Rainbow** option allows display of the probes in the Whole Genome tab in a different color for each chromosome:



The **Segment View Baseline** allows you to change the baseline indicator from 0 to some other value. For example, if 0.5 is specified here, the baseline will be shown as a dashed line at the 0.5 value in sample drill downs.

Shade call regions in chromosome drill down will shade areas in the log ratio and B allele frequency plots that contain calls. This makes it easier to see where calls have been in the zoomed out view of the plots. The shading color is a lighter transparent version of the color used to indicate the call (light blue –gain, pink – loss, light purple – allelic imbalance, light brown – LOH).



The resizing slider option allows you to place the panning ideogram (smaller ideogram) vertically or horizontally. You can also remove this ideogram completely from view by choosing “Hide”.

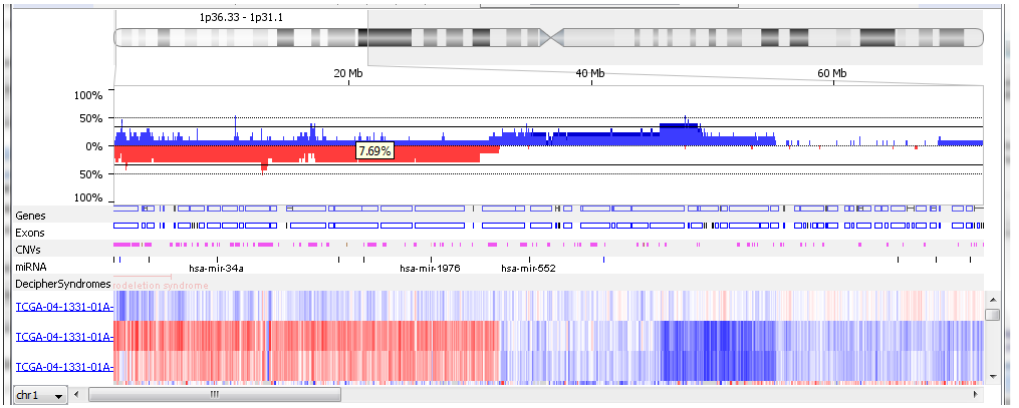
The option for numbering exons (either relative to gene or transcript) is covered in the section on *Annotation Tracks->Exons*, page 60.

The **Probe Heatmap range** controls the sensitivity of mapping log ratio intensities to the colors. For example, a range going from -0.5 to 0.5 will map any intensity below -0.5 to

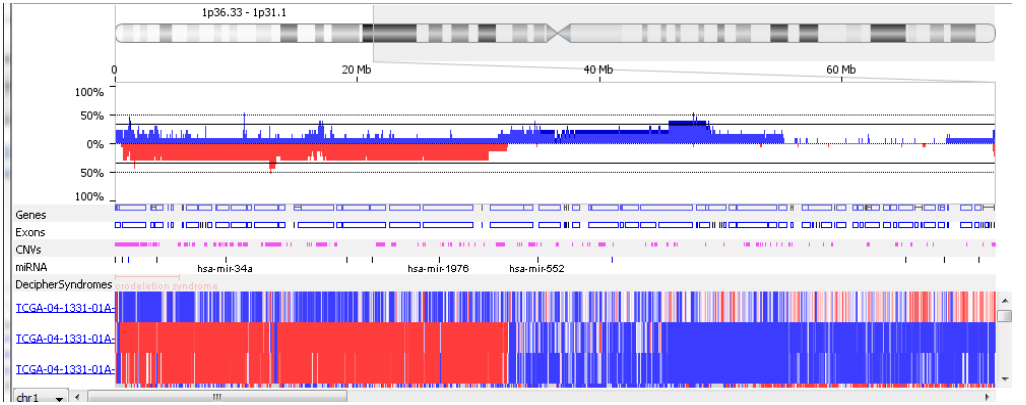
full red and any intensity above 0.5 to full blue. All values in between these two numbers will be displayed in corresponding lighter shades of red and blue. The heatmap is generated using **View->Probes** in the **Results** tab.

In the figures below, see how the larger range (-.5 to .5) has many more shades and is more sensitive to the intensity values when creating the heatmap than the smaller range (-.1 to .1).

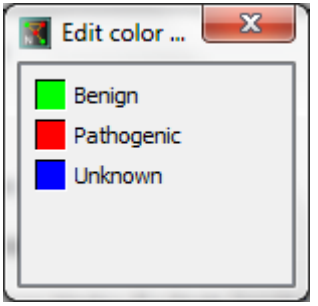
Probe heat map range -.5 to .5:



Probe heat map range -.1 to .1:



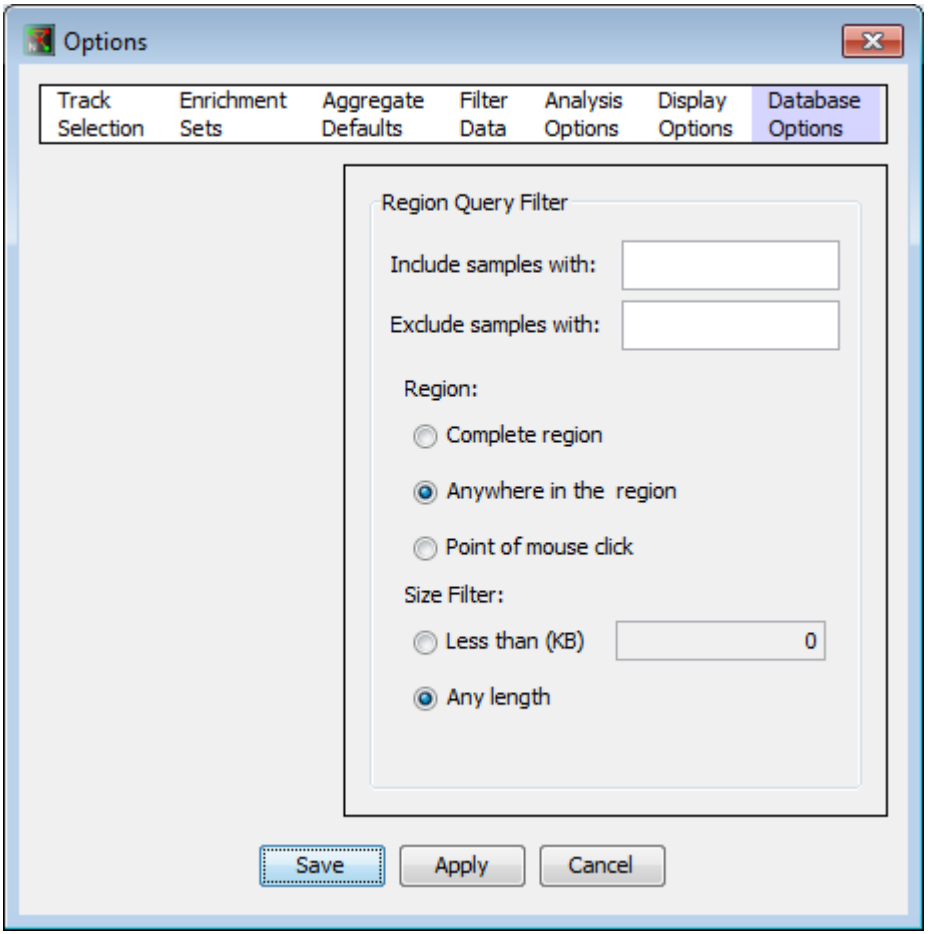
The **Call Classification Color Map** button allows editing of the colors used to represent call classifications (benign, malignant, unknown). Clicking on the button brings up the Edit colors dialog where each colored squared can be clicked to change its color.



DATABASE OPTIONS

Selecting **File->Options** menu item opens the Options window. In the Database Options tab, you can set some options for filtering queries when performing Region Queries using the **Query Database** tool. The **Include Samples With** parameter will only include in the query results, samples which have these specified keywords in their project keyword lists. The **Exclude Samples with** parameter will not display any samples which have these specified keywords in their project keyword lists. Keywords can be entered here separated with commas. The Region parameter allows specification as to the extent of

coverage of the queried region in the resulting matching regions (complete region, anywhere in region, point of mouse click) and a size filter allows for limiting the size of the returned results to a size less than what is specified.



COMMAND LINE FUNCTIONS

Some command line parameters to Nexus Express aid in automating workflows. There are two ways of running Nexus Express in batch: creating and processing a new project or processing an existing project. The GUI will not show up when running Nexus Express in batch mode.

CREATING AND PROCESSING A NEW PROJECT

To create and process a new project use the following (all in one line) in a .bat file, substituting the appropriate paths to the location of your new project and to the location of your descriptor file:

```
jre\bin\java -Xms400m -Xmx1200m -jar Nexus.jar
-newproject="C:\wherever\newProject"
-samples="C:\wherever\data\descriptor.txt"
-organism="Human" -build="NCBI Build 36.1" -batch
```

The descriptor.txt is a sample descriptor file that you need to create (see the *Sample Descriptor* section, page 258).

PROCESSING AN EXISTING PROJECT

An existing Nexus Express project can be opened and processed (with data that has already been loaded into the project) with the following command (all in one line) in a .bat file or by changing to the Nexus Express install folder in the command prompt console and running this:

```
jre\bin\java -Xms400m -Xmx1200m -jar Nexus.jar
-openproject="C:\location\of\the\specific\project" -batch
```

The **-openproject** parameter with **-batch** opens an existing project, processes it (if unprocessed sample are present) and closes down.

COMMAND LINE OPTIONS

Note: Make sure the "-" symbol in the above commands is the hyphen since copy and paste or autocorrect sometimes changes - to – (hyphen to an en dash) in some word processing applications. These two symbols (- and –) are different. Only the first (shorter one) is recognized by Nexus Express.

Entering this on the command line in the Nexus Express install folder will show and explain all the options available in batch:

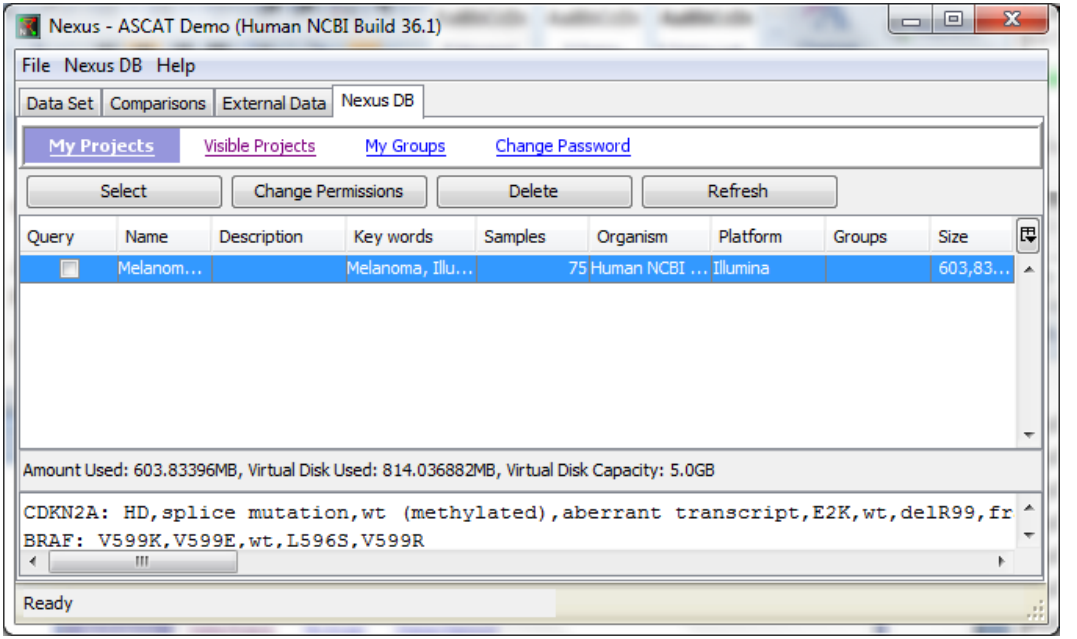
jre\bin\java -jar Nexus.jar –help

Allowed options:

Option	Description
-batch	Processes all samples and then shuts off
-build	Specifies the build version
-exporttable	Exports the calls table to a file after processing and then shuts off
-newproject	Creates a new project at the location specified
-nolog	Sends output messages to stdout instead of the log file
-openproject	Opens an existing project at the location specified
-organism	Specifies the type of organism
-samples	Specifies the sample descriptor file to load

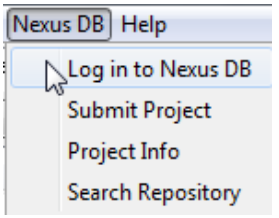
NEXUS DB

Nexus DB is a secure off-site repository for redundant back-up of Nexus projects and a collection of already processed public projects that users can download to add to their data sets. The site uses the most current internet security protocols to limit access to the data and only to the data owner. With data all in one place, samples and projects can be queried on keywords, project properties, or on common aberrations and samples or projects can be shared quickly with colleagues. Once a user is logged into the Nexus DB, an additional tab is displayed in the main Nexus window. Please note that Download features are not enabled in Nexus Express.

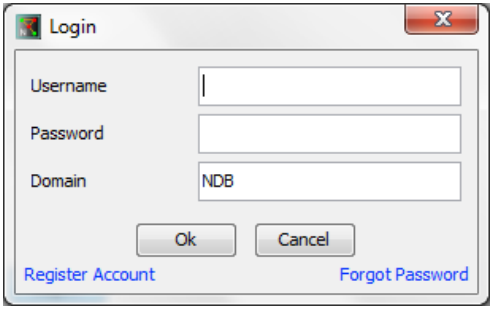


INTERFACE AND REPOSITORY ACCESS

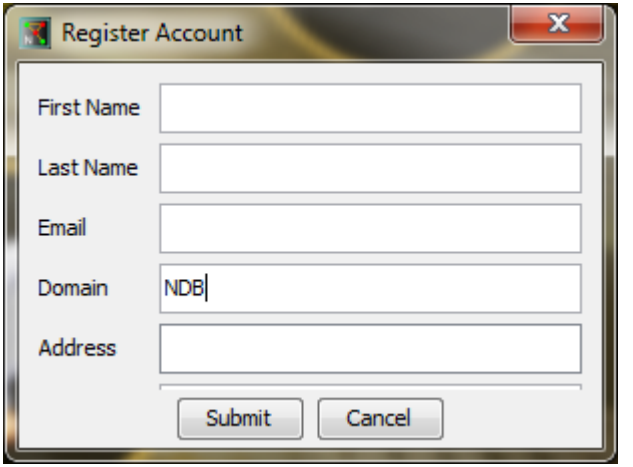
For Nexus Express licenses that have Nexus DB accessibility, the startup screen will display an icon to log in to the Nexus DB. One can log into Nexus DB by clicking on the icon or from the Nexus DB menu (see figures below).



The login dialog requires a username, password, and domain.



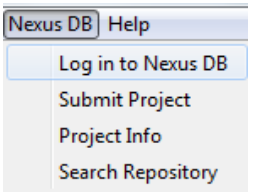
If you don't already have a Nexus DB account, click on the "Register Account" link and a dialog will appear. Once submitted, your information is verified by BioDiscovery and an email is sent to the license holder of your Nexus Express software for account approval. Once approval is received, BioDiscovery will create your Nexus DB account and assign you to the main user group for your license. You will receive an email containing your username, password, and group name. After your first log-in, you should change your password.



Any number of Nexus DB accounts can be associated with a single Nexus Express license so each member of a lab, for example, can and should request his own Nexus DB

account. He will be added to the lab's group and can then elect to share one or more of his projects with the group.

If you have already opened a project in Nexus Express and then would like to login in to the Nexus DB, you can do so from the **Nexus DB** menu.



GROUPS AND PERMISSIONS

There are many different Groups in Nexus. Some are solely for individuals using the same Nexus license and others are for anyone to join to access specific projects in the repository. For example, BioDiscovery has loaded a number of samples from The Cancer Genome Atlas (TCGA) project and users with a Nexus DB account can access this data by emailing NexusDBAdmin@biodiscovery.com and requesting to be added to the TCGA group. If you would like to create a new group for your consortium, department, specific disease area, etc., just request it by emailing NexusDBAdmin@biodiscovery.com.

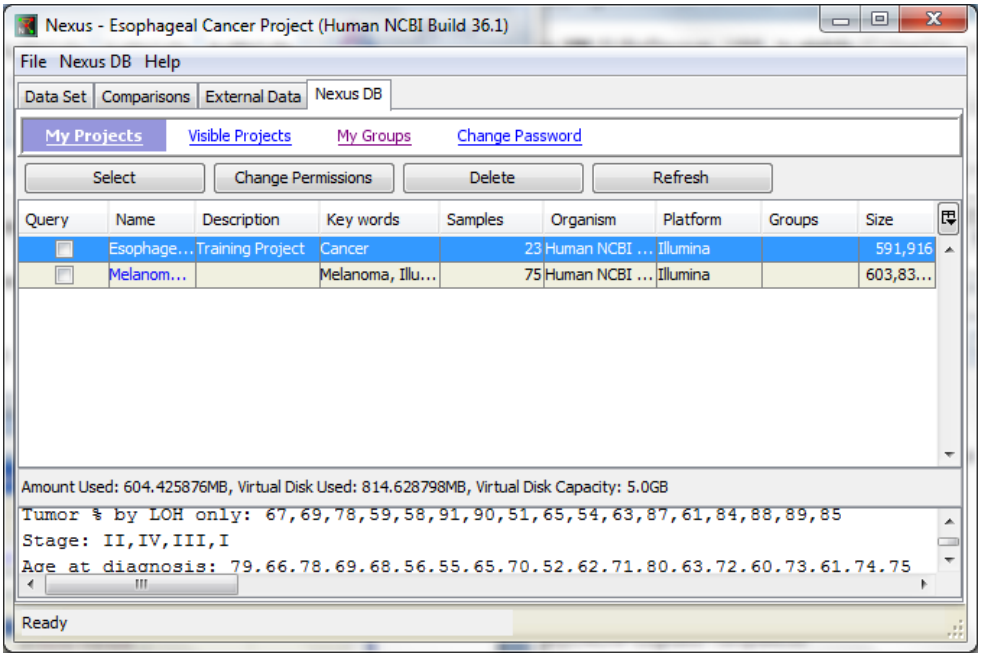
Clicking on **My Projects** lists all projects belonging to the user that are stored on the repository. Below this is a panel listing the amount of storage space available.

Amount used – lists amount of space used by the logged in user

Virtual disk used – lists the amount of space used by everyone whose space resides on this disk (a virtual disk corresponds to a product serial number and is shared by those sharing that particular license of Nexus)

Virtual Disk Capacity – lists the capacity of this disk. If this amount is not sufficient, additional space can be obtained at minimal cost by inquiring with BioDiscovery (customerservice@biodiscovery.com).

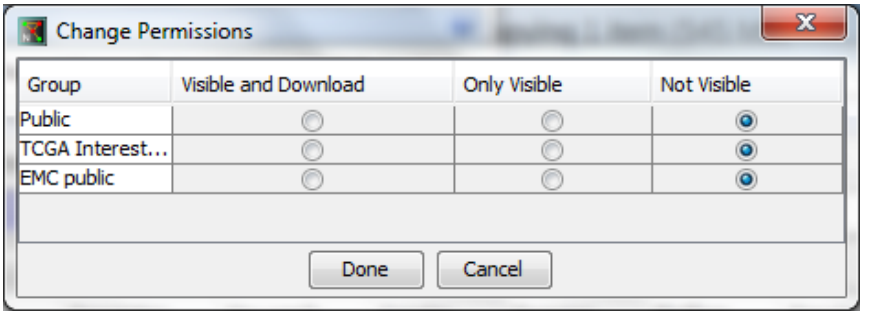
When a project is highlighted, the bottom panel displays all factors and factor values for this project.



PERMISSIONS

Clicking on the **Change Permissions** button will allow you to set project access. Highlight a project and then click on **Change Permissions** to modify access for that Group.

In the **Change Permissions** window, you will find the group names and whether that group has access to your projects. Click on the radio buttons to adjust who can access you project. Access to your data is on a per project basis and each project's accessibility must be set individually by highlighting the project in the **My Projects** list and then clicking **Permissions**.



Everyone with a Nexus DB account has access to the **Public** group. If you choose to make your project available to this group, your project is moved into this space and then frees up your own repository space.

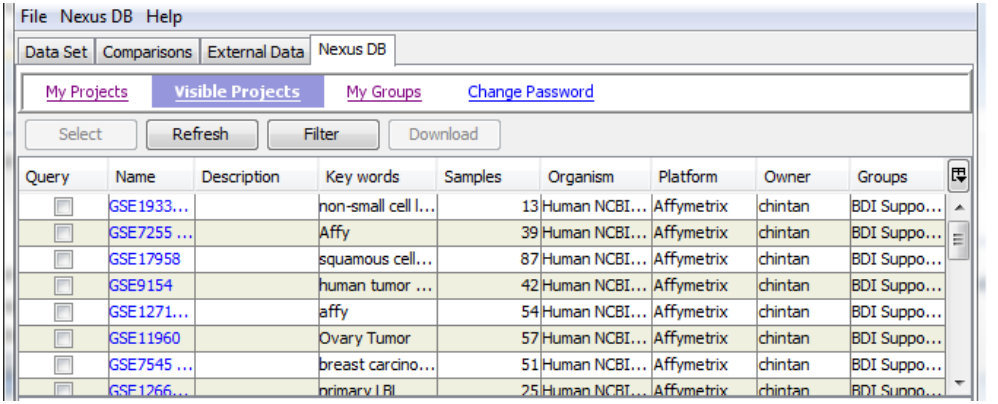
Visible and Download - Allows users to view and download the samples/projects

Only Visible - Samples will be visible in queries but users will not be able to download the samples

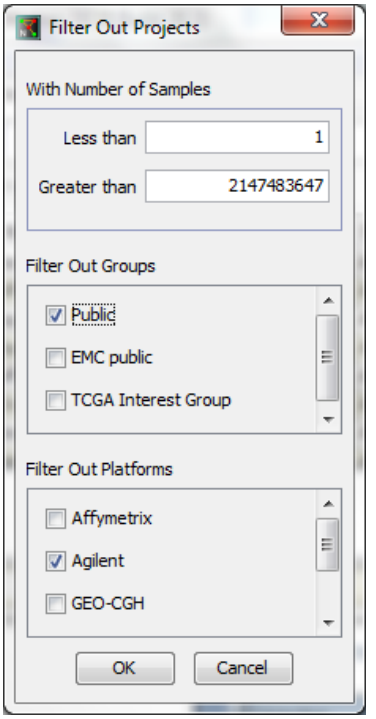
Not Visible – Samples are not visible/available to anyone

VISIBLE PROJECTS

The **Visible Projects** page in the Nexus DB tab lists all projects to which you have access. This includes projects visible to your groups as well as public projects.



Clicking on the **Filter** button brings up a window where you can select to remove from the list of projects those meeting certain criteria.



You can limit by the number of samples in a project, removing projects that are accessible by certain groups, and not displaying projects with samples from specific array platforms.

GROUPS

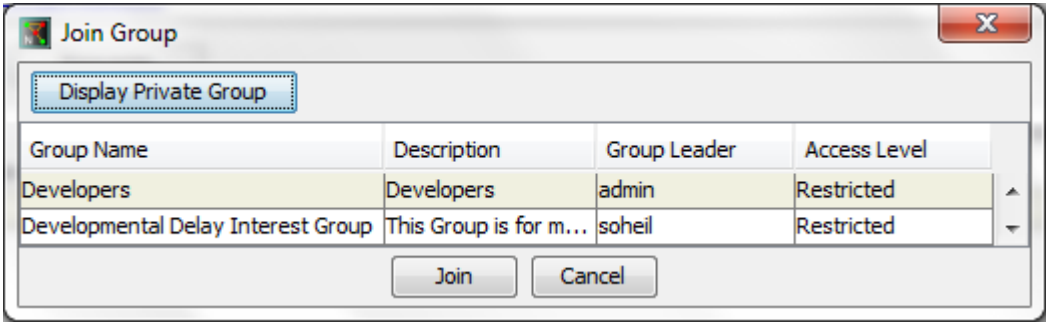
Clicking on **My Groups** will list names of all the groups to which you belong. Any user can create a new group and apply an access level:

- anyone can join
- restricted access - group owner will have to approve a user to join

- private - group is not visible to users and a user can join this only by invitation from the group owner

Joining a Group

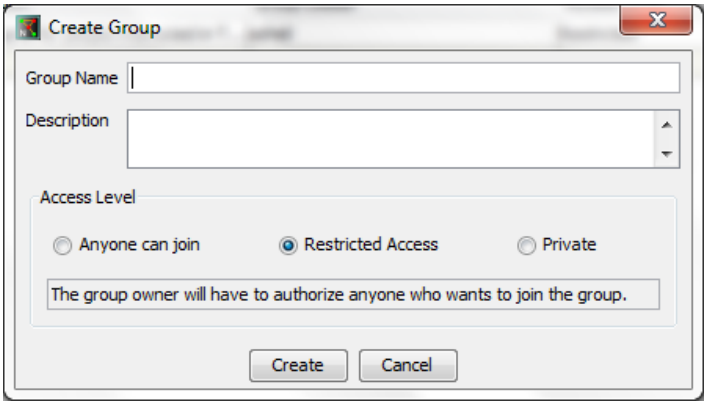
Any user can request to join a group by clicking the **Join Group** button and selecting the group from a list. If the access level is "Restricted" then the group owner must approve before the user will be added to the group. Private groups can be displayed only if a user enters an access key (provided by the group owner). If you click the **Display Private Group** button, you will be prompted to enter an access key.



When the owner approval is needed to join a group, the Group owner receives an email with the full name and Nexus DB username of the person requesting to join a certain group. If the owner accepts the request, the Group will now appear in the requestor's **My Groups** page.

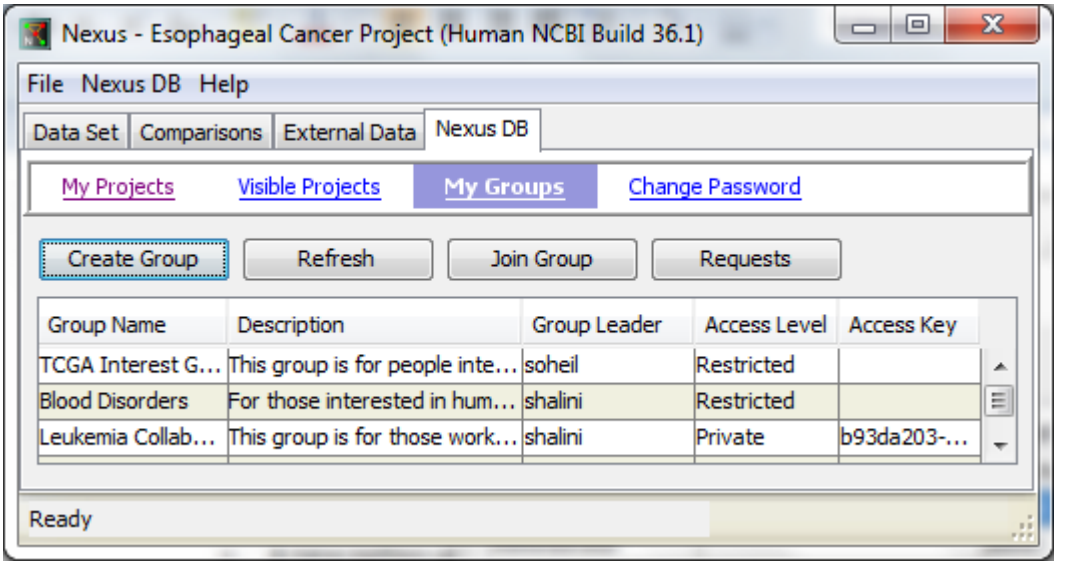
Creating a Group

Any user can create a group by clicking on the **Create Group** button.



One of three access levels can be chosen for each group. The information panel just below the radio buttons provides information on what that access level means.

Once a group is created, it is then placed on the **My Groups** page. If a group was set to Private, then an access key is provided in the **Access Key** column (see Leukemia Collaboration group in the figure below). The group owner then needs to provide this access key to anyone whom they invite to join this group. When the recipient then tries to join this group, he will have to enter the Access Key.



Clicking on the **Requests** button displays a list of those who have requested to join your group (name and institution will be listed). Here you can accept or deny the request.

QUERYING THE REPOSITORY

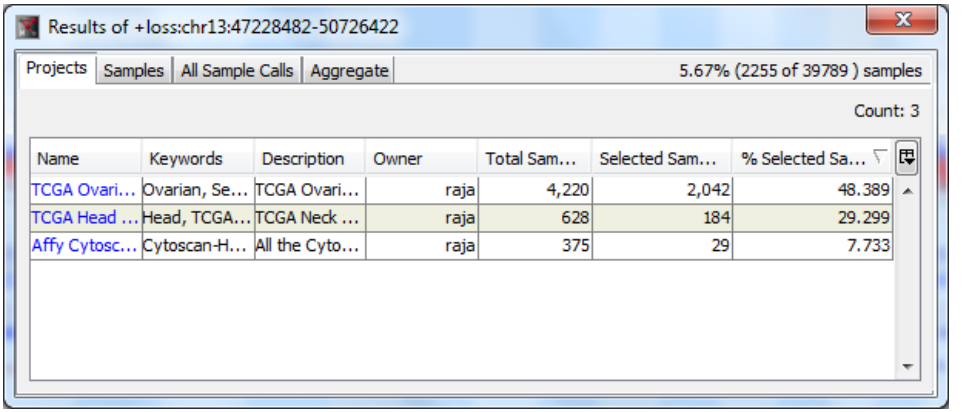
The repository can be searched by using the "Query Database" tool  in the **Genome** and **Chromosome** pages of the **Results** tab, **Chromosome** tab of the individual sample drill down window or by using the **Search Repository** button in the **Data Set** tab. The Query Database tool performs region queries and the **Search Repository** button allows keyword searches. Searches are filtered on the current open project's organism and build as you must have a project open in order to download data. You cannot add samples mapped to different genomes/builds into a single project. Thus, if your open project maps to human build 36.1, then your search results will show only those samples that map to this genome.

REGION SEARCHES WITH THE QUERY DATABASE TOOL

This search using the Query Database tool  allows identification of projects which share aberrations at a specific chromosomal region. To search for gain or loss events in other projects/samples on the repository, first click the **Query Database** tool to select/activate it (the tool will now be surrounded by a rounded rectangle and have a white background):

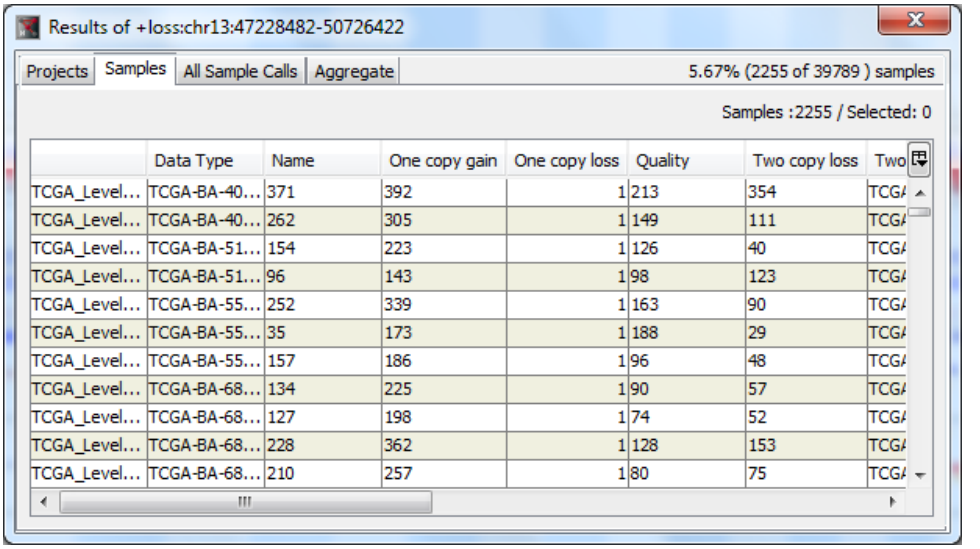
Then click on a gain or loss at a specific location by clicking on a single spot above the 0 line (for gains) or below the 0 line (for losses) on the frequency plot. Clicking on a loss on chr13 in the current local project, brings the following results from the repository where each project (to which the user has access) in the repository having that aberration is shown along with what percentage of the samples in each project have that aberration.

Also on the top right, the percent of all samples in the repository with the aberration is shown. This number includes all samples to which the user has access, including public data.



Name	Keywords	Description	Owner	Total Sam...	Selected Sam...	% Selected Sa...
TCGA Ovari...	Ovarian, Se...	TCGA Ovari...	raja	4,220	2,042	48.389
TCGA Head ...	Head, TCGA...	TCGA Neck ...	raja	628	184	29.299
Affy Cytosc...	Cytoscan-H...	All the Cyto...	raja	375	29	7.733

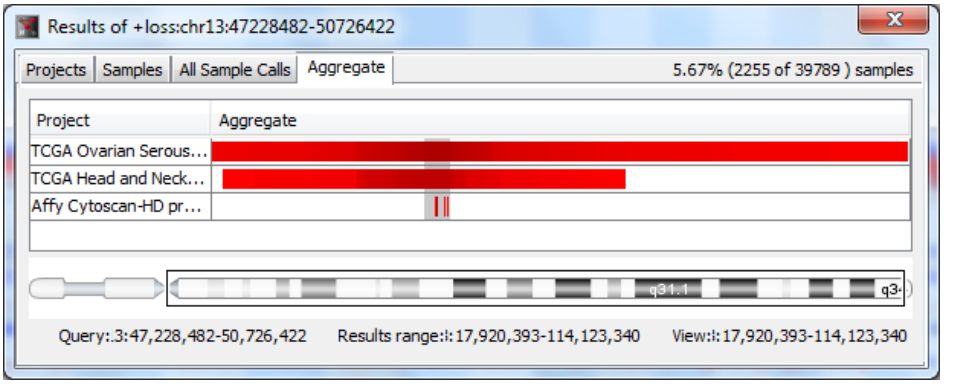
The samples tab list all samples that contain the aberration.



	Data Type	Name	One copy gain	One copy loss	Quality	Two copy loss	Two
TCGA_Level...	TCGA-BA-40...	371	392	1	213	354	TCGA
TCGA_Level...	TCGA-BA-40...	262	305	1	149	111	TCGA
TCGA_Level...	TCGA-BA-51...	154	223	1	126	40	TCGA
TCGA_Level...	TCGA-BA-51...	96	143	1	98	123	TCGA
TCGA_Level...	TCGA-BA-55...	252	339	1	163	90	TCGA
TCGA_Level...	TCGA-BA-55...	35	173	1	188	29	TCGA
TCGA_Level...	TCGA-BA-55...	157	186	1	96	48	TCGA
TCGA_Level...	TCGA-BA-68...	134	225	1	90	57	TCGA
TCGA_Level...	TCGA-BA-68...	127	198	1	74	52	TCGA
TCGA_Level...	TCGA-BA-68...	228	362	1	128	153	TCGA
TCGA_Level...	TCGA-BA-68...	210	257	1	80	75	TCGA

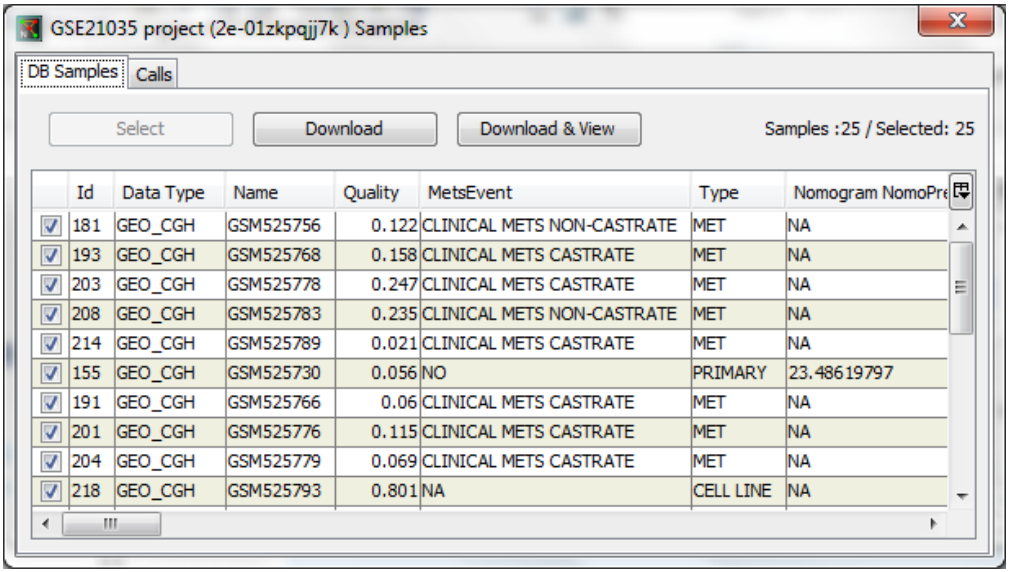
The **Aggregate** tab displays the results as a heat map. The darker the region, the greater the percentage of samples having that aberration at that point in the project. An

ideogram at the bottom shows which portion of the chromosome is depicted in the heat map. Below this, the query is listed along with the chromosomal location of the results and the location of current viewing field.



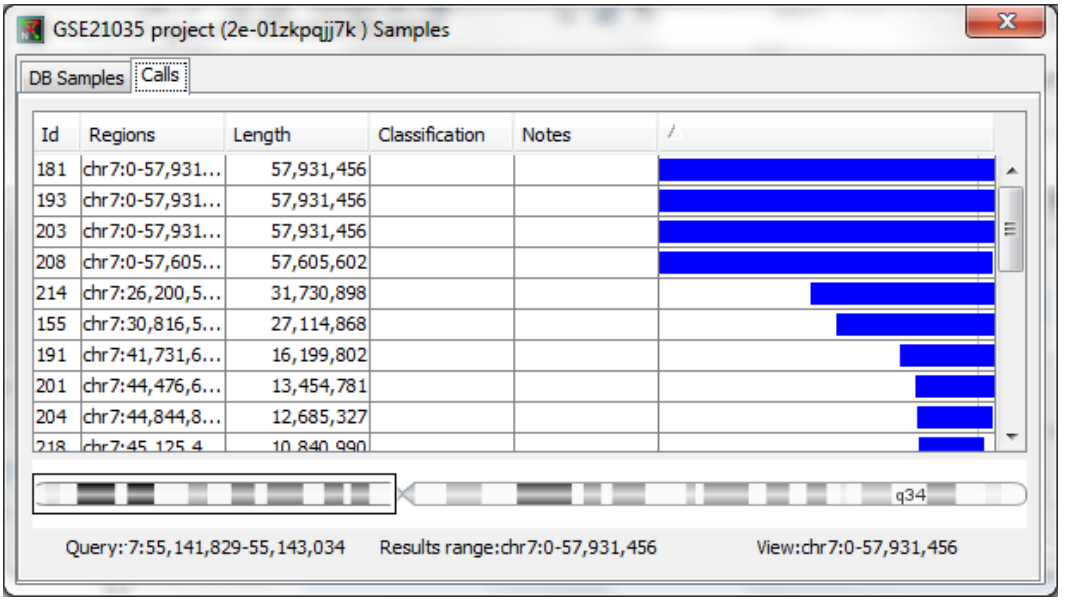
Clicking on the project name in the Projects tab will open up a new window with two tabs containing additional details on samples in a project.

The **DB Samples** tab lists the sample names, factors and quality scores.



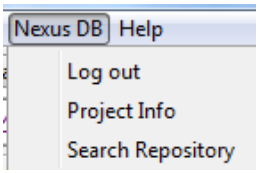
Id	Data Type	Name	Quality	MetsEvent	Type	Nomogram NomoPre
181	GEO_CGH	GSM525756	0.122	CLINICAL METS NON-CASTRATE	MET	NA
193	GEO_CGH	GSM525768	0.158	CLINICAL METS CASTRATE	MET	NA
203	GEO_CGH	GSM525778	0.247	CLINICAL METS CASTRATE	MET	NA
208	GEO_CGH	GSM525783	0.235	CLINICAL METS NON-CASTRATE	MET	NA
214	GEO_CGH	GSM525789	0.021	CLINICAL METS CASTRATE	MET	NA
155	GEO_CGH	GSM525730	0.056	NO	PRIMARY	23.48619797
191	GEO_CGH	GSM525766	0.06	CLINICAL METS CASTRATE	MET	NA
201	GEO_CGH	GSM525776	0.115	CLINICAL METS CASTRATE	MET	NA
204	GEO_CGH	GSM525779	0.069	CLINICAL METS CASTRATE	MET	NA
218	GEO_CGH	GSM525793	0.801	NA	CELL LINE	NA

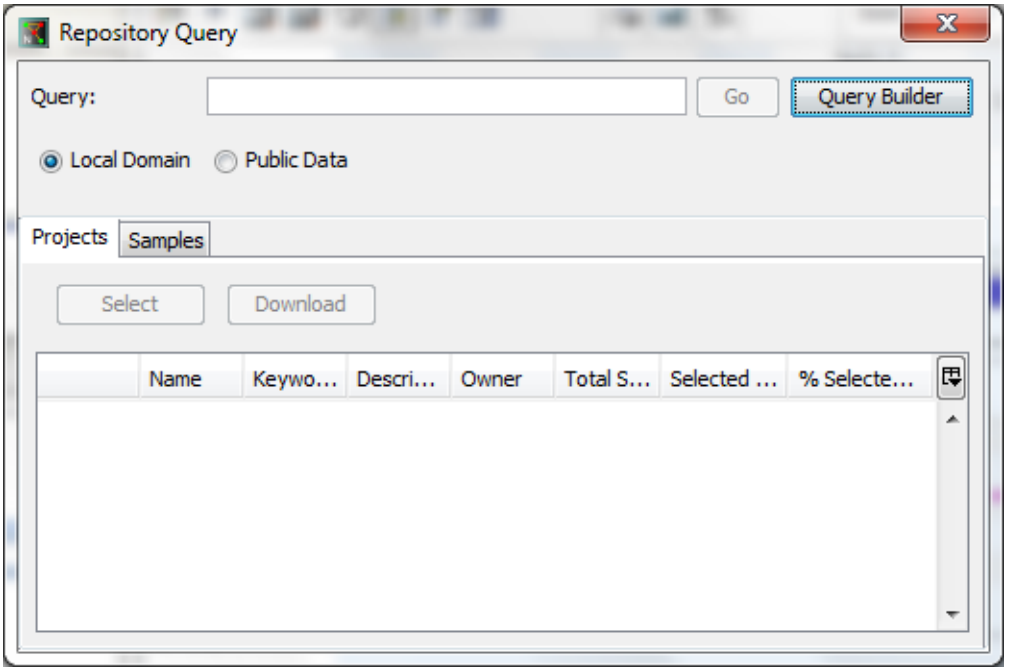
The **Calls** tab lists the regions, call classification, notes, and a graphical view of the aberration.



KEYWORD SEARCHES

Keyword searches for project properties (e.g. array platform) or sample factors (e.g. gender, tumor type, etc.) is achieved with the **Search Repository** item under the **Nexus DB** menu in the **Data Set** tab.

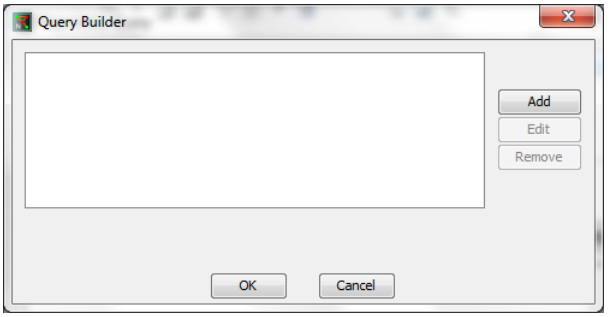




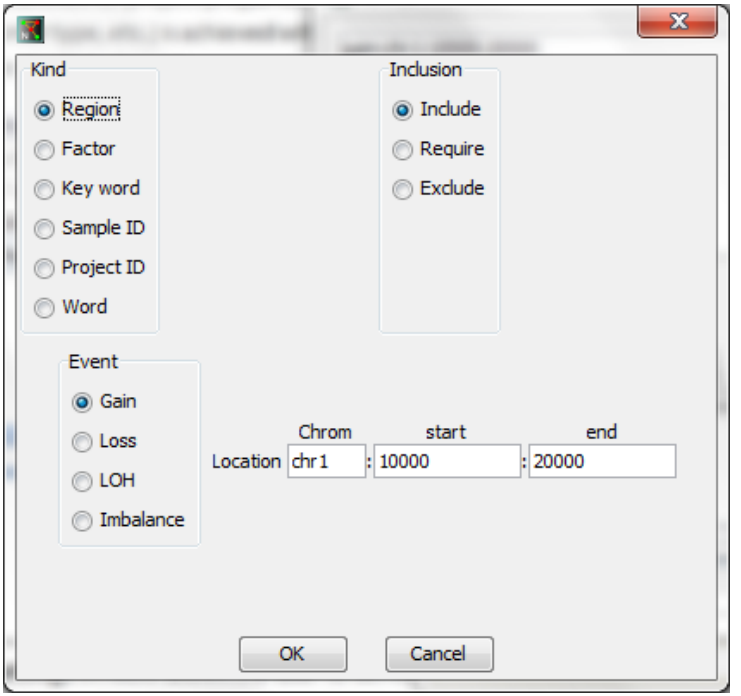
You can select to search only the projects to which you have access on your domain by selecting **Local Domain** or you can choose to search only public projects by selecting **Public Domain**. You can enter your search terms in the **Query** text box or use the **Query Builder** tool (via the **Query Builder** button) to help generate complex queries.

Query Builder Tool

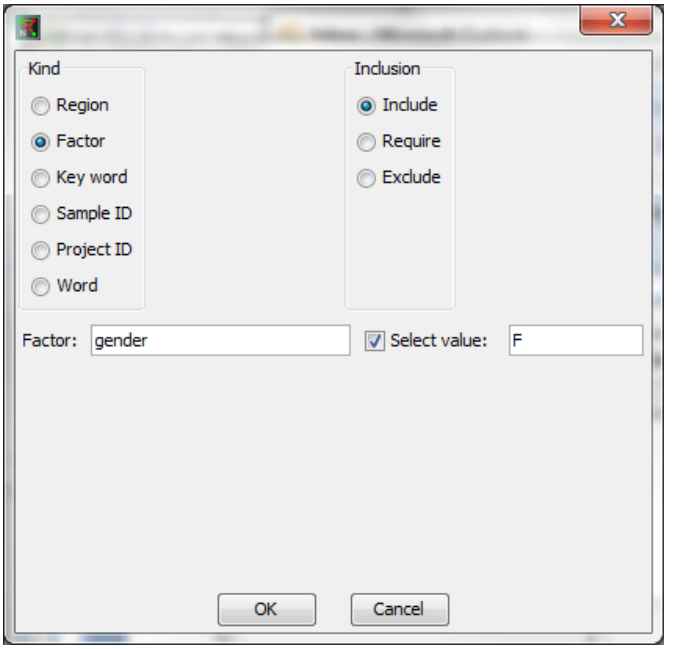
The query builder helps you to generate complex queries without having to use the notation described in the section below on manually creating queries. Clicking on the Query Builder button brings up a window where the query is generated.



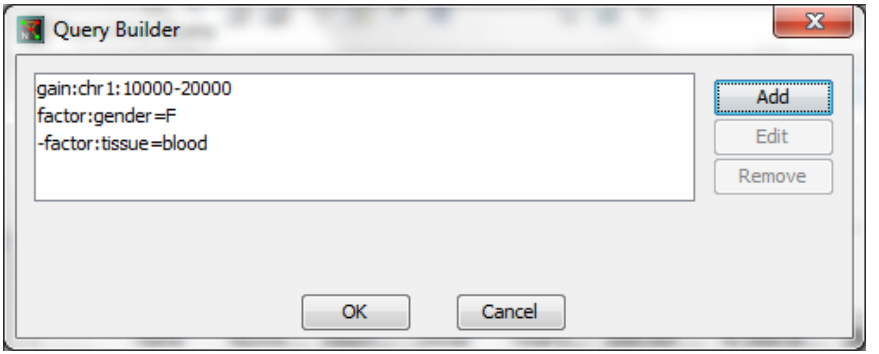
Click on **Add** to add a keyword. This brings up the window below where you select what you would like to query (a region, keyword, factor, etc.) and depending on the selection here, additional fields will appear. In the case of "region", the type of event (gain, loss, LOH, imbalance), and chromosomal location can be specified (see figure below).



If **Factor** is selected for **Kind**, then in addition to the factor name the factor value can also be specified:

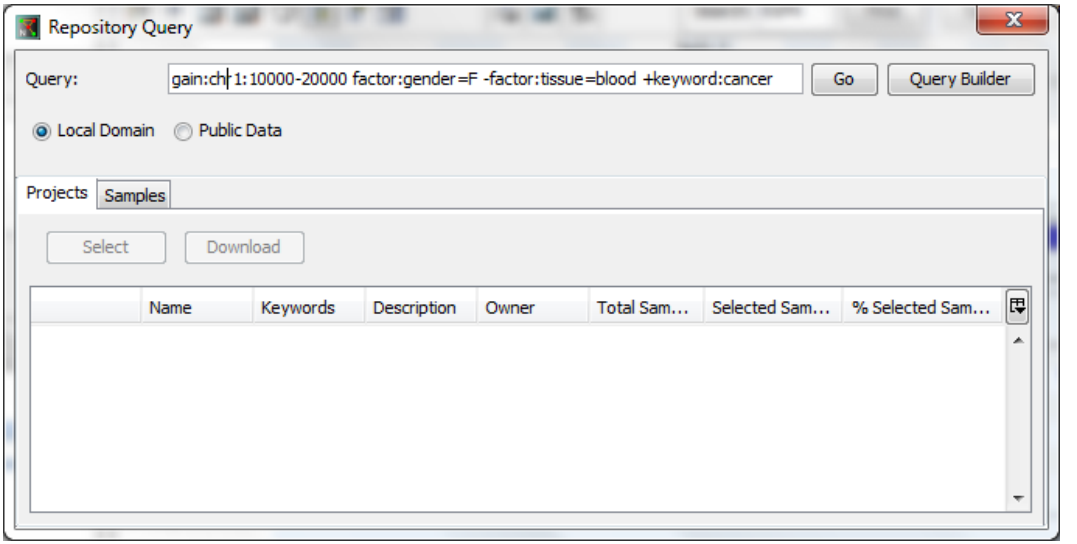


As different criteria are added, they are added to the query builder window:



When all criteria have been added, clicking on **OK** brings forward the Repository Query window where the query just created is entered into the Query field. Clicking on **Go**

executes the query and results meeting the criteria are shown in the **Projects** and **Samples** tabs.



MANUAL QUERY CREATION

Terms can be entered directly into the text box rather than using the query builder to create queries. When using multiple word terms, make sure to surround them with quotes. Searches are case-insensitive. The repository provides intelligent searching in that it has an internal synonym list such that if a Factor called “Gender” is searched for, the results will include Factor “Sex” since “Sex” is another way to specify “Gender”. If Factor:Gender=male is searched, the results will include all samples that have values “Male”, “M”, “male”, and “m”. After entering your criteria, press the **Go** button to execute the search.

To search for Factors and/or factor values, use the following format:

Factor:Gender

Factor:Gender=male

Factor:Gender finds all projects and samples which contain the Factor Gender. Factor:Gender finds all projects and samples indicating male as the gender. To add additional criteria add a space and "+" then specify the next criteria. E.g.

Factor:Gender=male +Factor:Survival

To search for keywords, just enter the words:

cancer +lung +"stem cell"

You can combine Factor and keyword criteria:

Factor:Gender=male +Factor:Survival +cancer

You can also use a minus sign to filter out results. E.g. If you want to search for the Factor Survival in all projects and samples except those associated with cancer then use the following notation:

Factor:Survival -cancer

The Results are displayed in two tabs – one tab lists all the Projects and the next tab lists samples.

SAMPLE DESCRIPTOR

*Note: A **Sample Descriptor** may be used to load data for special cases such as performing matched paired analysis. If you don't have such data, you do not need to create a sample descriptor and can skip this section.*

BASICS

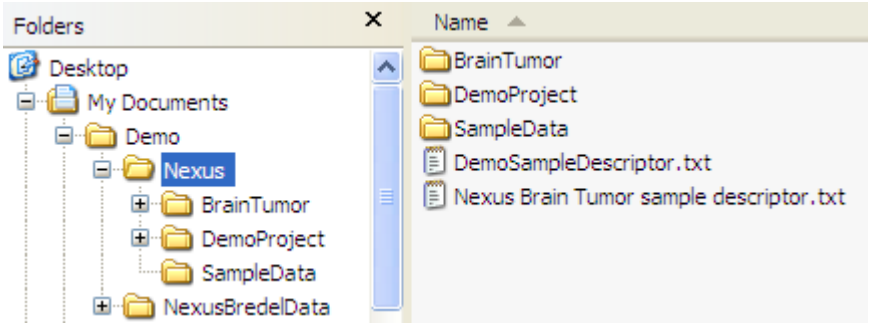
A **Sample Descriptor** is a tab-delimited text file that describes the type of data being loaded, the location of the data files, and any biological factors associated with the samples. If data from multiple data types are being loaded into one Nexus Project, a different Sample Descriptor file is created for each data type. The Sample Descriptor file for each type of data follows a similar format.

The first line of the Sample Descriptor file specifies the **Data Type** (e.g. Affymetrix OncoScan CBZIP or OncoScan OSCHP) that is going to be loaded into the Nexus Project. These files contain raw intensity values, log ratios, or copy number calls depending on the array platform and associated software. The first line needs to be **DataType:** followed by a tab and then the name of the data type. The next line is a header which describes the type of information needed for each sample in the data set. The remainder of the lines (one row for each sample) specifies the sample name, location of the data files, and factors.

Columns specifying the factor values are optional in the sample descriptor and there is no limit to the number of factors that can be added. These are specified by using **Factor:** followed by a name for the factor. The factor name is not limited to a single word and can be composed of a multiple word string. For example, **Age** and **Brain Tumor Classification** are both valid factor names and would be specified as **Factor:Age** and **Factor:Brain Tumor Classification**, respectively.

Any column in the Sample Descriptor file that is specifying an input data file location can contain either the full path or the relative path (relative to the location of the Sample

Descriptor file). In the example below, **My Documents->Nexus** contains a folder called **SampleData**, containing the sample files, and the Sample Descriptor file, **DemoSampleDescriptor.txt**.



Column headers that can be used in Sample Descriptor files:

Column	Description
Factor:	Optional. Any factor associated with the sample. Specified w/ **Factor:** followed by a term specifying the Factor. The term can be any text string and is not limited to a single word.
File	Specifies location of the input data file (containing raw intensities, log2 ratios. or copy number calls) to Nexus. *Full path name or path relative to the location of the Sample Descriptor file*.
Sample Name	Name of the sample

INPUT DATA FILES

Each input data file to Nexus Express Nexus needs to be in a specific format so that Nexus Express is able to load the data. The format of these input files as well as how to

generate these is detailed in *Appendix A: Data Type Format and Sample Descriptor Templates*, page 261.

TEMPLATE SAMPLE DESCRIPTOR FILES

For most data types, the sample descriptor file needs to be created manually by the user. A template Sample Descriptor file for each data type is located in the **Templates** folder in the Nexus installation directory. Within the **Templates** folder, each data type has its own folder (e.g. **Affymetrix**, etc). It is easiest to view and edit these files using a spreadsheet application such as Microsoft Excel. These templates contain only the required columns. If you have any biological factors that you want associated with each individual sample, please add additional columns (one for each factor) that begin with **Factor:** followed by the factor name. Once the template file has been populated, make sure to save this as a text file (with a **.txt** extension). See *Appendix A: Data Type Format and Sample Descriptor Templates*, page 261, for more information on Sample Descriptor files for each Data Type.

APPENDIX A: DATA TYPE FORMAT AND SAMPLE DESCRIPTOR TEMPLATES

In Nexus Express, the array type is limited to the OncoScan platform arrays.

Below are tables (one for each common data type supported by Nexus) detailing what is needed in the input file to Nexus as well as what needs to be in the Sample Descriptor file for a particular data type. The tables also indicate the appropriate Sample Descriptor template file to use (depending on what the input data is) to load data into Nexus. *Please note that a sample descriptor is not required to load data in most cases. It is only required in rare cases such as matched paired analysis.* Below the sample descriptor template file name are the headers used in the template file. The template file is tab delimited though in the tables below the headers, which are separated by commas. Please note that column headers in square brackets [ ] are optional columns. All other columns are required.

Affymetrix OncoScan CBZIP

Required format for input file to Nexus:

.CBZIP file provided by Affymetrix

Required columns in input file: NA

Sample Descriptor Template	Description
\Templates\Affymetrix\ AffyCBZIP-Template.txt Data Type: Affymetrix OncoScan CBZIP Sample Name, File, [Factor:]	Affymetrix OncoScan CBZIP files

Affymetrix OSCHP - TuScan

Required format for input file to Nexus:

.oschp file provided by Affymetrix

Required columns in input file: NA

Sample Descriptor Template	Description

\Templates\Affymetrix\ AffyOSCHP-TuScan-Template.txt Data Type: Affymetrix OSCHP-TuScan Sample Name, File, [Factor:]	Affymetrix OncoScan .oschp files

Affymetrix OSCHP-SNP-FASST2

Required format for input file to Nexus:

.oschp file provided by Affymetrix

Required columns in input file: NA

Sample Descriptor Template	**Description**
\Templates\Affymetrix\ AffyOSCHP-SNP-FASST2-Template.txt Data Type: Affymetrix OSCHP-SNP-FASST2 Sample Name, File, [Factor:]	Affymetrix OncoScan .oschp files

APPENDIX B: ADDING CUSTOM ANNOTATION TRACKS

Various types of custom annotation tracks can be added to a project. Some common examples are

-User-defined probe mappings

-Custom CNV tracks

-BED files from UCSC Genome Browser

See below for quick summaries on how to add such tracks and then read further below for more details on the required file formats. For further help in adding custom annotation tracks please contact BioDiscovery Support (support@biodiscovery.com).

QUICK SUMMARY

The quick summary outlines the steps required to create such files. If you need additional information, continue reading the other sections below.

CUSTOM CNV TRACKS

Steps need to create a custom cnv track:

1. Create a tab-delimited text file containing the cnv information. The structure of this file follows that of the **cnvs.txt** file in the organism build subfolder.
2. Place this file in the **customcnvs** folder within the appropriate Organism sub folder.
3. Restart Nexus and go to **File->Options->Track Selection.** The new probe mapping should appear in the **Select CNV Track** panel. Select the cnv track and click **Apply** to display this data in the viewer.

CUSTOM TRACKS WITH BEDFILES FROM THE UCSC GENOME BROWSER

Steps need to create a track using BED files:

1. Create a BED file following the format defined at http://genome.ucsc.edu/goldenPath/help/customTrack.html#BED

2. Place the BED file in the **customtracks** folder in the respective Organism subfolder.

4. Restart Nexus Express and go to **File->Options->Track Selection.** The new probe mapping should appear in the **Select Additional Tracks** panel. Select the BED file and click **Apply** to display this data in the viewer.

ADDING A CUSTOM CNV TRACK

If you want to display your own CNV data, you can do so by creating a new cnv file and placing it in the **customcnvs** folder in the organism build subfolder. The structure of this file follows that of the **cnvs.txt** file in the organism build subfolder. The following columns are required in a custom cnv file.

Column	Description
Chr	Chromosome number
Start	Start position (bp) of the cnv
End	End position (bp) of the cnv
VariationType	Values here can be **CopyNumber**, **InversionBreakpoint**, **Inversion**, and **Indel**
Reference	Publication reference
Method/platform	Technology used to determine the CNV
Gain	Number of gains found

Column	Description
Loss	Number of losses found
TotalGainLossInv	Total number of aberrations found
PubMedID	PubMedID for the reference
VariationID	Variation ID from the DGV database

The different types of CNVs will be displayed in a different color on the track:

CopyNumber – magenta

InversionBreakpoint – cyan

Inversion – pink

Indel - brown

Once you have created your file and saved it as a tab-delimited text file in the **customcnvs** folder, you must restart Nexus in order for your new cnv file to appear in the **Track Selection** options (**File->Options->Track Selection**). You can now select your file here and press **Apply**. The data from this file will now be displayed in the CNVs track.

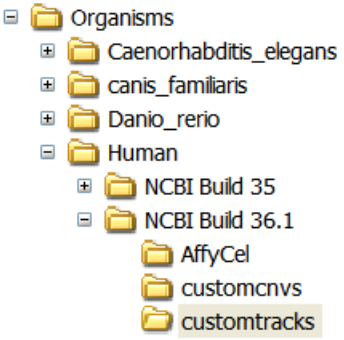
ADDING BED FILES FROM UCSC GENOME BROWSER

Data displayed in a UCSC Genome Browser annotation track can be imported into Nexus using the .BED format defined by UCSC Genome Browser. See the BED file format at the UCSC Genome Browser for more information (http://genome.ucsc.edu/goldenPath/help/customTrack.html#BED).

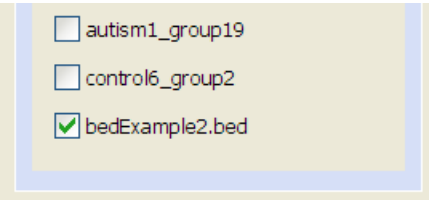
An example BED file:

	A	B	C	D	E	F	G	H	I	J	K
1	chr7	54028	73584	uc003sii.2	0	-	54028	54028	255,0,0	.	AL137655
2	chr7	60328	61569	uc010krx.1	0	-	60328	60328	255,0,0	.	PDGFA
3	chr7	62967	63529	uc003sij.2	0	-	62967	63366	255,0,0	.	DQ576410
4	chr7	64068	64107	uc003sil.1	0	-	64068	64068	255,0,0	.	DQ584609
5	chr7	65159	65220	uc003sim.	0	-	65159	65159	255,0,0	.	DQ600587
6	chr7	75460	116489	uc003sin.1	0	-	75460	75460	255,0,0	.	AL137655
7	chr7	244679	249951	uc003sio.1	0	+	244679	244679	0,255,0	.	AK024243
8	chr7	244826	249951	uc010kry.1	0	+	244826	244826	0,255,0	.	AK310146
9	chr7	288051	304059	uc003sip.1	0	+	288282	304057	0,255,0	.	FAM20C
10	chr7	322636	324640	uc003siq.1	0	-	322636	322636	255,0,0	.	AX746542
11	chr7	503422	526007	uc003sir.1	0	-	506606	525164	255,0,0	PDGFA	P04085
12	chr7	503422	526007	uc003sis.1	0	-	504726	525164	255,0,0	PDGFA	P04085-2
13	chr7	506940	519630	uc003sit.1	0	-	507195	518820	255,0,0	PDGFA	Q32M96
14	chr7	555912	718659	uc003siu.1	0	-	556592	717668	255,0,0	PRKAR1B	Q8N422

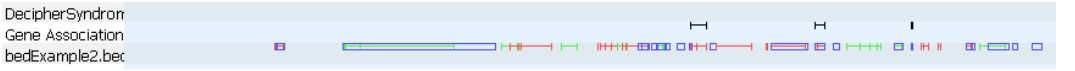
The .BED files are placed in the **customtracks** folder in the respective Organism folder.



The files in the customtracks folder will appear in the **Select Additional Tracks** list in the **Track Selection** tab of the **Options** window.



Marking the checkbox and clicking Apply will add the track to the Nexus display.



Moving the mouse over features in the track will display the BED line name. If RGB colors are in the file, features will be displayed in that color. Thick and thin lines are not supported at this time and neither are exon features in the BED files.

APPENDIX C: SOFTWARE LICENSE AGREEMENT

SOFTWARE LICENSE AGREEMENT AND LIMITED WARRANTY

BEFORE YOU GET STARTED:

1. This is a license to use the software for limited purposes: research and non-diagnostic purposes. Please understand those limits. For example, you cannot use this software under this license for diagnostic purposes or for any service that you provide.

2. Do not activate (or otherwise use) the software if you do not agree with the terms of this License Agreement. If you are reading this license agreement on a package, then it means that the terms of this License Agreement will take effect when you open the package; if you are reading this in a digital format (for, example, online), then the terms of this License Agreement will take effect once you activate the software.

3. If you do not agree with the terms of this License Agreement, then please contact us at support@affymetrix.com

THIS SOFTWARE LICENSE AGREEMENT AND LIMITED WARRANTY ("AGREEMENT") IS ENTERED INTO BY AND BETWEEN BIODISCOVERY, INC. ("LICENSOR") AND YOU WHETHER YOU ARE AN INDIVIDUAL OR AN ENTITY ("LICENSEE"). READ THE FOLLOWING TERMS AND CONDITIONS CAREFULLY BEFORE OPENING THE SEALED PACKAGE CONTAINING THE SOFTWARE (if you are reading this on a package) OR BEFORE ACTIVATING THE SOFTWARE (OR DOWNLOADING THE SOFTWARE), AND BEFORE PROCEEDING FURTHER WITH THE USE OR INSTALLATION OF THIS SOFTWARE.

BY YOUR OPENING OF THE PACKAGE CONTAINING THIS SOFTWARE, OR BY INSTALLING OR UTILIZING THE INSTANT SOFTWARE, YOU AGREE TO BE BOUND BY THE TERMS AND CONDITIONS SET FORTH HEREIN.

THIS SOFTWARE PRODUCT IS PROTECTED BY COPYRIGHT LAWS AND INTERNATIONAL COPYRIGHT TREATIES, AS WELL AS OTHER INTELLECTUAL PROPERTY LAWS AND

TREATIES, AND THIS AGREEMENT. THE SOFTWARE PRODUCT WHICH IS THE SUBJECT OF THIS AGREEMENT IS LICENSED UNDER THIS AGREEMENT, NOT SOLD.

1. LICENSE GRANT. For consideration promised and/or received, Licensor hereby grants to Licensee one nonexclusive, nontransferable, internal, end-user license (the "License") to use, except for any commercial diagnostic purpose or use, the basic software product entitled Nexus Express and those software modules expressly authorized in writing by Licensor, if any, (the "Software"), and the accompanying documentation in the form delivered to Licensee. If Licensee would like to use the Software and accompanying documentation for any commercial diagnostic purpose or use, Licensee will need to purchase a separate and additional license from Licensor.

Unless Licensee has requested and expressly obtained written permission from Licensor, and until such time that Licensee has paid a multiple licensee fee for the concurrent use of the Software, the Software is licensed as a single product and, notwithstanding the fact that the Software itself does execute and/or access multiple central processing units ("CPUs") concurrently, Licensee shall not separate, execute, or access the Software for use on more than one CPU at any one given time. Subject to Licensee's purchase of more than one License, the License granted hereunder is for use only upon a single stand alone computer and only one instance of the Software may be executed and/or accessed at any one time, where such computer upon which the Software is executed and/or accessed is owned, leased, or otherwise substantially controlled by Licensee. Subject to Licensee's purchase of more than one License, neither concurrent use on two or more computers nor use in a local area network or other network is permitted. Upon having purchased and obtained written consent from Licensor to hold more than one License to the Software, Licensee may concurrently load, use, or install the Software upon the number of computers or CPU's for which Licensee expressly holds a License.

Upon loading the Software, Licensee may retain the Software for backup purposes only. In addition, Licensee may make one copy of the Software on a second set of diskettes (or on compact disc or cassette tape) for the purpose of backup in the event the original diskette(s) or compact disc(s) for the Software is damaged or destroyed. Licensee may make one copy of the User's Manual for backup purposes only. Any such copies of the

Software or the User's Manual shall include Licensor's copyright and other proprietary notices. Except as authorized under this paragraph, no copies of the Software, the accompanying documentation, or any portions thereof may be made by Licensee or any person under Licensee's control or authority.

Licensee agrees that it will not use the Software for any purpose that is illegal or in any way that might result in any loss of its or any third party's property or information. Licensee agrees that it shall have the sole responsibility for protecting its data used in connection with the Software.

The terms and conditions of this Agreement shall apply to all additional, subsequent or multiple Licenses obtained by Licensee for the Software.

2. RESTRICTIONS. Licensee agrees that it will not assign, sell, sublicense, transfer, pledge, lease, rent, or share the Software, the accompanying documentation or Licensee's rights under this Agreement, nor delegate any of Licensee's obligations under this Agreement. Any attempted assignment, sale, sublicense, transfer, pledge, lease, rental or delegation shall be null and void. Licensee shall not utilize the Software to provide any analysis services directly to third parties for any compensation without first obtaining the express written consent of the Licensor. Licensee shall not, nor attempt to, reverse engineer, decompile, disassemble, modify, reproduce, reverse assemble, reverse compile or otherwise translate the Software or any part thereof.

3. LICENSOR'S RIGHTS. Licensee agrees and acknowledges that the Software and the accompanying documentation which are the subject of this Agreement are proprietary, confidential, and trade secret products of Licensor and/or Licensor's suppliers and that Licensee shall undertake all necessary steps and efforts to prevent unlawful or illegal distribution of such proprietary, confidential and trade secret information. Licensee further acknowledges and agrees that all right, title, and interest in and to the Software and the accompanying documentation, including associated intellectual property rights, are and shall remain with Licensor and/or Licensor's suppliers. This Agreement does not convey to Licensee any interest or rights in or to the Software and the accompanying documentation, except only a limited revocable right of use in accordance with the

terms of this Agreement. All rights not expressly granted are reserved by Licensor and its suppliers. Licensor may assign this Agreement to a third party at any time.

4. RESTRICTED RIGHTS. Licensee will comply with all applicable laws, regulations, treaties or other agreements in connection with its use of the Software and accompanying documentation. Without limiting the foregoing, Licensee hereby covenants that neither the Software and the accompanying documentation nor any information or know-how embodied in such Software and accompanying documentation will be directly or indirectly provided, transported or removed, or authorized to be provided, transported or removed, in contravention of any export laws, regulations or decrees of the U.S. Government or any agency thereof. This Agreement is subject to termination by Licensor in the event Licensee fails to comply with any such laws, regulations or decrees.

5. LICENSE FEES. Subject only to section 8 of this Agreement, the license fees paid by Licensee in consideration of the License granted under this Agreement are non refundable and shall not be returned or credited to Licensee under any circumstance, including, but not limited to any request for a pro rata refund by Licensee or any reason for termination of this Agreement.

6. TERM. This Agreement is effective upon Licensee's opening of the package containing the Software, or upon Licensee's acceptance of this Agreement. This Agreement shall continue thereafter until terminated. Licensee may terminate this Agreement at any time by returning the Software and all copies thereof and extracts therefrom to Licensor. Licensor may terminate this Agreement and revoke any License granted hereunder upon the breach by Licensee of any term hereof. If the License granted hereunder is terminated for any reason, upon notice of such termination, Licensee shall immediately de-install the Software from the computer on which it is installed and shall certify to Licensor in writing, under penalty of perjury of the laws of the United States of America, that the Software is de-installed and all copies thereof have either been destroyed or returned to Licensor. Any confidential, proprietary, or trade secret information or material provided to Licensee in connection with the Software shall be immediately returned to Licensor, unless otherwise specified by

Licensor. The provisions of sections 2-5, 7, 8-10, 12-16 and 19 shall survive any termination of this Agreement.

7. CONFIDENTIAL INFORMATION. Licensee hereby acknowledges that the Software and any accompanying documentation contain confidential, proprietary, and/or trade secret information belonging to Licensor. Licensee further acknowledges and agrees that it shall not disclose the Software and any accompanying documentation to any third party. Licensee further acknowledges and agrees that any written documentation provided by Licensor to Licensee which contains a legend upon such documentation, whether or not such legend be a single legend affixed upon a multiple page document, which legend identifies such document to be either proprietary, trademarked, registered, copyrighted, confidential and/or trade secret, shall impose a duty upon Licensee not to disclose to any third party the documentation or any information contained within such documentation, either in writing or orally, without the express written consent of Licensor. Notwithstanding the foregoing provision, Licensor may notify Licensee in writing within twenty (20) days after disclosure to Licensee of documents which do not contain a legend identifying such documents to be either proprietary, confidential and/or trade secret, that such documents disclosed were either proprietary, trademarked, registered, copyrighted, confidential and/or trade secret in nature. Such notice shall impose a duty upon the Licensee not to disclose to any third party such documentation or any information contained with such documentation, either in writing or orally, without the express written consent of Licensor. Licensee further acknowledges that any oral information provided by Licensor to Licensee which information is identified or summarized in writing within twenty (20) days after such oral disclosure to be either proprietary, trademarked, registered, copyrighted, confidential and/or trade secret in nature shall impose a duty upon Licensee not to disclose to any third party such information disclosed by Licensor to Licensee, either in writing or orally, without the express written consent of Licensor. The obligations of this section 7 shall not extend to any information which is lawfully known to Licensee prior to receipt from Licensor or its distributor; or enters the public domain through no wrongful act or breach of this Agreement by Licensee; or is received by Licensee from a third party having a legal right to disclose such information.

8. LIMITED WARRANTY. Licensor warrants for a period of 90 days from the date of commencement of this Agreement ("Warranty Period") that during the Warranty Period the Software shall operate substantially in accordance with the functional specifications in the User's Manual. LICENSOR FURTHER WARRANTS THAT DURING THE WARRANTY PERIOD THE MEDIA WHICH CONTAINS THE SOFTWARE SHALL BE FREE FROM DEFECTS IN MATERIAL AND WORKMANSHIP. LICENSEE'S SOLE AND EXCLUSIVE REMEDY, AND LICENSOR'S SOLE LIABILITY ARISING FROM BREACHES OF THE ABOVE WARRANTIES IS THE REPLACEMENT OF DEFECTIVE MEDIA OR, IF LICENSEE SHALL SO REQUEST, TO REFUND TO LICENSEE THE PURCHASE PRICE FOR THE DEFECTIVE SOFTWARE AND DOCUMENTATION, PROVIDED THAT LICENSEE NOTIFIES LICENSOR IN WRITING OF SUCH DEFECT AND RETURNS TO LICENSOR THE DEFECTIVE MEDIA CONTAINING THE SOFTWARE AND THE DOCUMENTATION, DURING THE ABOVE WARRANTY PERIOD.

EXCEPT AND TO THE EXTENT EXPRESSLY PROVIDED ABOVE, THE SOFTWARE AND DOCUMENTATION WHICH ARE THE SUBJECT OF THIS AGREEMENT ARE PROVIDED ON AN "AS IS" BASIS, WITHOUT ANY WARRANTIES OF ANY KIND, INCLUDING ANY AND ALL IMPLIED WARRANTIES OR CONDITIONS OF TITLE, NONINFRINGEMENT, MERCHANTABILITY, OR FITNESS OR SUITABILITY FOR ANY PARTICULAR PURPOSE, WHETHER ALLEGED TO ARISE BY LAW, BY REASON OF CUSTOM OR USAGE IN THE TRADE, OR BY COURSE OF DEALING. IN ADDITION, LICENSOR EXPRESSLY DISCLAIMS ANY WARRANTY OR REPRESENTATION TO ANY PERSON OTHER THAN LICENSEE WITH RESPECT TO THE SOFTWARE AND DOCUMENTATION WHICH ARE THE SUBJECT OF THIS AGREEMENT. LICENSEE ASSUMES THE ENTIRE LIABILITY FOR THE SELECTION AND USE OF THE SOFTWARE AND DOCUMENTATION, AND LICENSOR SHALL HAVE NO LIABILITY FOR ANY ERRORS, MALFUNCTIONS, DEFECTS, LOSS OF DATA, OR ECONOMIC LOSS RESULTING FROM OR RELATED TO THE USE OF SOFTWARE AND/OR DOCUMENTATION.

9. LIMITATION OF LIABILITY. Notwithstanding any other provision of this Agreement, the cumulative liability of Licensor and/or Licensor's suppliers, distributors and/or agents to Licensee or any other party for any loss or damages resulting from any claims, demands, or actions arising out of or relating to the Software, the accompanying documentation and/or this Agreement shall not exceed that license fee paid to Licensor by Licensee for the use of the Software and accompanying documentation. In no event

shall Licensor and/or Licensor's suppliers, distributors and/or agents be liable for any indirect, incidental, punitive, consequential, special, or exemplary damages, for any lost profits or savings, or for any loss of data or loss of use of equipment, even if Licensor and/or Licensor's suppliers, distributors and/or agents have been advised of the possibility of such damages and regardless of the form of action whether in tort, contract or otherwise. SOME STATES DO NOT ALLOW THE LIMITATION OR EXCLUSION OF LIABILITY FOR INCIDENTAL OR CONSEQUENTIAL DAMAGES, SO THE ABOVE LIMITATION OR EXCLUSION MAY NOT APPLY TO LICENSEE.

10. TRADEMARK. BioDiscovery (R), ImaGene(R), GeneSight(R), and GenePie(R) are registered trademarks of Licensor. No right, license, or interest to such trademarks is granted hereunder, and Licensee agrees that no such right, license, or interest shall be asserted by Licensee with respect to such trademarks.

11. NOTICE. All notices required or provided under the terms of this Agreement shall be given in writing to all parties and may be delivered by First Class U. S. Mail, postage prepaid; U.S. Registered Air Mail, postage prepaid; overnight air courier, courier charges prepaid; or facsimile. Notices shall be effective as follows: five (5) calendar days following mailing by First Class U.S. Mail, postage prepaid; seven (7) calendar days following mailing by U.S. Registered Mail, postage prepaid; two (2) business days following delivery by overnight courier; and two (2) business days following confirmation of transmittal by facsimile. Any notices provided under this Agreement shall be given at the address and/or facsimile number for the parties as set forth upon the sales document for this License, unless change of such address and/or facsimile number has been provided previously in writing.

12. GOVERNING LAW AND VENUE. This Agreement shall be construed and governed in accordance with the laws of the State of California, without giving effect to any choice or conflict of law provision or rule. Licensor and Licensee consent and agree that personal jurisdiction over them with respect to any dispute arising as to this Agreement shall rest solely with the state or federal courts of the State of California. Licensor and Licensee hereby expressly waive the right to bring an action in any state or federal court other than the California state or federal courts located within the County

of Los Angeles. This Agreement shall not be governed by the United Nations Convention on Contracts for the International Sale of Goods, the application of which is expressly excluded.

13. ATTORNEYS' FEES. If any action is brought by either party to this Agreement against the other party in an effort to enforce or effect any provision or language contained within this Agreement, the prevailing party shall be entitled to recover, in addition to any other relief granted, reasonable attorney fees and costs.

14. SEVERABILITY. If any provision of this Agreement shall be held illegal, unenforceable, or in conflict with any law of a federal, state, or local government having jurisdiction over this Agreement, the validity of the remaining portions or provisions hereof shall not be affected thereby.

15. NO WAIVER. The failure of either party to enforce any rights granted hereunder or to take action against the other party in the event of any breach hereunder shall not be deemed a waiver by that party as to subsequent enforcement of rights or subsequent actions in the event of future breaches.

16. ENTIRE AGREEMENT. Licensee acknowledges that it has read this Agreement, understands it, and agrees to be bound by its terms. This Agreement and any modifications made pursuant to it, constitutes the complete and exclusive written expression of all terms of the Agreement between Licensor and Licensee, and supersedes all prior or contemporaneous proposals, understandings, representations, conditions, warranties, covenants, and all other communications between Licensor and Licensee relating to the subject matter of this Agreement, whether oral or written. This Agreement may not in any way be explained or supplemented by a prior or existing course of dealing between Licensor and Licensee, by any usage of trade or custom, or by any prior performance between Licensor and Licensee pursuant to this Agreement or otherwise.

17. AMENDMENTS. No amendments or other modifications to this Agreement may be made except by a writing signed by Licensor and Licensee.

18. U.S. GOVERNMENT RIGHTS. If Licensee is the U.S. Government or if Licensee is a contractor or subcontractor (at any tier) of the U.S. Government and is licensing the Software for use by the U.S. Government or in connection with any contract or other transaction with the U.S. Government, Licensee acknowledges that by accepting delivery of the Software, the U.S. Government agrees that the Software qualifies as commercial computer software and that the accompanying documentation qualifies as commercial computer software documentation within the meaning of the acquisition regulations and contract clauses applicable to this procurement. The terms and conditions of this Agreement are fully and exclusively applicable to the Government's use and disclosure of the Software and accompany documentation, and shall supersede any conflicting terms or conditions. No license of any kind is granted in the case of acquisitions which contain or are subject to the clauses FAR 52-227.19 COMMERCIAL COMPUTER SOFTWARE-RESTRICTED RIGHTS (JUNE 1987) or DFARS 252.227-7013 RIGHTS IN TECHNICAL DATA AND COMPUTER SOFTWARE (OCT 1988) or any other clause which purports to grant to the U.S. Government rights greater than, or additional to those, set forth in this Agreement, or which purports to impose additional requirements upon Licensor to make this Agreement effective, unless Licensor specifically so consents by separate written agreement. Failing such agreement, and if this Agreement fails to meet the U.S. Government's stated needs or is inconsistent in any respect with federal law, the U.S. Government agrees to return the Software and accompanying documentation, unused, to Licensor. The Contractor/Licensor/Manufacturer is BioDiscovery, Inc., 2301 Rosecrans Blvd., Suite 2185, El Segundo, CA 90245.

19. INDEMNITY. Licensee acknowledges that Licensor has no knowledge of, or control over, the uses of the Software and accompany documentation made by Licensee. Licensee agrees to defend, indemnify and hold Licensor harmless from and against any and all losses, liabilities, damages, costs and expenses (including but not limited to reasonable attorney's fees) arising out of or related to any suit, claim or proceeding relating to the use of the Software and accompanying documentation, including without limitation any loss related to Licensee's failure to conform to the requirements of section 4 of this Agreement.

20. UPDATING. Licensee acknowledges and agrees that the Software may communicate via the Internet or other communications systems with Licensor's computer systems for the purpose of checking the status of the License, checking the status of maintenance for the Software, and/or determining if any fixes, updates and/or upgrades to the Software are available. Licensee agrees and consents to any such communications between the Software and Licensor's computer systems and the transfer of data between the Software and Licensor's computer system. Licensee also agrees and consents to, if applicable, remote configuration of the Software on Licensee's computer systems.

Licensee acknowledges and agrees that the Software permits downloading of Software fixes, updates and upgrades, and that such fixes, updates and upgrades may occur without notice. Licensee agrees to install and use any and all such fixes, updates and upgrades and, if applicable, discontinue use of the previous version of the Software. Licensee agrees that any and all such fixes, updates and upgrades are and shall be governed by this Agreement unless superseded by an agreement associated with such a fix, update and/or upgrade.

21. PRIVACY. Licensor believes that the personal information Licensee provides to Licensor must be both kept private, and used in a responsible fashion. Licensor consistently puts its best efforts towards achieving both these objectives. The updating feature of the Software provides information (such as product serial number, operating system information, product language, etc.) to Licensee for the purpose of checking the status of the License, checking the status of maintenance for the Software, and/or determining if any fixes, updates and/or upgrades to the Software are available and Licensee agrees and consents that Licensor may store, process and use such information for that purpose. Licensee agrees that in the event that Licensor sells, merges or otherwise re-organizes all or part of its business with another entity, the other entity may have access to and process, store and use Licensee's information, including the information collected through the updating feature of the Software, on the same terms and conditions as set forth in this Agreement. Licensee agrees that Licensor may process and/or store Licensee's information in United States databases. Licensee permits Licensor to share Licensee's information within Licensor and transfer it to countries in

the world where Licensor does business. Licensee permits Licensor to disclose Licensee's information when Licensor is required to do so by law.

APPENDIX D: THIRD PARTY LICENSE AGREEMENTS

JOPT SIMPLE VERSION

Academic Free License ("AFL") v. 3.0

This Academic Free License (the "License") applies to any original work of authorship (the "Original Work") whose owner (the "Licensor") has placed the following licensing notice adjacent to the copyright notice for the Original Work:

Licensed under the Academic Free License version 3.0

1. Grant of Copyright License. Licensor grants You a worldwide, royalty-free, non-exclusive, sub licensable license, for the duration of the copyright, to do the following:

1. to reproduce the Original Work in copies, either alone or as part of a collective work;

2. to translate, adapt, alter, transform, modify, or arrange the Original Work, thereby creating derivative works ("Derivative Works") based upon the Original Work;

3. to distribute or communicate copies of the Original Work and Derivative Works to the public, under any license of your choice that does not contradict the terms and conditions, including Licensor's reserved rights and remedies, in this Academic Free License;

4. to perform the Original Work publicly; and

5. to display the Original Work publicly.

2. Grant of Patent License. Licensor grants You a worldwide, royalty-free, non-exclusive, sublicensable license, under patent claims owned or controlled by the Licensor that are embodied in the Original Work as furnished by the Licensor, for the duration of the patents, to make, use, sell, offer for sale, have made, and import the Original Work and Derivative Works.

3. Grant of Source Code License. The term "Source Code" means the preferred form of the Original Work for making modifications to it and all available documentation describing how to modify the Original Work. Licensor agrees to provide a machine-readable copy of the Source Code of the Original Work along with each copy of the Original Work that Licensor distributes. Licensor reserves the right to satisfy this obligation by placing a machine-readable copy of the Source Code in an information repository reasonably calculated to permit inexpensive and convenient access by You for as long as Licensor continues to distribute the Original Work.

4. Exclusions From License Grant. Neither the names of Licensor, nor the names of any contributors to the Original Work, nor any of their trademarks or service marks, may be used to endorse or promote products derived from this Original Work without express prior permission of the Licensor. Except as expressly stated herein, nothing in this License grants any license to Licensor's trademarks, copyrights, patents, trade secrets or any other intellectual property. No patent license is granted to make, use, sell, offer for sale, have made, or import embodiments of any patent claims other than the licensed claims defined in Section 2. No license is granted to the trademarks of Licensor even if such marks are included in the Original Work. Nothing in this License shall be interpreted to prohibit Licensor from licensing under terms different from this License any Original Work that Licensor otherwise would have a right to license.

5. External Deployment. The term "External Deployment" means the use, distribution, or communication of the Original Work or Derivative Works in any way such that the Original Work or Derivative Works may be used by anyone other than You, whether those works are distributed or communicated to those persons or made available as an application intended for use over a network. As an express condition for the grants of license hereunder, You must treat any External Deployment by You of the Original Work or a Derivative Work as a distribution under section 1(c).

6. Attribution Rights. You must retain, in the Source Code of any Derivative Works that You create, all copyright, patent, or trademark notices from the Source Code of the Original Work, as well as any notices of licensing and any descriptive text identified therein as an "Attribution Notice." You must cause the Source Code for any Derivative

Works that You create to carry a prominent Attribution Notice reasonably calculated to inform recipients that You have modified the Original Work.

7. Warranty of Provenance and Disclaimer of Warranty. Licensor warrants that the copyright in and to the Original Work and the patent rights granted herein by Licensor are owned by the Licensor or are sublicensed to You under the terms of this License with the permission of the contributor(s) of those copyrights and patent rights. Except as expressly stated in the immediately preceding sentence, the Original Work is provided under this License on an "AS IS" BASIS and WITHOUT WARRANTY, either express or implied, including, without limitation, the warranties of non-infringement, merchantability or fitness for a particular purpose. THE ENTIRE RISK AS TO THE QUALITY OF THE ORIGINAL WORK IS WITH YOU. This DISCLAIMER OF WARRANTY constitutes an essential part of this License. No license to the Original Work is granted by this License except under this disclaimer.

8. Limitation of Liability. Under no circumstances and under no legal theory, whether in tort (including negligence), contract, or otherwise, shall the Licensor be liable to anyone for any indirect, special, incidental, or consequential damages of any character arising as a result of this License or the use of the Original Work including, without limitation, damages for loss of goodwill, work stoppage, computer failure or malfunction, or any and all other commercial damages or losses. This limitation of liability shall not apply to the extent applicable law prohibits such limitation.

9. Acceptance and Termination. If, at any time, You expressly assented to this License, that assent indicates your clear and irrevocable acceptance of this License and all of its terms and conditions. If You distribute or communicate copies of the Original Work or a Derivative Work, You must make a reasonable effort under the circumstances to obtain the express assent of recipients to the terms of this License. This License conditions your rights to undertake the activities listed in Section 1, including your right to create Derivative Works based upon the Original Work, and doing so without honoring these terms and conditions is prohibited by copyright law and international treaty. Nothing in this License is intended to affect copyright exceptions and limitations (including “fair use” or “fair dealing”). This License shall terminate immediately and You may no longer

exercise any of the rights granted to You by this License upon your failure to honor the conditions in Section 1(c).

10. Termination for Patent Action. This License shall terminate automatically and You may no longer exercise any of the rights granted to You by this License as of the date You commence an action, including a cross-claim or counterclaim, against Licensor or any licensee alleging that the Original Work infringes a patent. This termination provision shall not apply for an action alleging patent infringement by combinations of the Original Work with other software or hardware.

11. Jurisdiction, Venue and Governing Law. Any action or suit relating to this License may be brought only in the courts of a jurisdiction wherein the Licensor resides or in which Licensor conducts its primary business, and under the laws of that jurisdiction excluding its conflict-of-law provisions. The application of the United Nations Convention on Contracts for the International Sale of Goods is expressly excluded. Any use of the Original Work outside the scope of this License or after its termination shall be subject to the requirements and penalties of copyright or patent law in the appropriate jurisdiction. This section shall survive the termination of this License.

12. Attorneys' Fees. In any action to enforce the terms of this License or seeking damages relating thereto, the prevailing party shall be entitled to recover its costs and expenses, including, without limitation, reasonable attorneys' fees and costs incurred in connection with such action, including any appeal of such action. This section shall survive the termination of this License.

13. Miscellaneous. If any provision of this License is held to be unenforceable, such provision shall be reformed only to the extent necessary to make it enforceable.

14. Definition of "You" in This License. "You" throughout this License, whether in upper or lower case, means an individual or a legal entity exercising rights under, and complying with all of the terms of, this License. For legal entities, "You" includes any entity that controls, is controlled by, or is under common control with you. For purposes of this definition, "control" means (i) the power, direct or indirect, to cause the direction or management of such entity, whether by contract or otherwise, or (ii) ownership of

fifty percent (50%) or more of the outstanding shares, or (iii) beneficial ownership of such entity.

15. Right to Use. You may use the Original Work in all ways not otherwise restricted or conditioned by this License or by law, and Licensor promises not to interfere with or be responsible for such uses by You.

16. Modification of This License. This License is Copyright © 2005 Lawrence Rosen. Permission is granted to copy, distribute, or communicate this License without modification. Nothing in this License permits You to modify this License as applied to the Original Work or to Derivative Works. However, You may modify the text of this License and copy, distribute or communicate your modified version (the "Modified License") and apply it to other original works of authorship subject to the following conditions: (i) You may not indicate in any way that your Modified License is the "Academic Free License" or "AFL" and you may not use those names in the name of your Modified License; (ii) You must replace the notice specified in the first paragraph above with the notice "Licensed under <insert your license name here>" or with a notice of your own that is not confusingly similar to the notice in this License; and (iii) You may not claim that your original works are open source software unless your Modified License has been approved by Open Source Initiative (OSI) and You comply with its license review and certification process.