Importing & Exporting Data

Introduction

This chapter describes how to import data and prepare it for analysis. Partek can access data from many sources. The first part of this chapter will describe the following:

- Importing Text Files (CSV or TXT)
- Pivoting (transposing) data during import
- Importing from the clipboard
- Importing from Microsoft Excel Workbooks (XLS)
- Creating a new spreadsheet
- Merging two spreadsheets
- Opening Partek format files
- Saving your data

Data Types

Since there are no software imposed limits for the number of rows and number of columns in a spreadsheet, Partek can easily handle very large data sets. Note: usually, in Partek, observations (e.g. sample, subject) are on rows and variables (e.g. sample attributes like name, gender; measurements like genes) are on columns. Partek automatically imports the following data types:

- Text files (tab, comma, and user-specified delimiters)
- Windows Clipboard
- Affymetrix® CEL, CHP, and EXP files, see the **Genomics Specific Files** section below
- Agilent, GPR, and other genomics file formats. See the **Genomics Specific Files** section below
- Plate-based data (Low, High, or Ultra-high Throughput Screening)
- Partek Format Files (.fmt)

Data Import Steps

Data import typically requires three steps:

- Select a data source
- Select the layout for your data and specify the column properties
- Save the data to create a *Partek Format File*. This will allow you to open the same data without having to repeat the import process each time

Partek provides an import wizard that will guide you through the import process for many types of data. The following sections will describe the importing data process.

Column Types and Attributes in Partek

The Analytical Spreadsheet® requires rectangular data, much like a database. It also requires homogeneous column types. Note: usually, in Partek, observations (e.g. sample, subject) are on rows and variables (e.g. sample attributes like name, gender; measurements like genes) are on columns. If your data has observations on columns and variables on rows, you need to pivot or transpose your data.

Homogeneous column types are defined as either text, categorical, or numerical. Numerical columns can be doubles (4 bytes), floats (2 bytes), or integers (1/2/4 bytes). If you have extremely large data you may want to consider loading the data as floats to reduce the initial memory requirement.

Column Types

There are seven types of columns in Partek, but only four are generally used; they are bolded below.

Type	Description
text	**variable length string**
categorical	**variable length nominal**
double	**double precision floating point (8 bytes) (-1.7E308 to 1.7E308)**
float	**single precision floating point (4 bytes) (-3.8E38 to 3.4E38)**
integer	integer (4 bytes) (-2147183648 to 2147483647)
short	short integer (2 bytes) (-32768 to 32767)
byte	1 byte (0 to 255)
snp	genotype data can only be AA, BB, AB, NC values

Table 4. 1: Identifying column types in Partek

Note: By default, numerical columns are automatically imported into Partek as response and double precision. Text columns are automatically imported as variable length string columns.

Column Attributes

Type	Description
factor	a variable that causes or influences another variable
response	a variable that is caused by or influenced by another variable

Table 4. 2: Identifying column attributes in Partek

When importing text file, Partek will automatically detect text, categorical, and numerical column types. You should always take a quick look at the column types that are determined during the import process. If a column of numerical values

(gene expression, blood pressure, weight, IC50, etc.) is automatically detected as text or categorical, this is an indication that a non-numerical character exists in that column and the data should be examined in more detail.

Each section illustrates some of the common scenarios you may encounter when first importing your data. These are not intended as detailed step-by-step instructions for importing data but rather to give some examples of typical data layouts.

Partek Format Files

During the import process, Partek creates a companion file called the *Partek Format File*. This file has the same name as the data file with an *.fmt* extension, for example, if you import a data named **MyData.txt**, a companion file **MyData.txt.fmt** will be created. The format file describes the contents of the data file for Partek. The advantage of the format file is that data only needs to be imported into Partek once. Subsequent analysis of the data can be done by opening the *.fmt* file using **File > Open**. Table 4.3 shows the contents of the format file for the data in examples below**.**

Note: the Partek format file does not contain the raw data; it only contains meta-information about the data.

```
ascii
records 60
offset 1
delimiter "  "
missing ?
data vstring[6] double[22000]
field 1 label
field 2 dependent
cl 1 Subject
cl 2 100001_at
……————————
```

Table 4. 3: Viewing part of the Partek format file (.fmt)

Common File Formats

Importing Text Files

If the data is stored as a text file (CSV or TXT) you will need to import the data using the import utility. If the data was previously imported and a Partek format file was created, the data can be loaded using **File > Open**.

Steps for Importing Text Files

To import text files, follow these steps:

- Select **File > Import > Text (csv, txt**) from the Partek main menu
- Select the file to import
- Verify or select the column delimiter and whether to pivot the data
- Select the column labels, start of data, and missing data symbol (if necessary)
- Verify and/or change the column types and attributes

Example: Importing Data with Multiple Column Labels

While Partek has no software imposed limits on the number of rows or columns, it does restrict the number of non-numeric rows that can be used as column labels to one. The columns in Table 4. 4 represent samples and the rows are measurements on those samples (in this data the columns are gene expression measurements).

Strain	**Control**	**Control**	**Control**	**Control**
Tissue	**Cerebellum**	**Cerebellum**	**Cerebellum**	**Cortex**
Subject	**3396**	**3405**	**3406**	**3396**
100001_at	5.43923	5.42716	5.2433	5.75647
100002_at	8.7084	9.11436	8.89326	7.56515
100003_at	6.14255	6.26519	6.65075	5.80349
100004_at	7.02487	7.07921	7.09967	6.98329

Table 4. 4: Viewing three rows of sample information followed by gene expression data

In Table 4. 4, the data begins in row 4; however, the three rows of sample information (the rows beginning with *Strain*, *Tissue*, and *Subject*) could be used as column labels. It is recommended that you use the *Subject ID* in this case to uniquely identify each column of data. The next example will illustrate pivoting (transposing) the data on input to provide a more powerful look at this data.

- Select **File > Import > Text Files (.csv, .txt)…** browse to the folder containing the .txt file and open it

The *File Type* panel, shown in Figure 4. 1, shows that Partek correctly determined that the file is a *Tab Delimited* file.

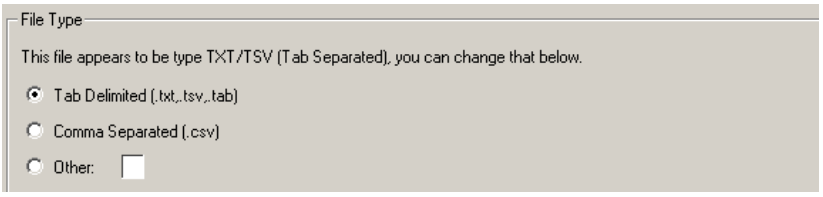


Figure 4. 1: Specifying the tab delimited option

- Click **Next >** to continue

The *Identify Column Labels*, *Start of Data* group box is shown in Figure 4. 2. This data would be imported by selecting row 3 (Subject ID) as the column label and

row 4 for the beginning of the data. This data has no missing values so the *Missing Data Representation* group box (not shown) can be ignored.

- Click **Next >** to continue

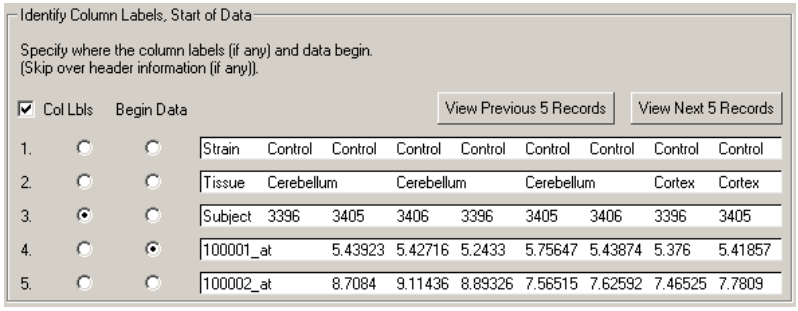


Figure 4. 2: Selecting the column labels and where the data begins

The final page of the import dialog is shown in Figure 4. 3. In this case, Partek has correctly identified column 1 as a text column. The spreadsheet will contain 12,488 rows and 19 columns. The numerical values are shown beginning in column 2.

In this dialog, you can use the left mouse button, <Control> + left button, or <Shift> + left button to select single, multiple, or contiguous column(s). Right click on the column's type (e.g. double, categorical) to assign a new type. Right click on the column's attribute (e.g. factor, response) to assign a new attribute.

- Select **Import** to import the data (Figure 4. 3)

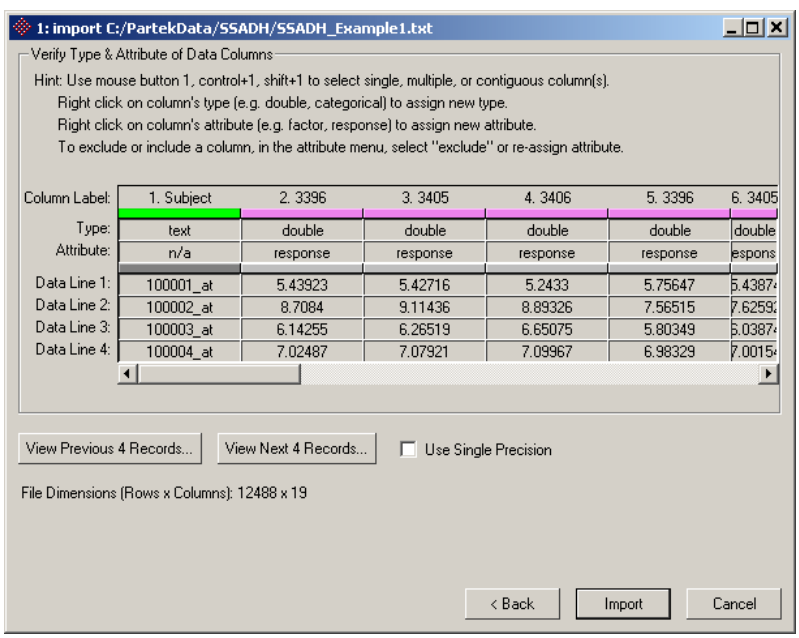


Figure 4. 3: Configuring column types and attributes

General File Information

General file information can be found by selecting **File > Info…** The file's information is sorted into three tabbed panels: *General Info*, *Comments*, and *Format File*. These panels will be discussed below.

The *General Info* tab gives a quick look at the information regarding the file in the active spreadsheet. It includes *Filename*, *Size of Spreadsheet*, *Variable Attributes*, *Row/Column Filters Applied*, and *Other Info* (Figure 4. 4).

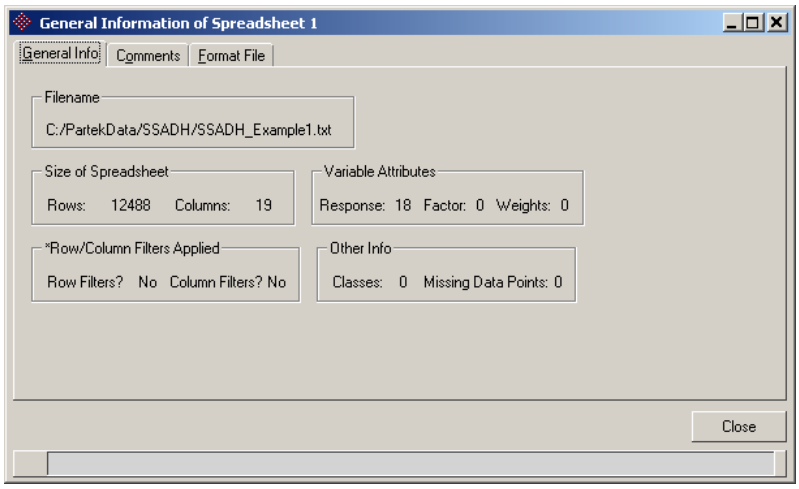


Figure 4. 4: General Info tab

The *Comments* tab offers a place to make comments about the active file. The comments will be saved when the spreadsheet is saved (Figure 4. 5). The comment "Imported January 1, 2005" has been manually added to this file.

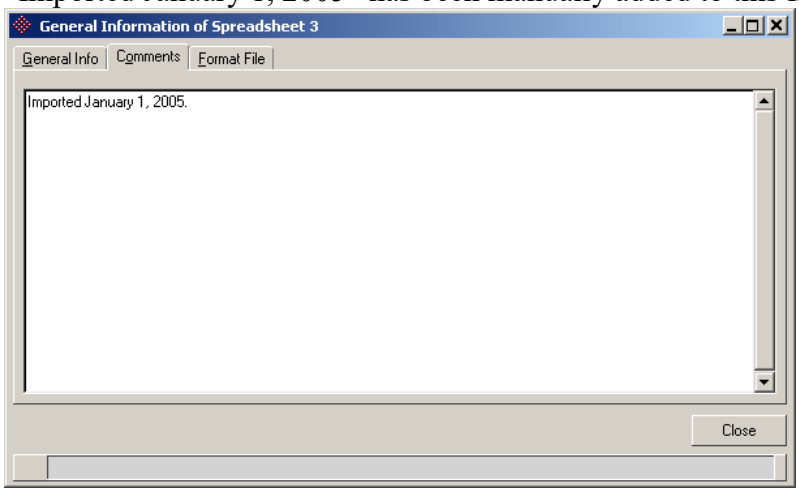


Figure 4. 5: Comments tab

The *Format File* tab shows the format of the active file (Figure 4. 6).

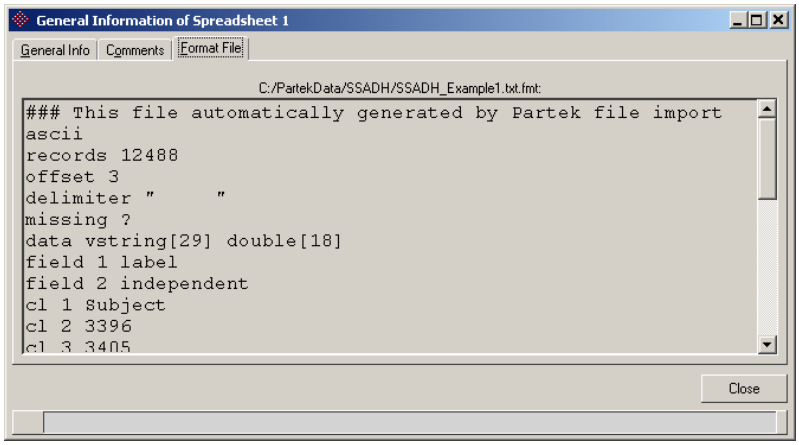


Figure 4. 6: Format File tab

Example: Pivoting (Transposing) Data on Import

Some genomic data is stored with the genes/proteins being represented by the rows of the spreadsheet and the samples being represented by the columns. However, most statistical software products operate with the assumption that the rows represent the observations of interest (samples) and the column the variables measured on them (genes/proteins). Partek provides the ability to pivot data on import to easily load the data into an appropriate format for analysis.

- Select **File > Import > Text (.csv, .txt)** ,browse to the folder containing the text file, and open it

Figure 4. 7 shows the *Import* dialog for importing text files.

- Check the **Transpose the file to** check button (Figure 4. 7)

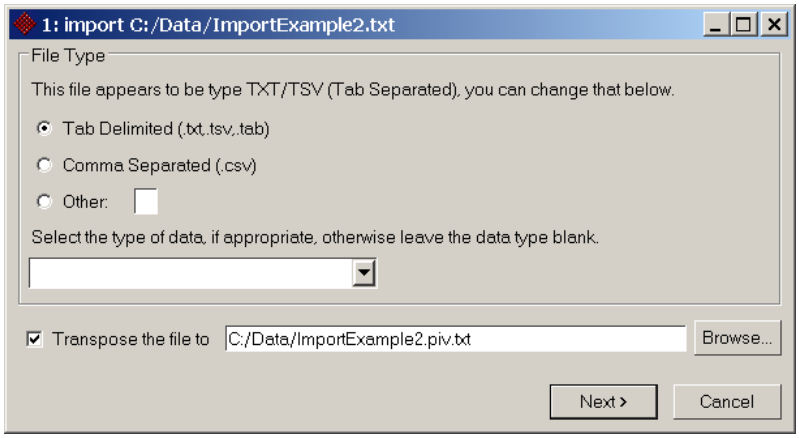


Figure 4. 7: Selecting to Transpose (pivoting) the data on import

When this is selected, the original data will be transposed and stored in the same folder with a .piv extension. In this example, the original data is transposed and a new file is created containing the transposed data. The import procedure will continue using the newly created file containing the pivoted data.

The *Identify Column Labels, Start of Data* dialog is shown in Figure 4. 8. For Affymetrix GeneChip data, it is recommended that Affymetrix probe set identifiers are used as the column labels.

- Select the *Column Labels* as row **1** and the *Begin Data* as row **3**
- Click **Next >** (Figure 4. 8)

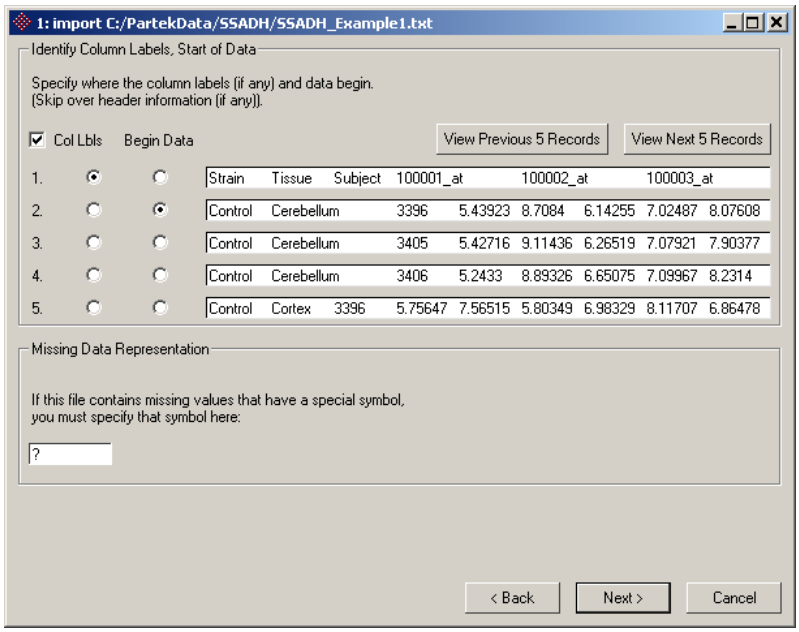


Figure 4. 8: Selecting the column labels and start of data after pivoting

The *Verify Type & Attributes of Data Columns* dialog is shown in Figure 4. 9. Notice that the column types for *Strain* and *Tissue* are set correctly for this data. The *subject IDs* in this data are numerical so Partek correctly coded them as such. In our example, the column *Type* and *Attribute* needs to be changed, to do so:

- Change the column *Type* to **categorical (random effect)**
- Change the column *Attribute* to **factor**
- Select **Import** (Figure 4. 9)

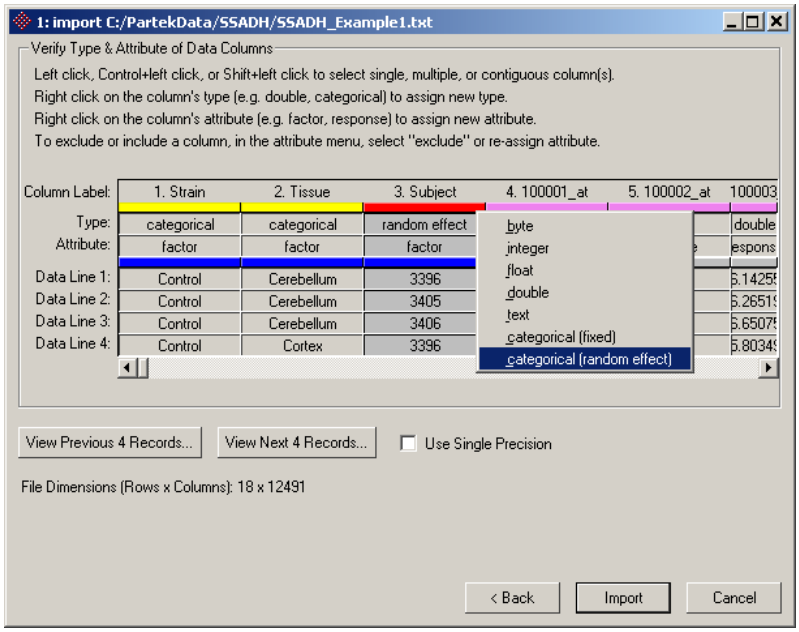


Figure 4. 9: Verifying column types and attributes after pivoting

Figure 4. 10 shows the data from Table 4. 3 successfully pivoted and imported into Partek. Columns 1, 2, & 3 contain the *Strain*, *Tissue*, and *Subject* information for each sample. For data stored in this format, this is a convenient way to include any non-expression data with the expression data.

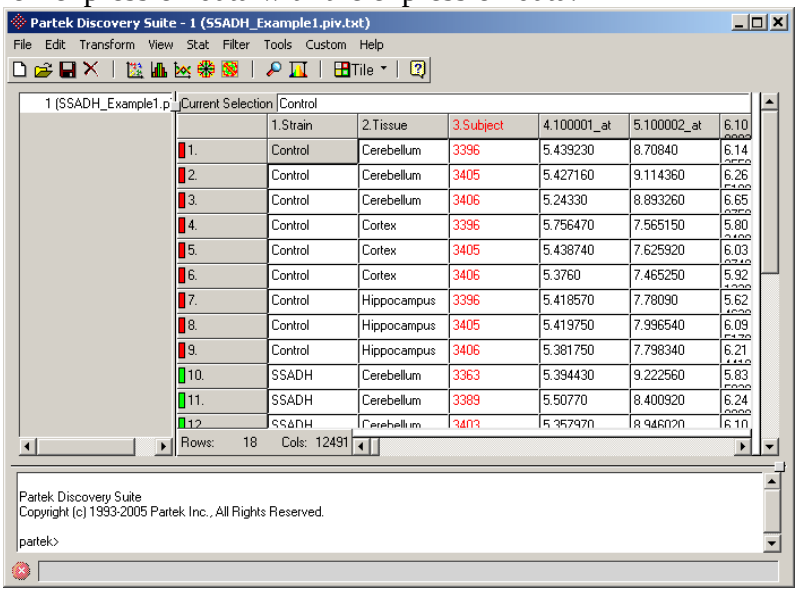


Figure 4. 10: Viewing data from Table 2 after importing into Partek

Example: Potential Problem when Pivoting on Import

The data displayed in Table 4. 5 illustrates a potential problem when pivoting data on import.

ID	Sample1	Sample2	Sample3	Sample4	Description
Treatment	Treated	Treated	Control	Control	

Gender	F	M	F	M	
AFFX-r2-P1-cre-5_at	1144.3	876.2	465.3	8834.4	Bacteriophage
AFFX-r2-Bs-thr-5_s_at	1223.4	1105.5	629.7	7347.1	B. subtilis
AFFX-b-ActinMur/M12481_5_at	1023.5	671.7	301.8	6310.9	M. musculus

Table 4. 5: Potential problem pivoting on import

The pivoted data is shown in Table 4. 6.

ID	Treatment	AFFX-r2-P1-cre-5_at	AFFX-r2-Bs-thr-5_s_at	AFFX-b-ActinMur/M12481_5_at
Sample1	Treated	1144.3	1223.4	1023.5
Sample2	Treated	876.2	1105.5	671.7
Sample3	Control	465.3	629.7	301.8
Sample4	Control	8834.4	7347.1	6310.9
Sample5	Control	0.617411	13.6	0.239005
Description		**Bacteriophage**	**B. subtilis**	**M. musculus**

Table 4. 6: Pivoted data of Table 4. 5

The two rows of column header information are in the first and last rows and are highlighted in yellow. In this case, there is a row of column IDs followed by 5 samples with a trailing row of column descriptions. Because Partek requires homogenous column types, the software will interpret all 5 columns as non-numeric during import.

This data must be preprocessed prior to importing into Partek if you intend to pivot the data during the import. Prior to importing this data into Partek, delete column 6 (Description) displayed in Table 3 or move it adjacent to column 1 (ID).

Pasting from the Clipboard to a Spreadsheet

You can copy and paste the text contents from other software directly into the Analytical Spreadsheet®. From other software like a text editor or Microsoft Excel®, select the contents of the file, and select **Copy**. In the Analytical Spreadsheet, select **File > Paste to Spreadsheet…** to paste contents into the spreadsheet.

Example: Creating a Sample Information File

The following steps show how to create a sample information file:

1. Open any text editor
2. Type .CEL file names and other subject/sample information in the following format:

```
ChipType1        ChipType2...       Attribute1       Attribute2...
CT1_File1.cel    CT2_File1.cel      Subject1_Attr1   S1_Attr2...
CT1_File2.cel    CT2_File2.cel      Subject2_Attr1   S2_Attr2...
...
```

All the delimiters are Tabs, and the first line contains the column headers. You may specify one, two, or more *ChipType* columns. For example, if all the chips you have are Affymetrix® GeneChip HG_U133plus2, then you will have only one *ChipType*. If you run every subject/sample on, for instance, Affymetrix® Mapping 250K Nsp and Mapping 250K Sty (combined to make up Mapping 500K), you must use two *ChipType* columns to specify which two chips are from the same subject. The rest of the columns specify the subject's attributes (e.g. gender, tissue, age etc.).

3. Copy all the edited content to the clipboard

You can also use Microsoft® Excel™ to do steps 1-3.

4. In Partek, select **Edit > Paste to New Spreadsheet**
5. Verify the correctness in the Partek Analytical Spreadsheet®. Note: you can edit the content in Partek
6. Select **File > Save** to save the spreadsheet
7. Select **File > Close** to close the spreadsheet. Note: Partek will create 2 files e.g. ABC and ABC.fmt. ABC contains the content and ABC.fmt contains the format. Use ABC.fmt as the sample information file.

Creating a New Spreadsheet

The *Create New* dialog configures a new spreadsheet based on your specifications. You can specify the number of *Rows* as well as the number of *Data Fields*. For each data field, you can specify the *Field Type, Size,* and *Field Attribute* (Figure 4. 11).

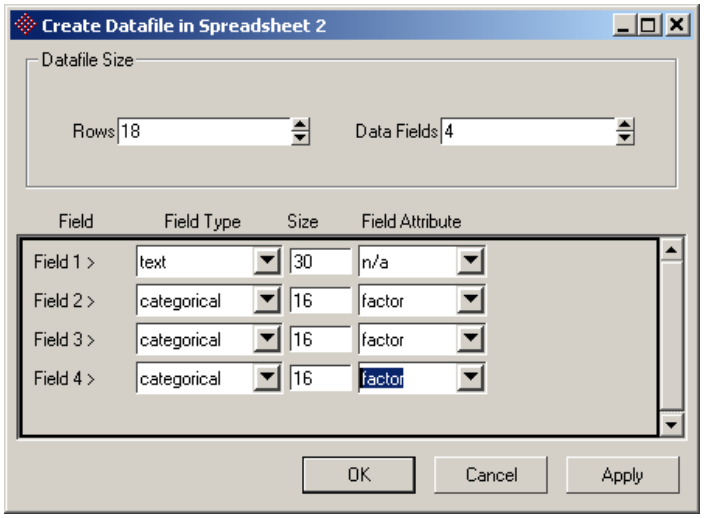


Figure 4. 11: Creating a Datafile in the Spreadsheet dialog

Field Types to choose from are **double, text, categorical, float, integer, short,** and **byte.** *Field Attributes* include **response, factor,** or **label.** After you have finished configuring the dimensions, select **OK** to create the new spreadsheet and dismiss the dialog or select **Apply** to create the new spreadsheet but keep the dialog open.

After creating the spreadsheet, you can type and edit the content in the spreadsheet, then save the spreadsheet. Note: those steps can be used to create a sample information file from scratch.

Merging Spreadsheets

You can merge two spreadsheets within Partek (Figure 4. 12).

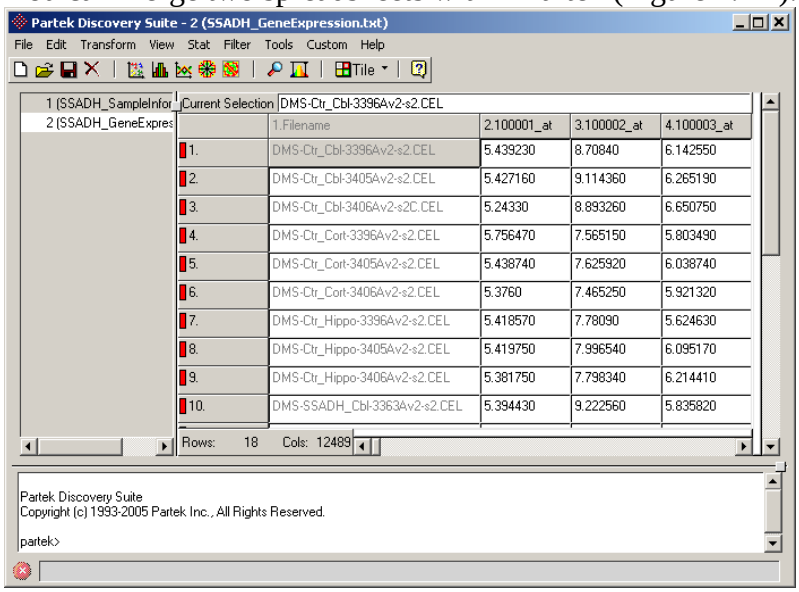


Figure 4. 12: Opening two different files in the Analytical Spreadsheet

- Click **File** > **Merge Spreadsheets** from the Partek main menu to open the *Merge Spreadsheet* dialog, a dialog similar to Figure 4. 13 will appear

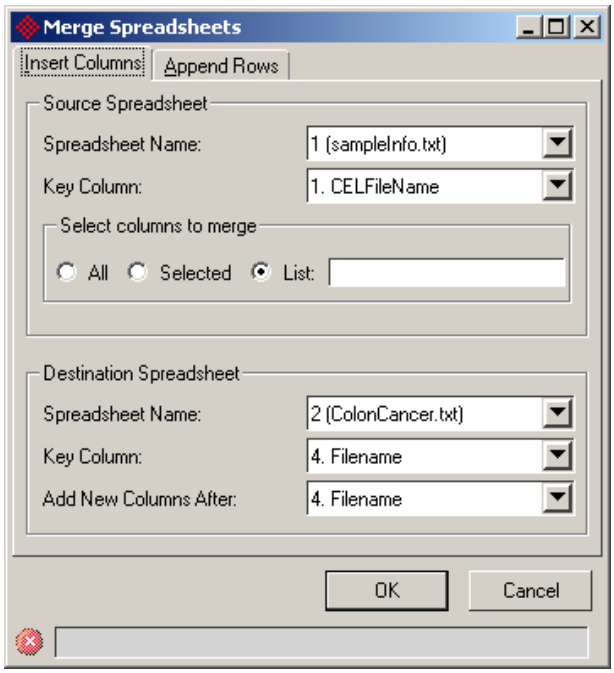


Figure 4. 13: Configuring the Merge Spreadsheets dialog

- If you want to add more columns in the destination spreadsheet, use the default tab **Insert Columns. Note:** The two spreadsheets must have a common *key* column to be able to merge the spreadsheets.
- Choose the *Source Spreadsheet* from which to copy information, from the *Spreadsheet Name* drop-down list (Figure 4. 14). To be more efficient, choose the smaller spreadsheet as the source
- The *Key Column* is the unique ID of each row in the spreadsheet; the values in this column should match the values in the *Key Column* of the *Destination Spreadsheet*. Keys are case-insensitive
- Finally, **Select columns to merge** in the *Source Spreadsheet* to copy to the *Destination Spreadsheet*

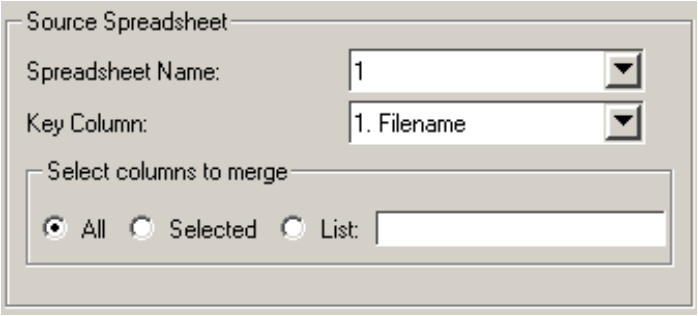


Figure 4. 14: Configuring the Source Spreadsheet dialog

- To configure the *Destination Spreadsheet*, choose the *Spreadsheet Name* and **Key Column** from the drop-down list
- Specify where to add the new information from the *Add New Column After* drop-down list (Figure 4. 15)

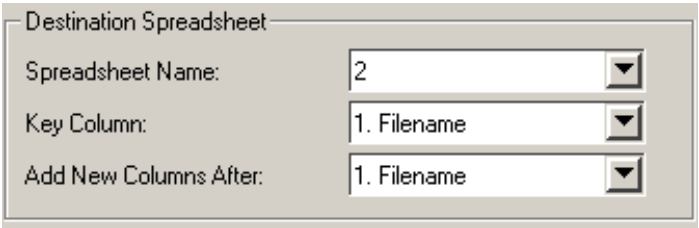


Figure 4. 15: Configuring the Destination Spreadsheet dialog

- Click **OK**

This will copy the sample information columns from spreadsheet 1 and insert them into spreadsheet 2. The order of the samples in the spreadsheet does not matter. The number of samples can be different in those two files.

- If you want to add more rows in the destination spreadsheet, select the tab **Append Rows** tab from the *Merge Spreadsheets* dialog **(**Figure 4. 16**)**

- Choose the *Source Spreadsheet* and *Destination Spreadsheet* from the *Spreadsheet Name* drop-down list. To be more efficient, choose the smaller spreadsheet as the source spreadsheet
- Click **OK**

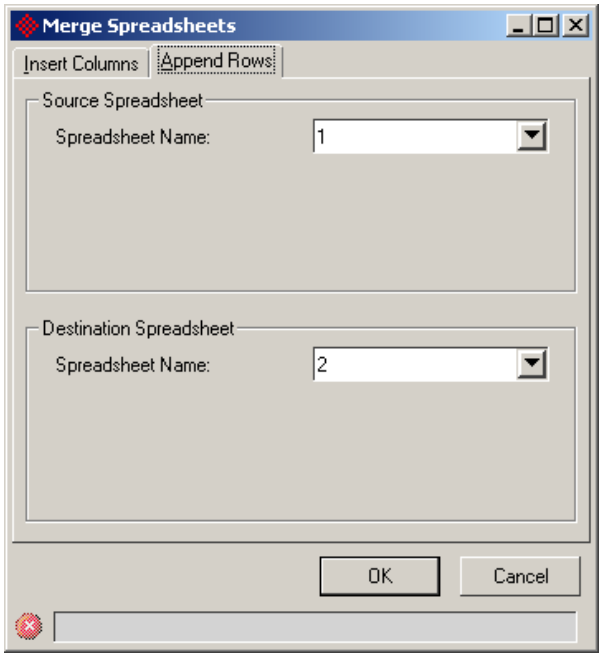


Figure 4. 16: Adding more rows dialog

Note: Appending rows requires the source and destination spreadsheets to have the same number of columns.

Genomic Specific Files

Introduction

Partek® Genomics Suite™ can import two-color microarray data, Affymetrix ARR CEL, CNT, CHP, EXP Files, Experiment Results Summary, and data from the NCBI GEO database.

Importing Affymetrix CEL Files

Partek can load Affymetrix CEL files using a variety of methods, like RMA (Robust Multi-chip Average), GC content adjustment, probe sequence adjustment, fragment length correction, etc.

Selecting the File to Import

- To import Affymetrix® CEL files, select **File > Import > Affymetrix Files... > CEL Files…**from the Partek main window (Figure 4. 19)

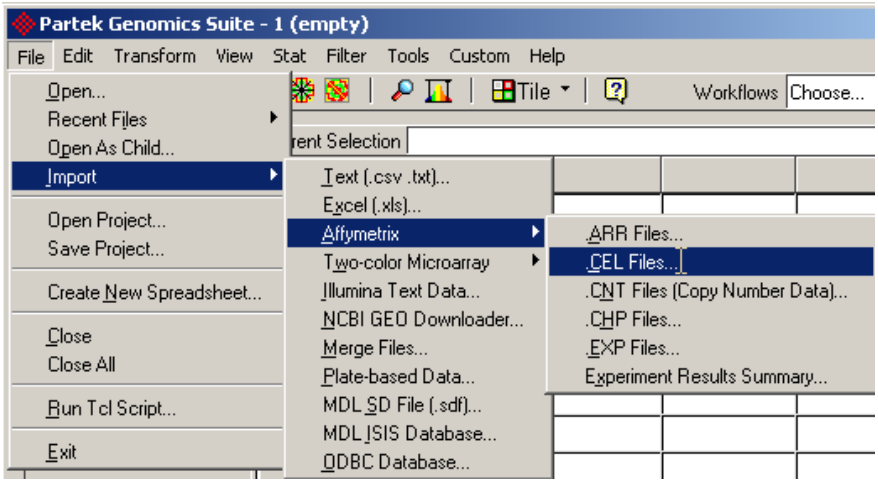


Figure 4. 17: The File > Import > Affymetrix > CEL Files menu item

- Select the CEL files to process (Figure 4. 18)

CEL File Selection

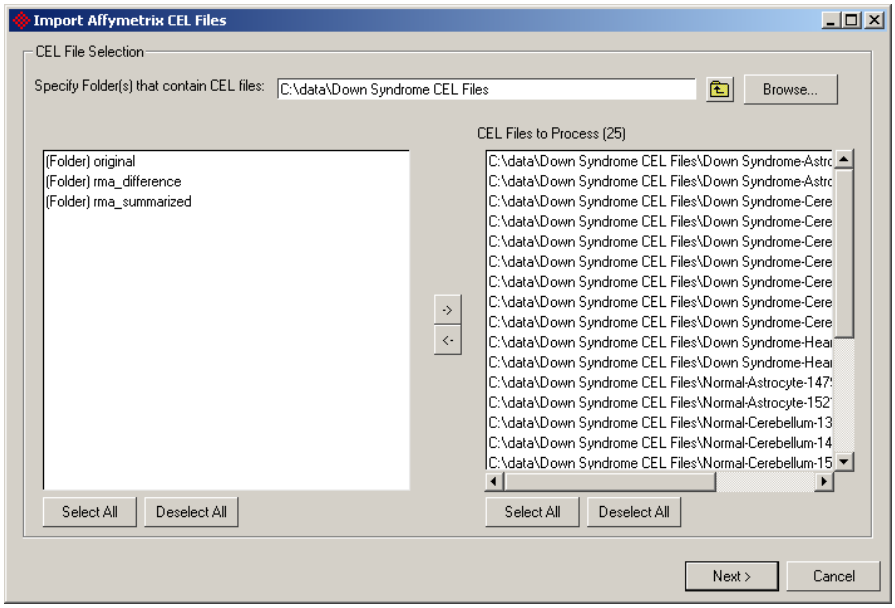


Figure 4. 18: Selecting the CEL files

From the *Import Affymetrix CEL Files* dialog, browse to the folder that contains the CEL files that will be used. By default, when moving to a new directory, all CEL files will be selected but not chosen. To enter a new directory, either enter the address in the *Address* bar, or double click on a directory. To go up to the parent directory, click on the up directory button ( ).

Configuring the Additional Files

- Click the **Next** > button, a dialog similar to Figure 4. 19 will appear

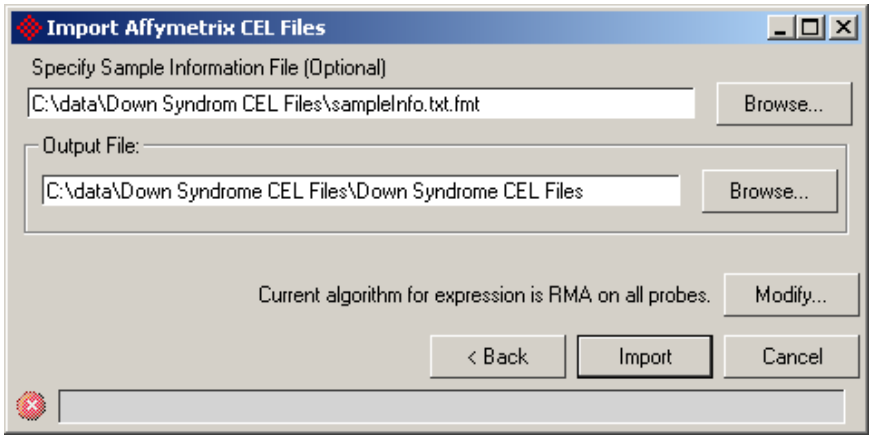


Figure 4. 19: Specifying the Sample Information File and Output File

Specify Sample Information File is optional if you are importing only one chip type. If Partek discovered a file in the same directory as the CEL files, then that file will automatically be used as the sample information file for the import. If that is not the correct file, it can be changed. Directions for how to create and use Sample Information file are described in *Example: Creating a Sample Information File* section above.

- Click **Import** to start the import process (Figure 4. 19)

You may be asked to **Specify the Library File Root Folder** (Figure 4. 20). You can select to use your current Affymetrix library folder if you have already installed software from Affymetrix like Expression Console™, Genotyping Console™, etc.

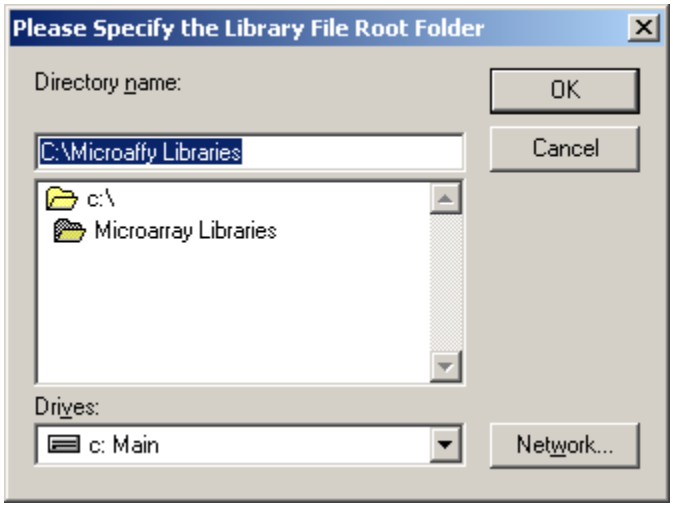


Figure 4. 20: Specifying the default library root folder

When the process successfully completes, the results are loaded into a Partek spreadsheet (Figure 4. 21).

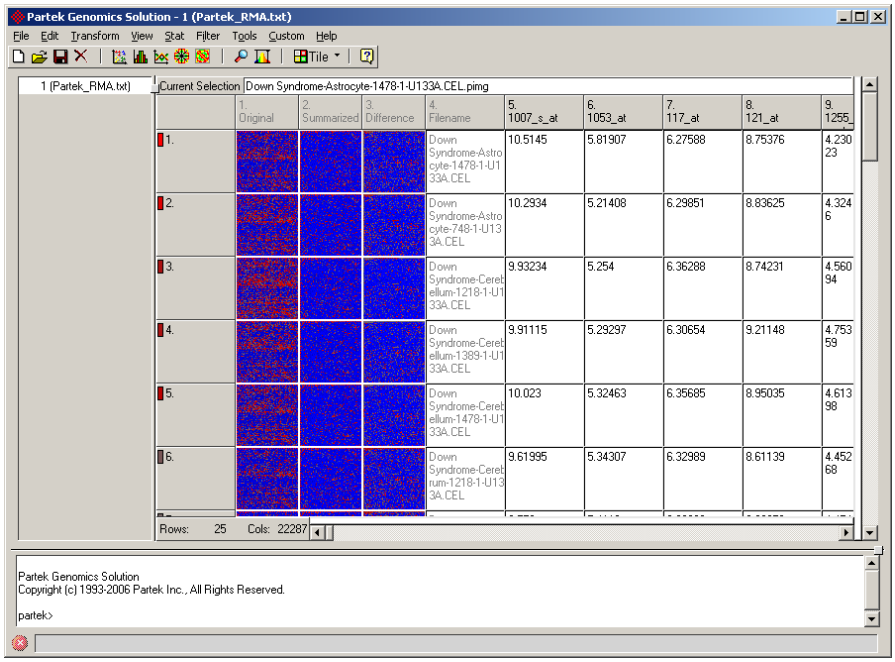


Figure 4. 21: Viewing the RMA results after they are loaded into a Partek spreadsheet

Configuring the Options for the Modify Button

The options in the *Modify* button from the *Importing Affymetrix CEL Files* dialog can be configured to suite the importing process. Settings for handling the algorithm and file output are available. The default settings will be in place for the chip type being imported if no import settings were changed.

Configuring the Algorithms Panel

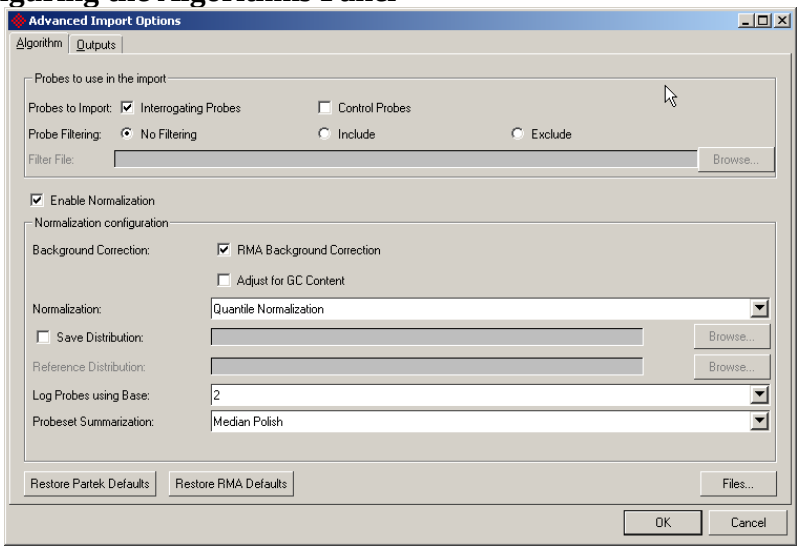


Figure 4. 22: Configuring the Algorithms panel

In the *Probes to Import* option, *Interrogating Probes* and/or the *Control Probes* can be selected. Interrogating Probes are probes that target on certain locations on a

genome. *Control Probes* can be checkerboard QC probes on 3'IVT arrays, genomic background probes on Exon arrays, etc. Please refer to Affymetrix array support web page on the types of *Control Probes* of a particular array.

For *Probe Filtering*, you can choose to have *No filtering*, which is the default setting, or you can specify a list file and choose to include or exclude probes/probe sets in the list file.

File Formats for Filtering

- Affymetrix GCOS® .MSK files can also be used as the list file. Please refer to GCOS' help on how to create probe set mask and probe mask files
- Exon meta-probeset annotation files can be used to filter probes or probesets
- You can create your own list file. The format is a text file. Each line represents a probe set. In a probe set line, you can further specify individual probes. The probe number starts from 1

NOTE: If you are going to import probe level data (*No Summarization*), the imported probe indices will be re-numbered starting from 1. Here is an example list file that will be used to exclude probe sets and probes:

```
#Comment: this list file will be used to exclude probes and probe sets
#Probes and probe sets that are in this file will be excluded, in other words,
#Probes and probe sets that are NOT in this file will be kept
#
#Exclude the whole AFFX-BioB-3_at probe set
AFFX-BioB-3_at
#Exclude the first probe of 222384_at
#A single probe can be separated by '.' from its probe set
222384_at.1
#Exclude the last probe of 222384_at
#Use tab or ' ' as delimiter
200000_s_at 11
#Exclude several probes of 202495_at (probe 5 and 6 will be kept)
#Format 1 (use ' ' as delimiter):
202495_at 1 2 3 4 7 8 9 10 11
#Exclude several probes of 202495_at (probe 5 and 6 will be kept)
#Format 2 (use , as delimiter, GCOS .MSK compatible):
202495_at 1,2,3,4,7,8,9,10,11
#Exclude several probes of 202495_at (probe 5 and 6 will be kept)
#Format 3 (also GCOS .MSK compatible):
202495_at 1-4,7-11
```

Here is another example list file that will be used to include probe sets and probes:

```
#Comment: this list file will be used to include probes and probe sets
#Probes and probe sets that are in this file will be kept, in other words,
#Probes and probe sets that are NOT in this file will be excluded
#
#Include the whole AFFX-r2-P1-cre-5_at probe set
AFFX-r2-P1-cre-5_at
#Include only the first probe of 200000_s_at
#A single probe can be separated by '.' from its probe set
200000_s_at.1
#Include only the probe 5 of 222384_at
```

```
#Use tab or ‘ ’ as delimiter
222384_at 5
#Include several probes of 202495_at
#Format 1 (use tab or ‘ ‘ as delimiter)
202495_at 5 6
#Include several probes of 202495_at
#Format 2 (use , as delimiter, GCOS .MSK compatible)
202495_at 5,6
#Include several probes of 202495_at
#Format 3 (also GCOS .MSK compatible)
202495_at 5-6
```

Background Correction

RMA

For Background Correction, Partek can perform the RMA background correction, and/or adjust probe intensities for a number of properties such as Fragment length, GC content, and Sequence allele position.

Probe-level intensity is known to be significantly dependent on the GC content of the sequence. Partek's GC adjustment uses a model fit on all imported probes to remove the effects of GC content on probe-level intensities. Interrogating probes and control probes are both used and treated identically during fitting and adjustment. This procedure is performed before any other background correction, or normalization, such as RMA background correction or quantile normalization.

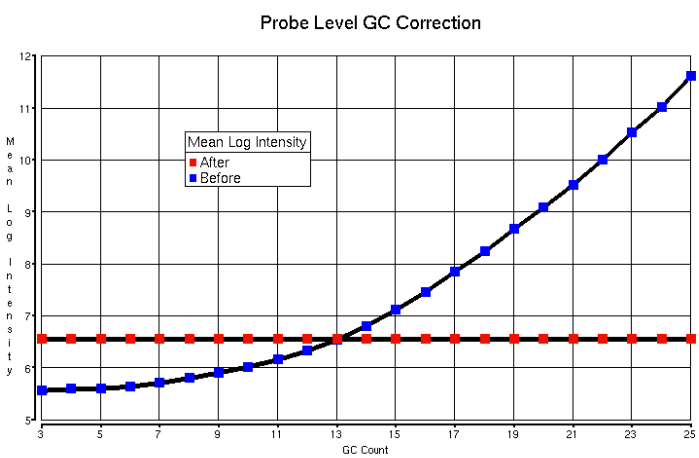


Figure 4. 23: Mean log probe intensities before and after GC correction for a Human Exon chip

Figure 4. 23 shows the mean of log probe intensities before and after GC adjustment plotted against the GC count.

Fragment length and sequence correction are performed in a similar fashion, adjusting for any intensity correlations. Sequence correction adjusts for all effects of GC correction, so it is not necessary to perform both.

After importing probe intensities, samples are normalized by scaling the samples to have the same overall intensity. The probes are then summarized using the target allele information to account for probes within the same SNP probe set targeting different sequences. Creating copy number from the summarized intensities is accomplished by normalizing each sample to the reference—either paired references or a pooled reference depending on paired or unpaired workflow.

For *Normalization*, you can choose to perform *No normalization, Quantile normalization, Normalize to reference distribution, Shift to reference Median, or Scale to reference Median, or Normalize to Reference Distribution*. If the RMA algorithm is going to run several times on difference batches of chips, it's critical to quantile normalize all batches to the same distribution, so signals from any 2 chips are comparable. In Partek, you can first select *Quantile normalization* and *Save reference distribution* on one batch then *Normalize to (*the saved*) reference distribution* on other batches.

When logging the data, you can choose or type the base.

In *Summarization*, Partek allows you to use different types of summarization techniques to compile the data of a probe set down to a single number that represents a central tendency for that probe set. The different algorithms for probe set summarization are arranged from most conservative to least conservative, with respect to their statistical efficiency. For a description of the different summarization algorithms used here, see **Chapter 9 Descriptive Statistics, Correlation, and Measures of Similarity and Dissimilarity**.

If you choose to have *No summarization*, you will import the probe level data. NOTE: If you've specified to filter probes based on a list file (*Step 1*), and to *import without summarization* then the imported probe indices will be re-numbered starting from 1.

Depending on the array type, Partek recommends different algorithm settings. Selecting the *Restore Partek Defaults* button will restore the pre-defined settings Selecting the *Restore RMA Defaults* button will set the current algorithm to RMA.

The Partek implementation of RMA is tuned for speed and decreased memory usage. There are four steps involved in the RMA importing method; only PM values are used in this method:

- Background correction on the PM values
- Quantile normalization across all the chips in the experiment
- Log2 transformation. Note: the log is base 2, and if the input value <= 0 the transformed value will be marked as missing
- Median polish summarization. Note: Median polish might give the same summarized values for all/most samples if your sample size is very small. For more information, please go to:

https://stat.ethz.ch/pipermail/bioconductor/2003-September/002498.html

The chapter **References** section has more material on the RMA algorithm.

GCRMA

The Partek implementation of GCRMA uses GCRMA background correction and then the same normalization (quantile Normalization) and summarization (Median Polish) methods as RMA to convert background corrected data into expression measures. There are four steps involved in the GCRMA importing method; both PM and MM values are used in this method:

- GCRMA Background correction on the PM values by fitting a loess curve through MM values ~ MM affinities
- Quantile normalization across all the chips in the experiment
- Log2 transformation. Note: the log is base 2, and if the input value <= 0 the transformed value will be marked as missing
- Median polish summarization. Note: Median polish might give the same summarized values for all/most samples if your sample size is very small. For more information, please go to: https://stat.ethz.ch/pipermail/bioconductor/2003-September/002498.html

The chapter **References** section has more material on the GCRMA algorithm.

NOTE: GCRMA only works for Gene Expression but not Exon, Tiling and SNP, because GCRMA needs MM values.

Output Options

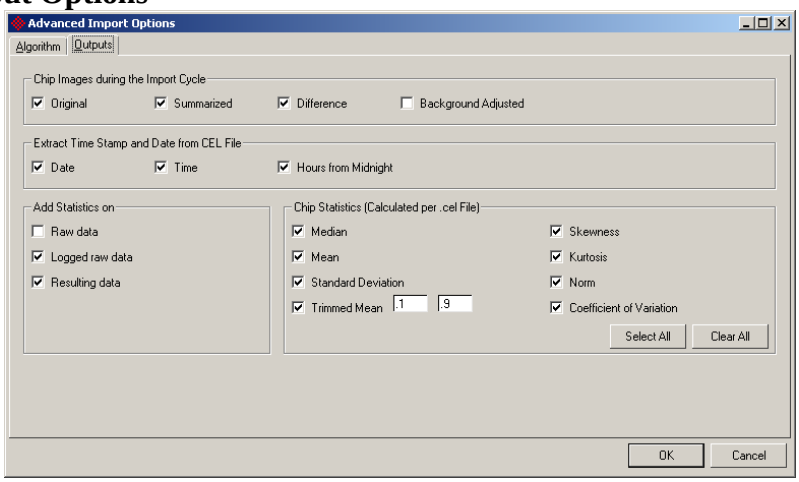


Figure 4. 24: Choosing Outputs

Partek can generate full-sized original, summarized, difference, and background corrected images, along with thumbnail images that are used in the resulting spreadsheet.

By default, the images will be stored in subfolders in the results file folder. When the default folder for image files is used, Partek remembers the relative path between the CEL data and images. Thus, the folder containing the results and images can be copied to a new location and the image links will still work.

In the *Date and Time* section, specify how to convert date/time fields such as hybridize date and scan date. As an example, the format of *Date* is **Apr 16 2004**, and the format of *Time* is **09:34:00 AM**. *Hours from Midnight* is the time one experiment happened relative to 12:00AM of that day.

In the *Statistics (Calculated per .CEL File*) section, you can perform statistical analysis on the raw data both before and after logging the data and before and after the summarization of the data. The different options for the statistical calculations are described in **Chapter 9 Descriptive Statistics, Correlation, & Measures of Similarity & Dissimilarity**. This is done to each .CEL file separately.

Linking to Affymetrix Annotation Files

Linking to Affymetrix annotation files allows views of a probe set's information, as included in an Affymetrix annotation file, on demand. It is created automatically by the import process.

To invoke the link, right-click on the column header of the desired probe set and select **Probe Set Details** (Figure 4. 25). The probe set's information appears in a new dialog (Figure 4. 26) that includes useful links to various relevant websites.

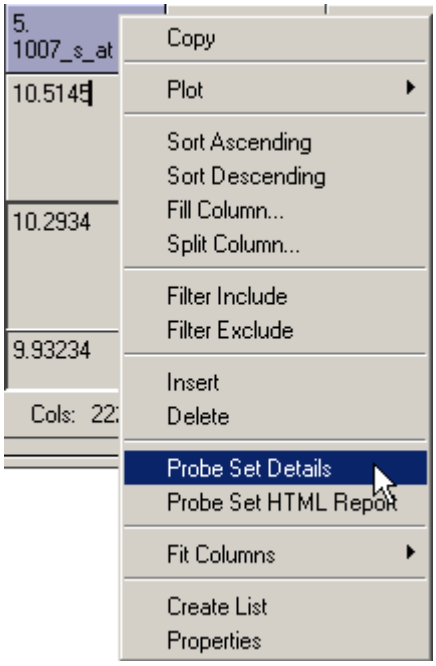


Figure 4. 25: Invoking the AffyInfo link

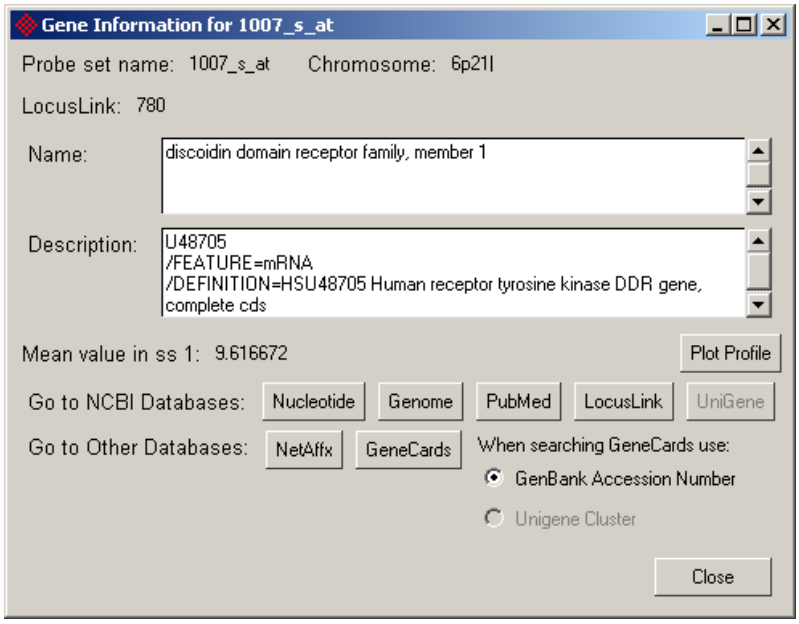


Figure 4. 26: Displaying the probe set's gene information

Viewing Images

Images can be added to the final spreadsheet by checking the desired buttons in the **Image Selection** panel of the dialog. Images will be stored in *.pimg* files under the CEL file directory.

Figure 4. 27 shows an example of the RMA resultant spreadsheet with thumbnail images. Here the *Original image* represents the raw probe-level expression values on the chip; the *Summarized* image corresponds to the RMA results; and the *Difference* image can be considered as *Summarized – Original.*

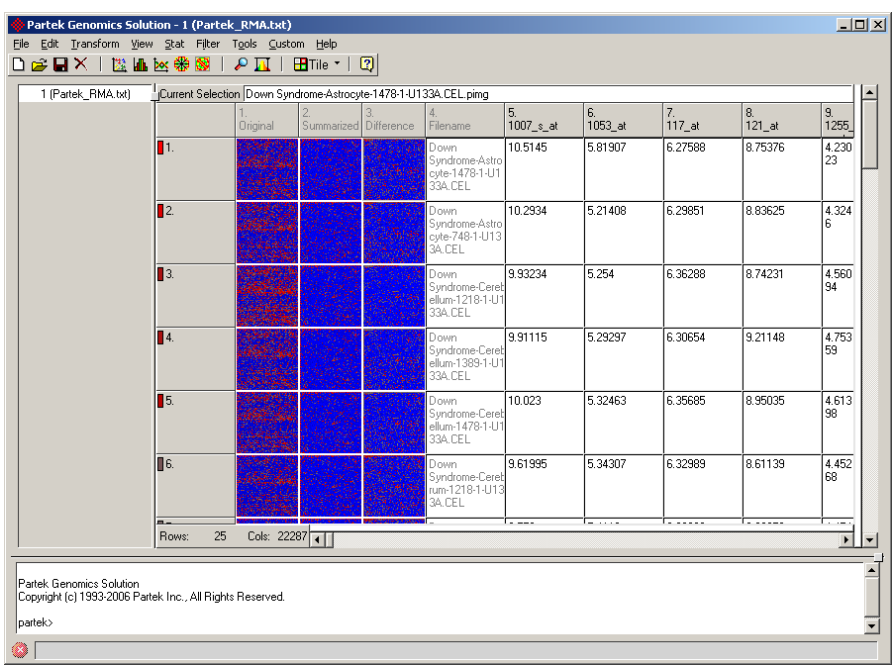


Figure 4. 27: Viewing the RMA results with thumbnail images after they have been loaded into a Partek spreadsheet

- Double-click on a thumbnail image to view its full sized image (Figure 4. 28)

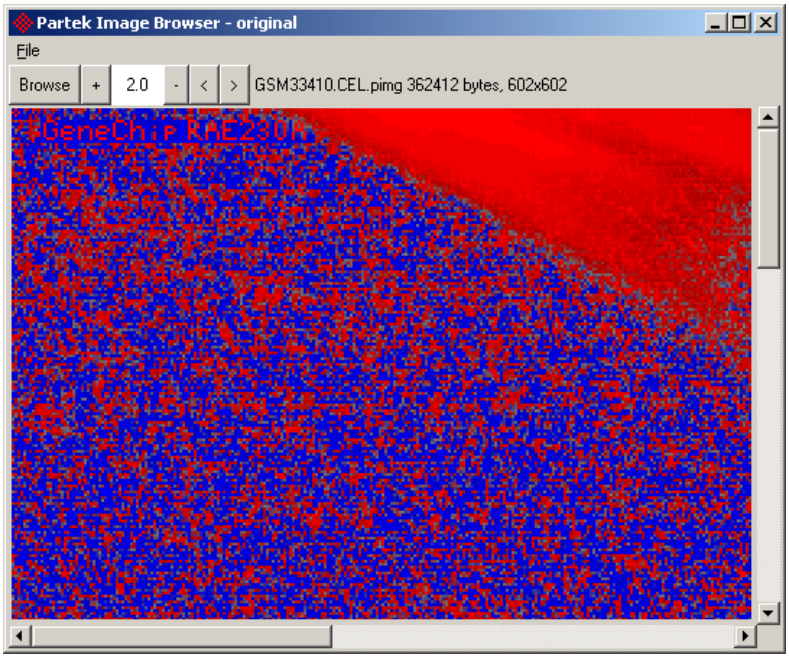


Figure 4. 28: Viewing the full sized image shown in the Partek® Image Browser ™

The *Browse* button of the *Partek Image Browser* (Figure 4. 28) is used to change to another folder and view the image there. The + and – buttons are used to zoom in and out, respectively, and the < and > buttons are used to show the previous or next image in the same folder.

There is a limit to how much you can zoom in on an image. Using the *Preferences* dialog (invoked from **Edit > Preferences > Other Settings**), the *Maximum image size (pixels)* value can be changed to a size (e.g. 1024 x 1024 = 1048576) that fits a computer's screen and memory size (Figure 4. 29).

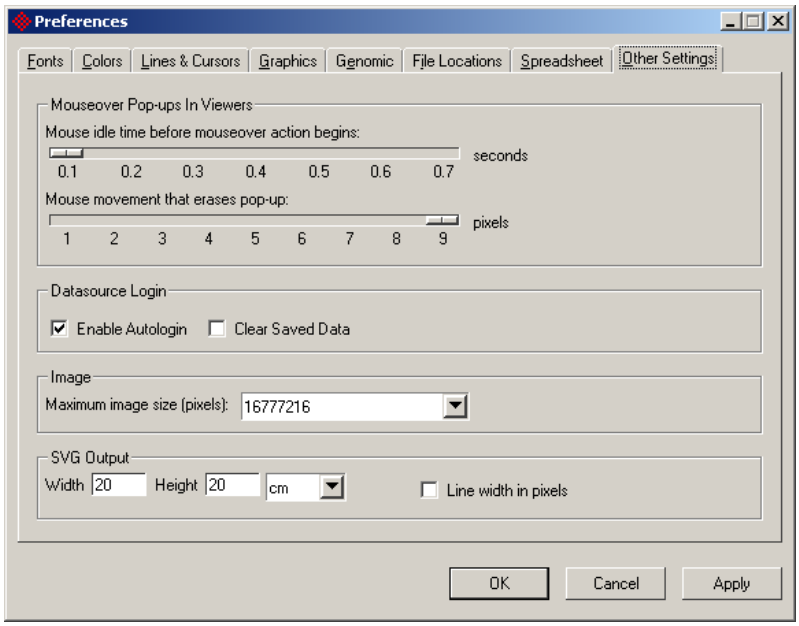


Figure 4. 29: Selecting the Maximum image size the Preferences dialog

On the *Spreadsheet* tab of the same *Preferences* dialog, the *Maximum Number of Images Per Spreadsheet* value can be changed based on screen and memory size (Figure 4. 30).

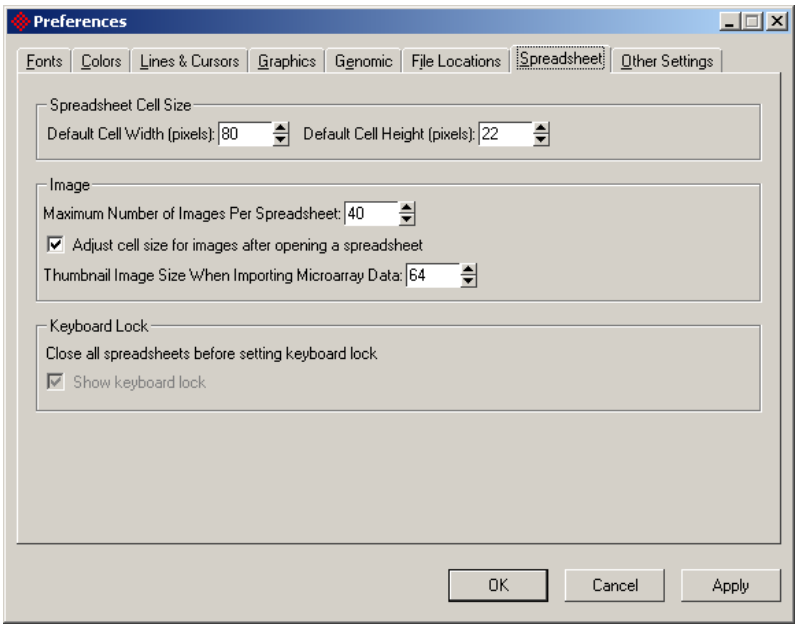


Figure 4. 30: Selecting the Maximum number of images per spreadsheet

When images are saved under the same folder as the data, Partek uses a relative path so copying the entire experiment folder to another location will not break the image links. Using a different folder could break the image links after moving the tree. In that case, go to **Tools > External Link Manager...** from the main Partek window, select **Original Image Viewer** (top of Figure 4. 31), and click **Edit...**. In

the pop-up dialog that appears (middle of Figure 4. 31), you can click **Browse...** and select the location of the original image folder (bottom of Figure 4. 31). The same steps can be repeated for *Corrected Image Viewer* and *Residual Image Viewer* in the *External Link Manager* dialog (top of Figure 4. 31).

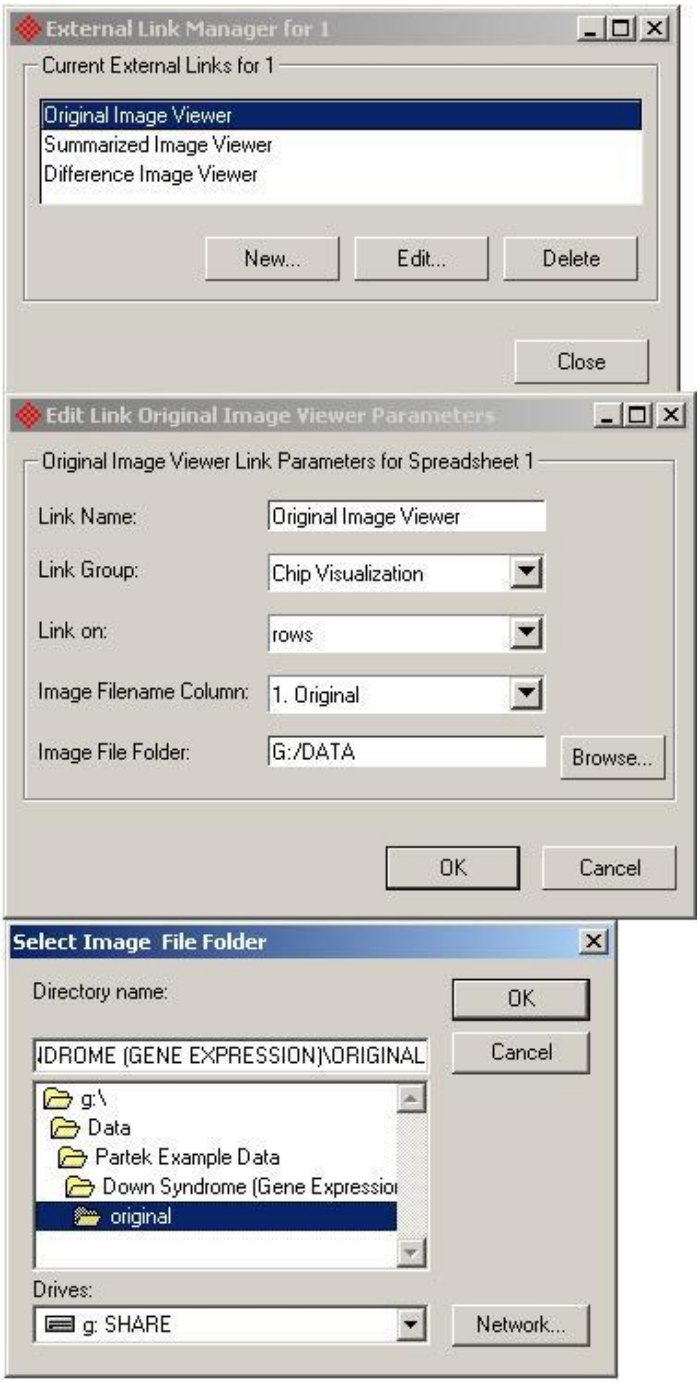


Figure 4. 31: Fixing a broken image link

Configuring the Image Color Map

Partek chip images use the current color palette to color the false image of the .CEL file. You may create a new color palette using the *Color Palette Manager* from the **Tools > Color Palette Manager...** menu item of the Partek main window. Within the *Continuous* tab of the dialog (Figure 4. 32), click the **Create New** button.

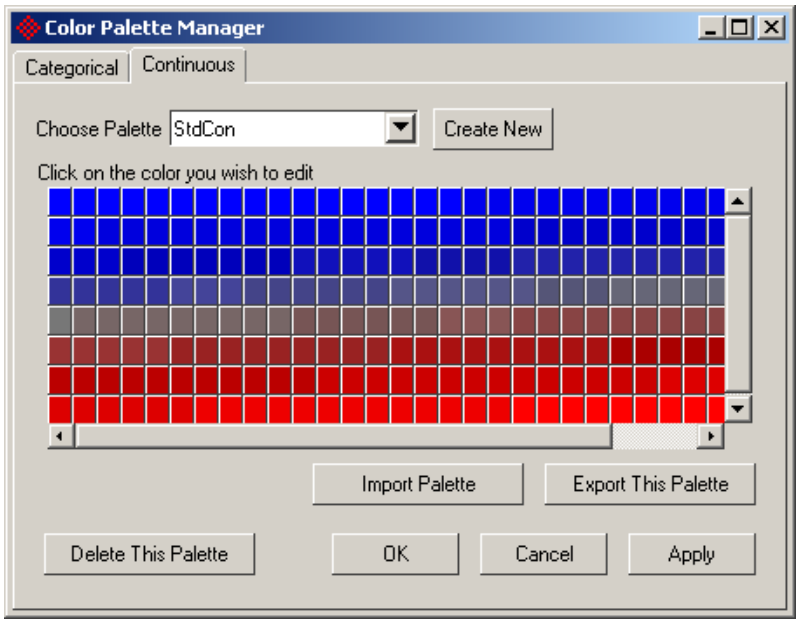


Figure 4. 32: The Color Palette Manager

- Specify a name for the new palette (e.g. Blue-White-Red), specify the *Palette Size* as **256**, *Interpolation Points* as **3**, and then choose colors for the three bars (Figure 4. 33)
- Click **Create New Palette**

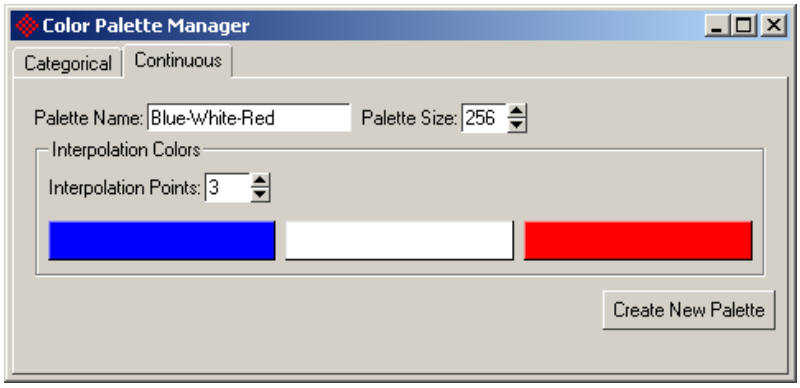


Figure 4. 33: Creating a new palette

- Click **OK** (Figure 4. 34) to make the newly created color map the default palette

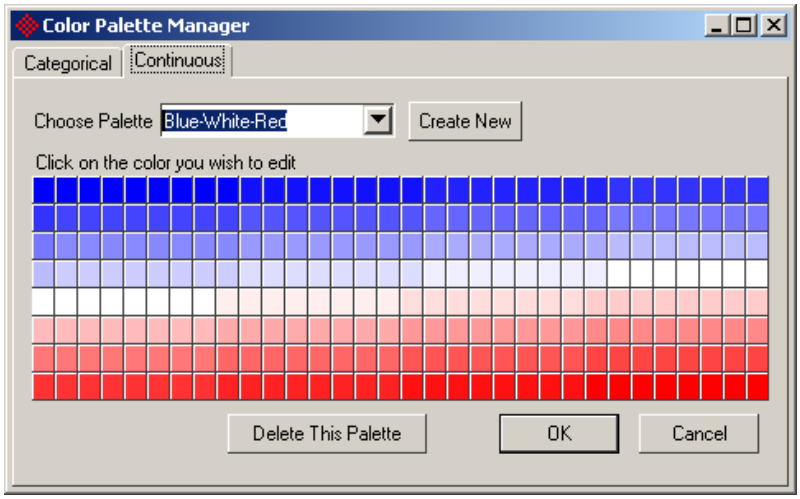


Figure 4. 34: New color palette

Note: You will need to save, close, and reopen the spreadsheet for the new color palette to become effective.

Importing Affymetrix CHP Files

Introduction

This document describes how Partek imports Affymetrix CHP files. Partek handles the three main formats of these files: the text format, the binary XDA format, and the new AGCC format. Partek also handles two kinds of CHP results: the expression results and the genotyping results.

The import process takes one or more CHP files and creates one final Partek spreadsheet, where each row represents a file and each column represents a probe set. The spreadsheet's data values consist of gene expression analysis results, as extracted from the files.

All CHP files are expected to be of the same Affymetrix array type or the import process will generate an error.

Importing CHP Files

The import dialog is invoked from **File** > **Import** > **Affymetrix** > **.CHP Files**.

Selecting CHP Files

Figure 4. 35 shows the *File Selection* panel of the import dialog.

- Click the **Browse…** button to specify the CHP file folder

All files found in the folder with a .CHP extension (not case-sensitive) are selected for import by default.

- Click the -> button to select the highlighted .CHP files
- Click **Next >**

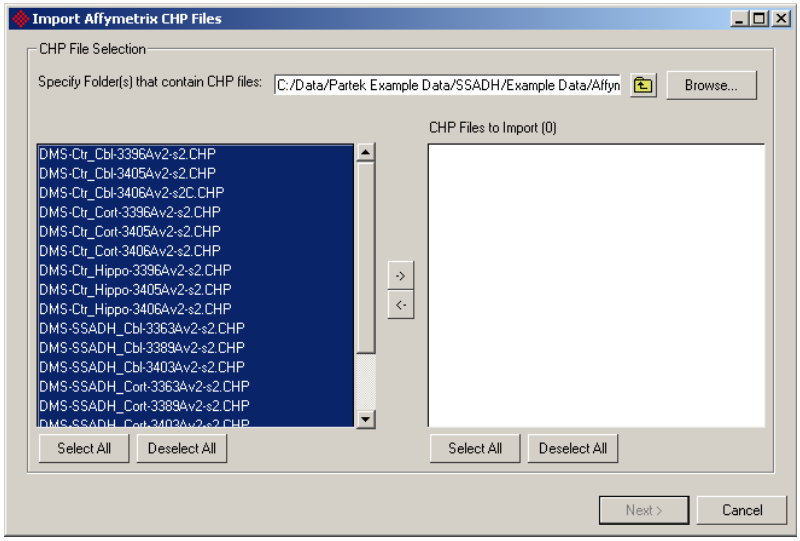


Figure 4. 35: Selecting CHP files

Selecting Files and Import Type

Partek will look for a file called *SampleInfo.txt.* If your sample information has a different name, you can select **Browse** to specify it (Figure 4. 36).

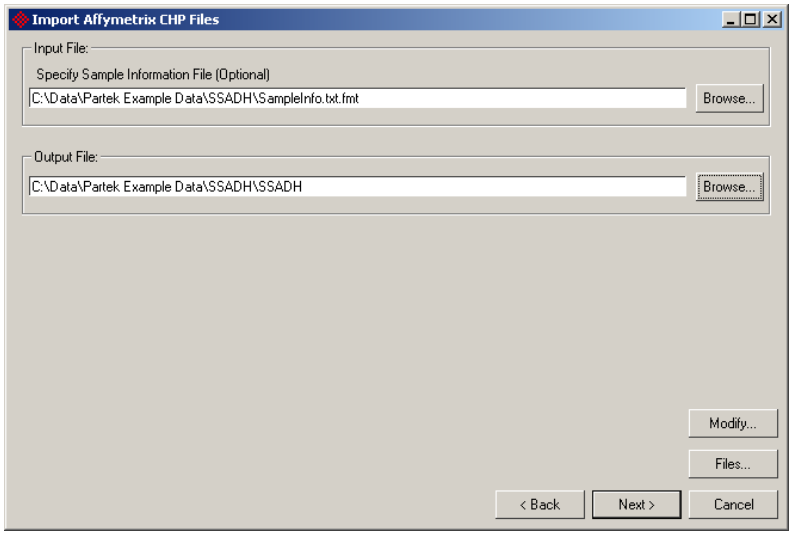


Figure 4. 36: Selecting Folders, Files, and Import Type

Expression Array CHP Probe Set Absent/Present Detection and Statistics

If the CHP files are expression arrays, you can click the *Modify…* button to specify the import criteria based on Absent/Present calls. Note: for genotyping CHP files, the *Modify* button will be disabled. In the *Probe Sets to Import* panel, you can choose to import *All probe sets* or *Only probe sets* that are *present*, *marginal*, *absent*, or any combination thereof on at least a certain number of chips. You can

Calculate Probe Set Statistics on different groups of probe sets, e.g. only on probe sets that are present on all chips, when importing CHP files (Figure 4. 37).

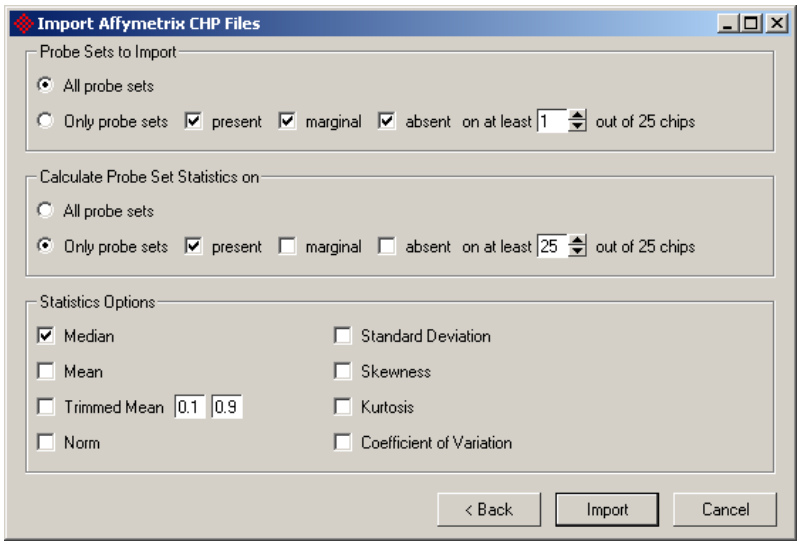


Figure 4. 37: Configuring the Probe set absent/present detection and statistics dialog

Importing Process

The *Affymetrix CHP Import Status* panel in Figure 4. 38 displays the progress of the CHP file import process. Upon successful completion, the results are loaded into a spreadsheet. Figure 4. 39 shows the imported CHP results for expression arrays.

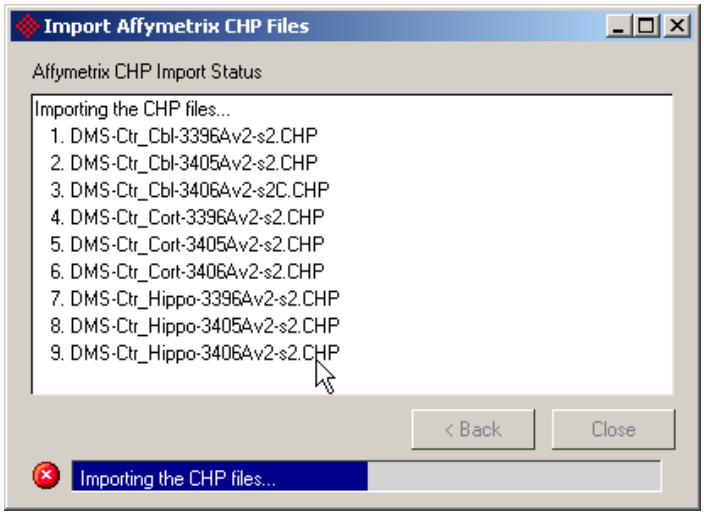


Figure 4. 38: The Import Affymetrix CHP Files progress panel

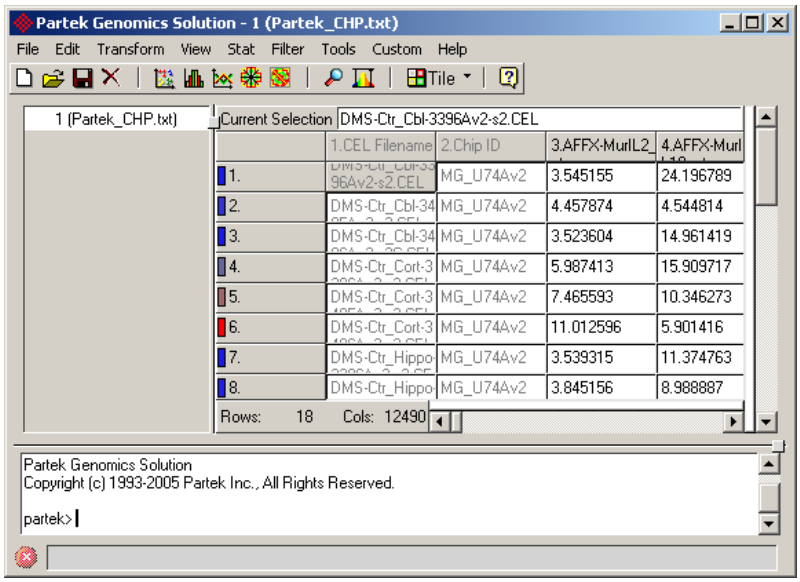


Figure 4. 39: Viewing the imported Affymetrix expression CHP results

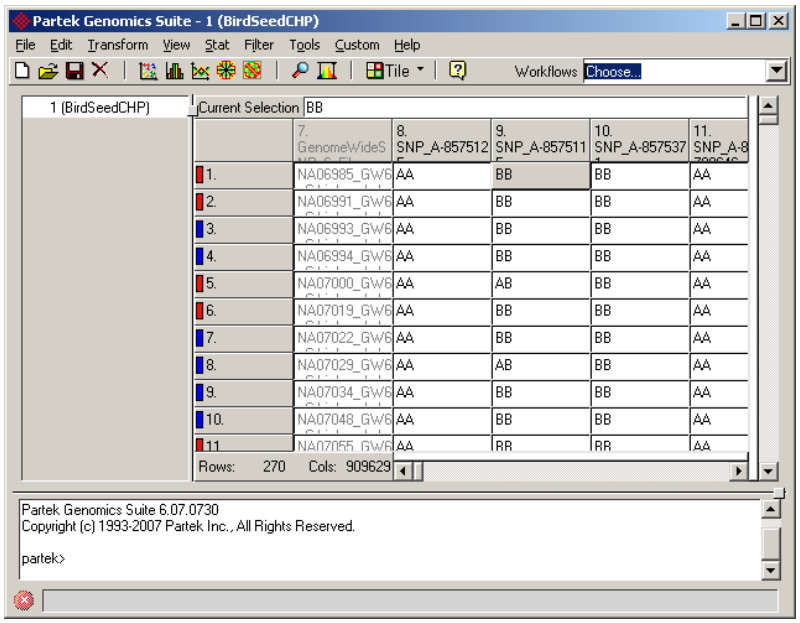


Figure 4. 40 Viewing the imported Affymetrix genotyping CHP results

Importing Agilent Data

This section describes how to import data that is created by Agilent's Feature Extraction Software.

The Import Process

Invoking the Dialog

The process is invoked by selecting the **File -> Two-color Microarray -> Agilent...** menu item (Figure 4. 41).

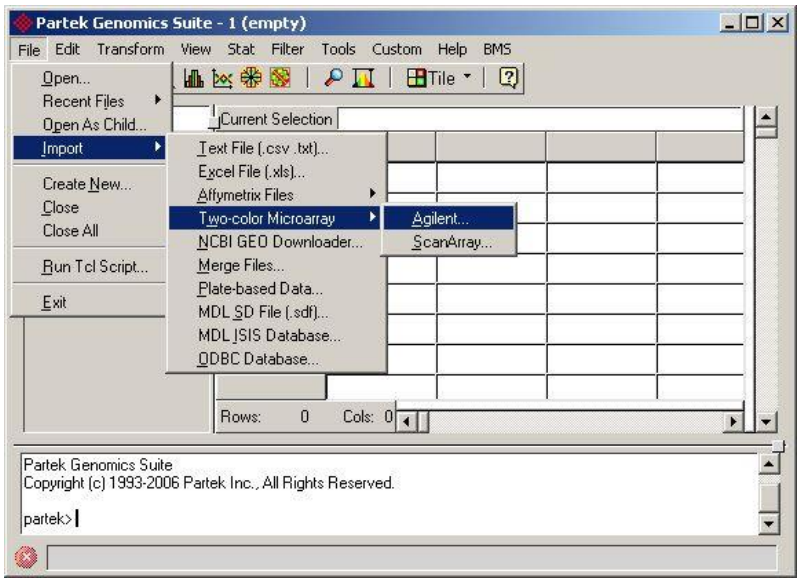


Figure 4. 41: Invoking the Agilent Import dialog

Import Type

The first step is to specify whether paired data (that is, data for both red and green channels) or non-paired data (such as ratios) are to be imported (Figure 4. 42).

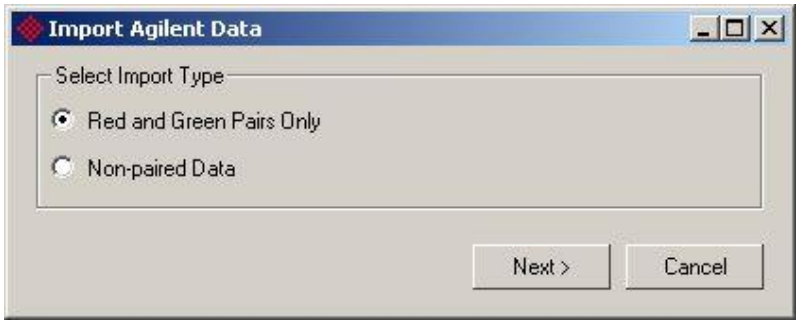


Figure 4. 42: Choosing between paired or non-paired data

File Selection

Next, specify (Figure 4. 43) the folder where the Agilent data resides. Once chosen, the importer automatically sets the *Results File* to *Partek_AgilentData.txt* and in the same folder. If this file already exists, a different one can be chosen or it can be overwritten when prompted.

Specification of an annotation file is optional; however, in order to link probe sets in the final imported spreadsheet to gene annotations in an Agilent annotation file, this value must be specified.

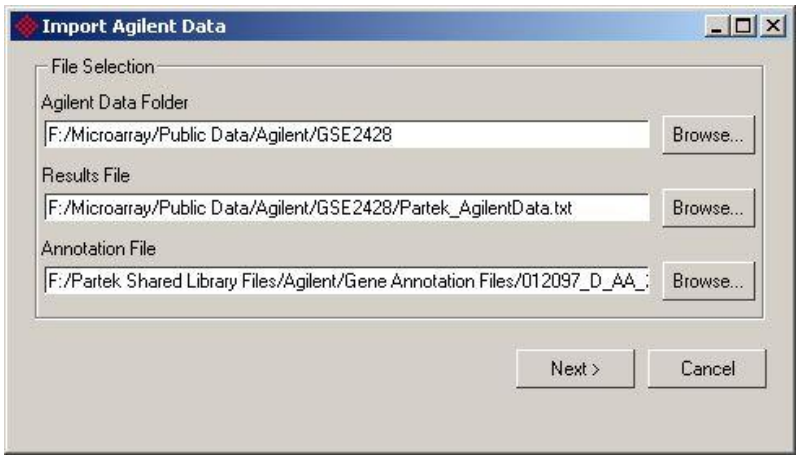


Figure 4. 43: Specifying the location of the data, results, and annotations

The first time this dialog is used, these entries are empty; however, after adding or modifying values, they are stored as Partek preferences and are retained upon subsequent use.

The next phase of the import is selecting the files to import (Figure 4. 44).

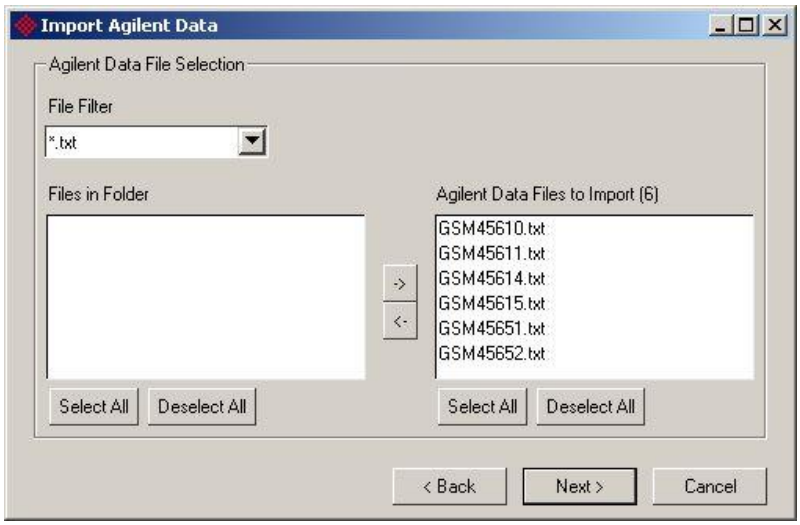


Figure 4. 44: Selecting the files to import

Column Selection

At this point, choose the columns, as found in the FEATURES table of the files created by the *Feature Extraction Software*, to import. If it was specified (in the dialog of Figure 4. 42) to import paired data, then each item in the *All* panel of Figure 4. 45 represents data from two columns – one that corresponds to the red channel, and one that corresponds to the green channel. For example, if *Mean Signals* is chosen, then both the mean signal of the red channel and mean signal of the green channel is imported.

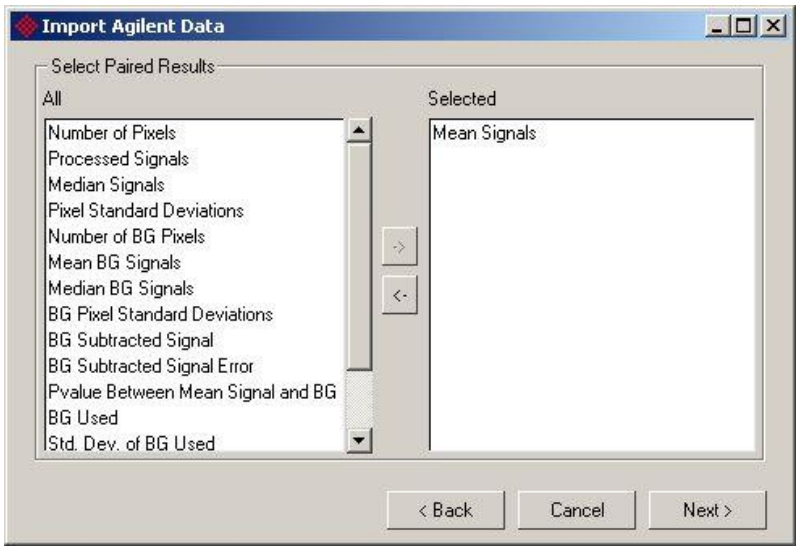


Figure 4. 45: Selecting paired data columns

If non-paired data (such as ratios) are being imported, then each item in the *All* panel of Figure 4. 46 will represent exactly one column to import.

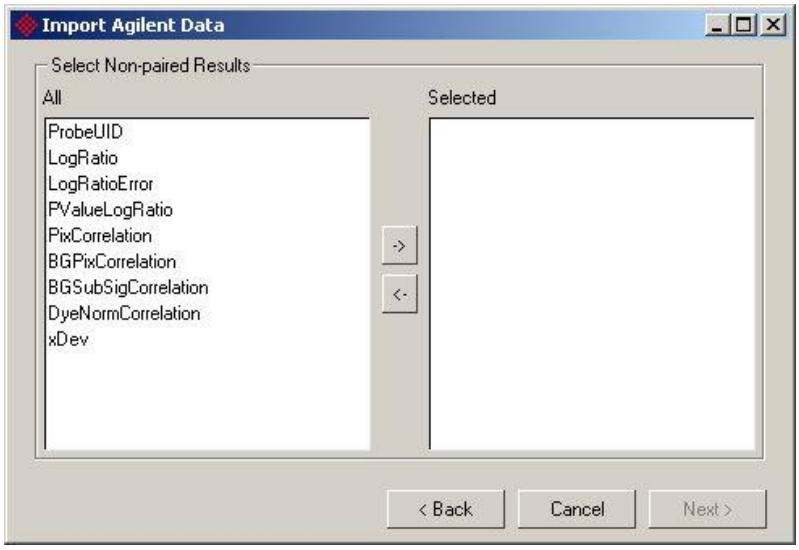


Figure 4. 46: Selecting non-paired data columns

Only one item may be selected for import in the *Import Agilent Data* dialog. Items are chosen by double-clicking, or by highlighting with a single-click and then using the -> and <- buttons.

Loading the Data

Partek will import the data and load it into a spreadsheet (see Figure 4. 48). During this time, the status of the import is displayed (Figure 4. 47).

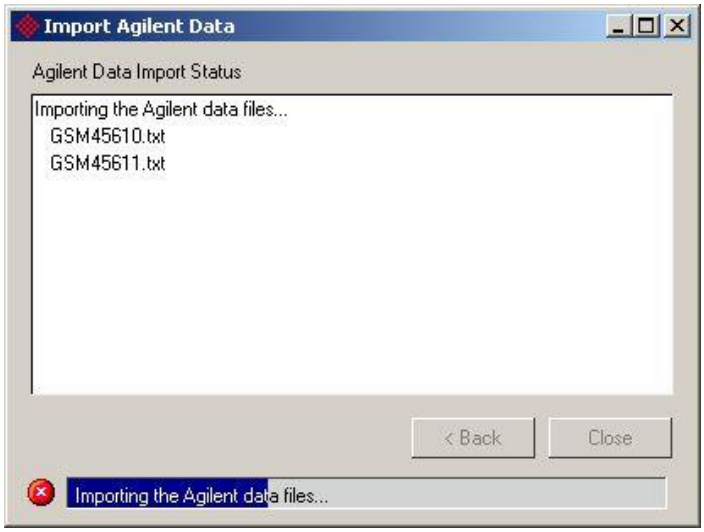


Figure 4. 47: Importing Status

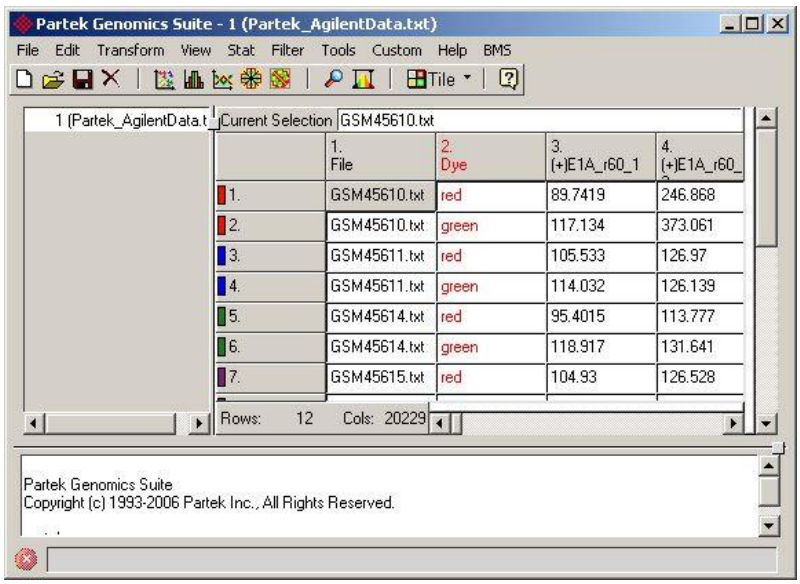


Figure 4. 48: Viewing the final results loaded into the Partek spreadsheet (for paired data)

Annotations

To link a gene with corresponding information in an Agilent annotation file, right-click the desired column header of the spreadsheet, and select **Probe Set Details**. The gene's information appears in a new dialog (Figure 4. 49) that includes useful links to various relevant websites.

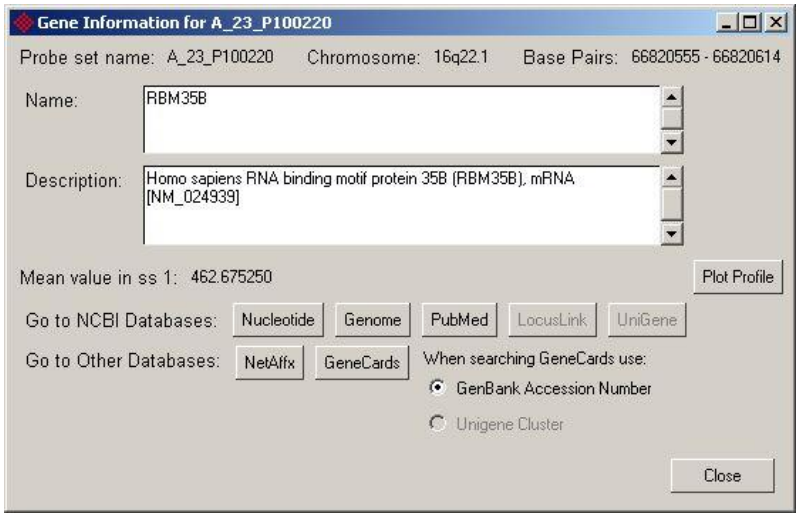


Figure 4. 49: Viewing gene information

Importing from the NCBI GEO Database into Partek

The National Center for Biotechnology Information's *Gene Expression Omnibus* (NCBI GEO) is an on-line database for gene expression information. The *GEO Downloader* can import data from *GEO* into Partek.

Importing from NCBI GEO

In order to download data from Gene Expression Omnibus, you need to know either the GSE number or the GSM numbers of the file.

- Select **File > Import >NCBI GEO Downloader…** from the Partek main menu to open the dialog (Figure 4. 50)

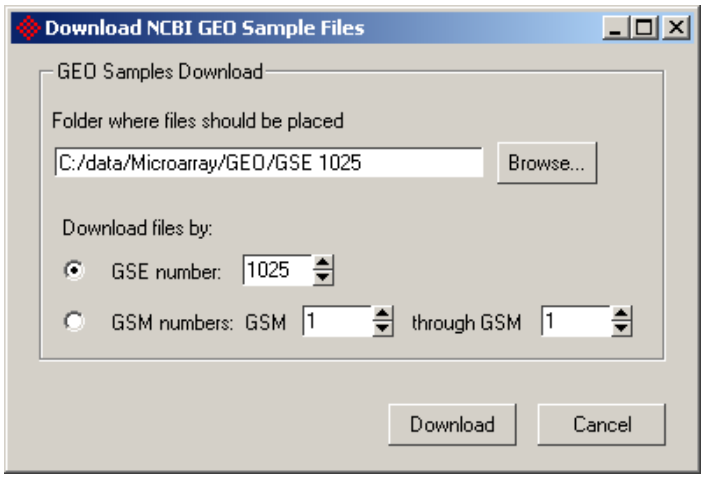


Figure 4. 50: Configuring the Download NCBI GEO dialog

Specify the folder where the files are, either the GSE number or the GSM number, and click **Next**. When it is done, you will be asked if you want to merge the file and import it to Partek.

Merging Files: Keys in Columns/Keys in Rows

- Select **File > Import… > Merge Files…** from the Partek main menu
- Using the *File Merge – Path Selection* dialog (Figure 4. 51), specify the folder where the files to be merged reside and specify the name (full path) of the final merged file. Use the top *Browse…* button to select the folder where the files reside and the bottom *Browse…* button to select the final merged filename
- Using the same dialog, under *File Format Selection*, select **Keys in columns** as the final merged file format

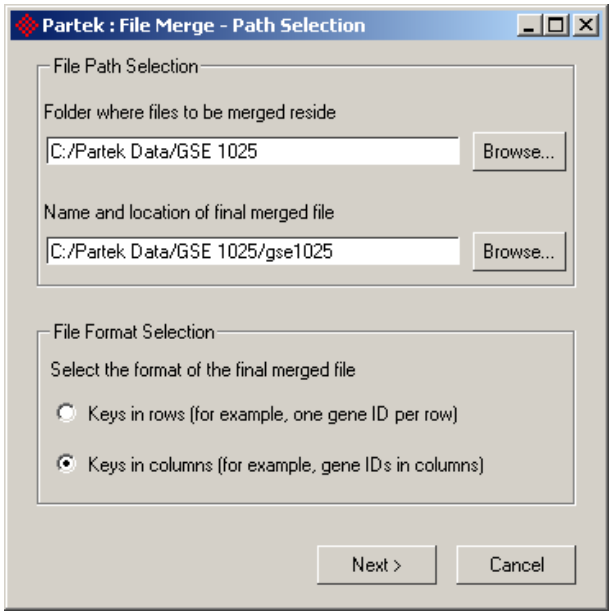


Figure 4. 51: Configuring the File Merge – Path Selection dialog with example folder and file names filled in

- Click **Next >** to advance to the next step of file merge: file selection
- Use the *File Merge – File Selection* dialog (Figure 4. 52) to select the files to be merged

The *File Filter* pull-down list has several common file extensions that can be used to filter the list of files in the folder, or you can type in your own wildcard text string (for example, *.dat). To add files to the list of those to be merged, select the file in the *Files in Folder* list and then click the right arrow button to move them to the *Files to Merge* list. The *Select All* and *Deselect All* buttons will quickly select or deselect items in either list.

- Click **Next >** to advance to the next step of file merge

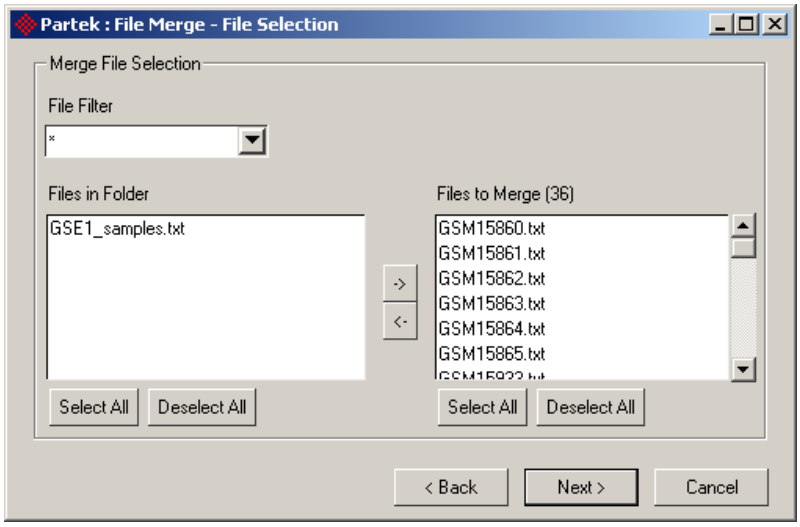


Figure 4. 52: Configuring the File Merge – File Selection dialog with example files selected for merging

Use the *File Merge – Column Selection dialog* (Figure 4. 53) to identify the columns you want included in the final merged file. This is done in the same manner as selecting files in the previous step. Use the right arrow button to add columns to the list of columns to include in the merge

!	File Merge will run faster the fewer columns you select to include in the merged file. Select only those columns that are necessary for the final merged file.

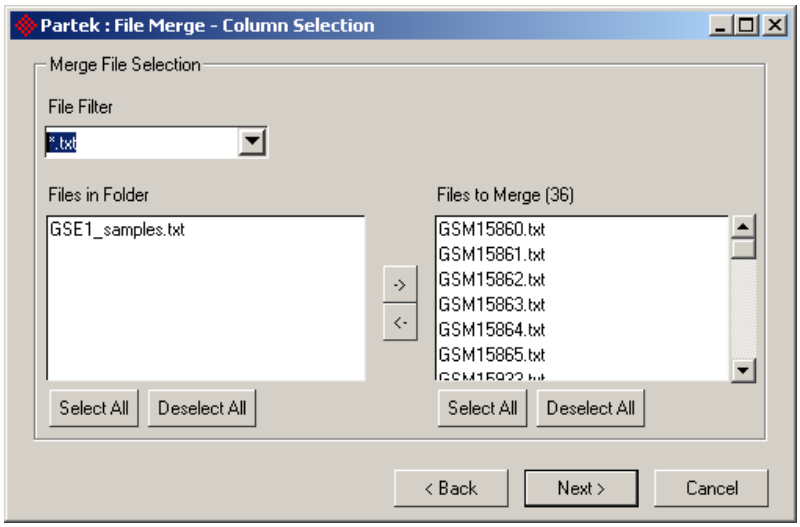


Figure 4. 53: Configuring the File Merge – Column Selection dialog with example columns included in the merge and a key column identified

Use the same dialog (Figure 4. 53) to identify the key column and alternate key columns. The first column in the "Columns used to join files" list is used as the

primary key for the merge. Second and subsequent columns in that list are treated as alternate keys

- Click **Add** to add additional key columns or click **Remove** to remove key columns

!	If you do not add your primary key column to the **Columns to Include in Merge** list, it will be automatically added during the merge process because the merge process needs that column to do the merge. However, any alternate key columns you identify must be explicitly added to the **Columns to Include in Merge** list in order for them to appear in the final merged file.

If the text files to be merged include any special characters or text strings that represent missing values, use the *Missing Data Text Add* button to add them one at a time. Any missing data text you specify will be converted to question marks (the default Partek missing data symbol) in the final merged file.

- Click **Next >** to advance to the next step of the file merge process or click **<Back** if to go back to a previous file merge dialog

Use the *File Merge – Data Types & Duplicates* dialog (Figure 4. 84) to define column types and the handling of duplicate keys, such as gene names and patient IDs. For each column, verify and/or change the type to either text, numeric or nominal. For specific file formats (other than CSV and tab-delimited), the File Merge utility will set reasonable defaults for column types, but each column's type should be verified.

Using the same dialog (Figure 4. 54), specify how to handle the duplicate keys. Remember that "duplicate key" means a particular primary key value appears more than once within the same file. If you know that the files do not have any duplicates, ignore *Duplicate Key Handling*. Note that the options for handling duplicate key numeric columns differ from non-numeric columns. The column types must be specified correctly in order for duplicate key handling to work properly.

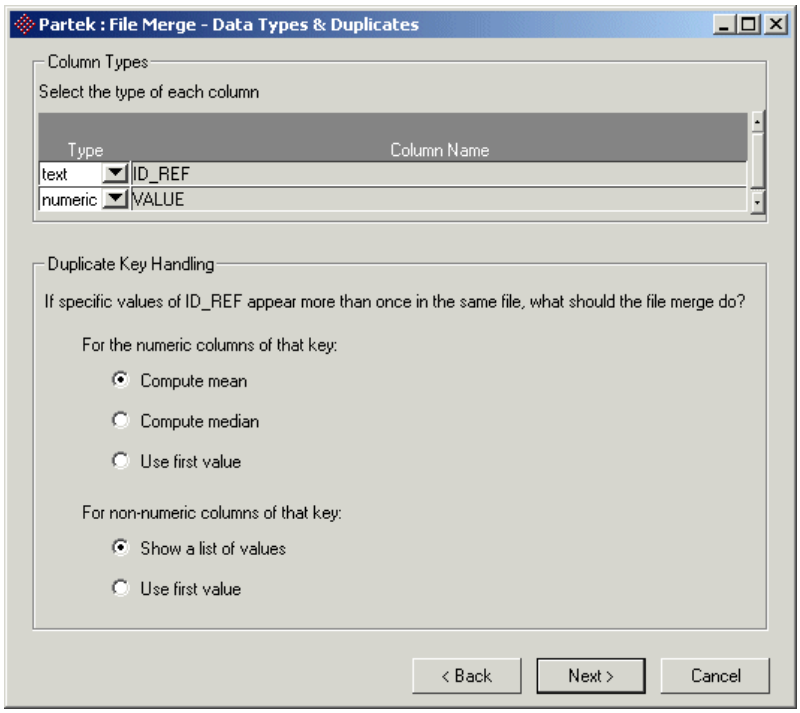


Figure 4. 54: Configuring the File Merge – Data Types & Duplicates dialog showing example column types and duplicate handling

The *Keys In Columns* setup is now finished.

- Select **Next >** to begin merging the files or click **<Back** to return to previous *File Merge* dialogs to make corrections

The *File Merge – Merge Status* dialog (Figure 4. 55) will show the status of the file merging process. When merging is complete, the final merged file will be automatically loaded into a Partek spreadsheet.

- Select **Close** on the *File Merge – Merge Status* dialog (Figure 4. 55) to close the file merge utility

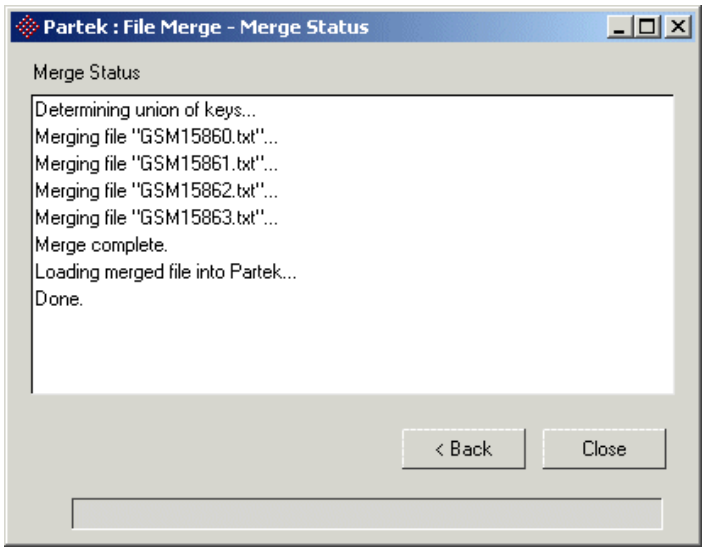


Figure 4. 55: Viewing the File Merge – Merge Status dialog after a completed file merge

Exercise: Downloading Data from NCBI GEO (Gene Expression Omnibus)

In this exercise, you will learn how to import NCBI GEO data into Partek.

Importing NCBI GEO Data

The NCBI Gene Expression Omnibus (GEO) is a large repository of public domain gene expression data. The GEO Importer, part of the Partek® Genomics Suite™, easily accesses this data in a usable format. In addition, the importer automatically downloads the data from the repository and formats the data so that it can be easily analyzed using Partek, Excel™, or any other tool that can read tab-delimited data.

Available experiments are listed at http://www.ncbi.nlm.nih.gov/geo/. Click on the **GSE Series** button under the **Public Data** heading on the GEO main page. This example will download GSE1025, along with others, from GEO.

Opening the GEO Downloader:

- Select **File > Import > NCBI GEO Downloader…** from the main menu (Figure 4. 56)

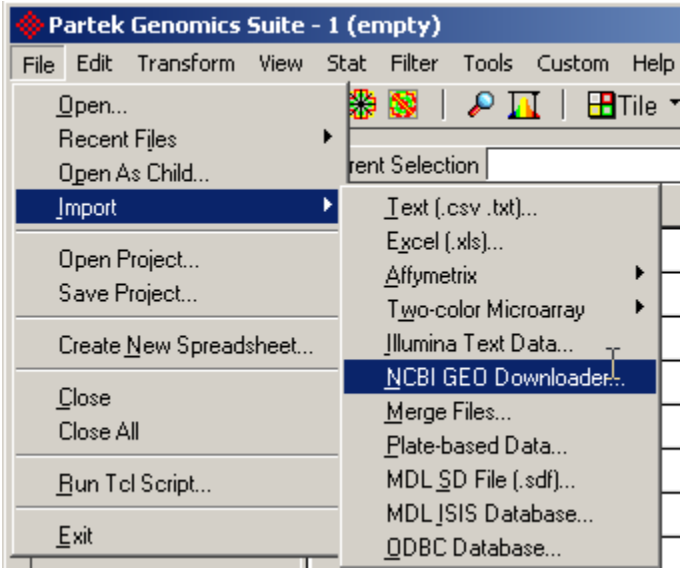


Figure 4. 56: Invoking the NCBI GEO Downloader

Specifying the Experiment to Download:

- Specify a folder to place the downloaded files
- Specify the **GSE number** to be **1025**

The dialog should look like Figure 4. 57.

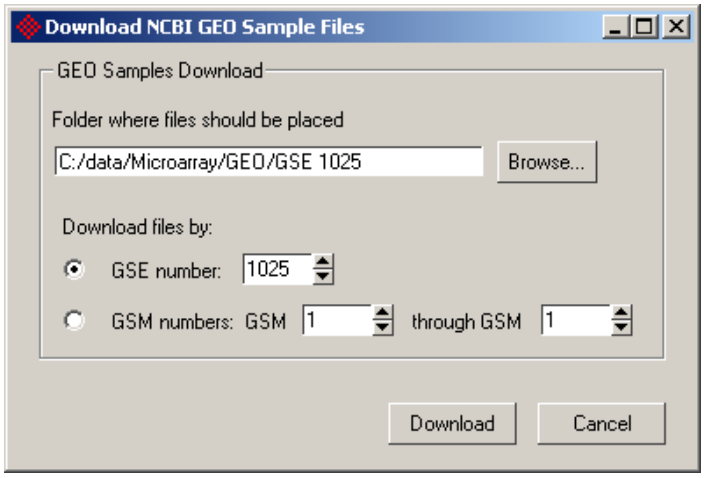


Figure 4. 57: Configuring the GEO Importer to download and import experiment GSE1025

- Click **Download…** to begin downloading the sample files from GEO

By specifying the GSE experiment ID, Partek automatically determines which sample files need to be downloaded. Figure 4. 58 shows the progress of the download. These data files are relatively large, so ideally, you should have a fast Internet connection for downloading these files. Depending on your Internet speed, the download can take several minutes or more.

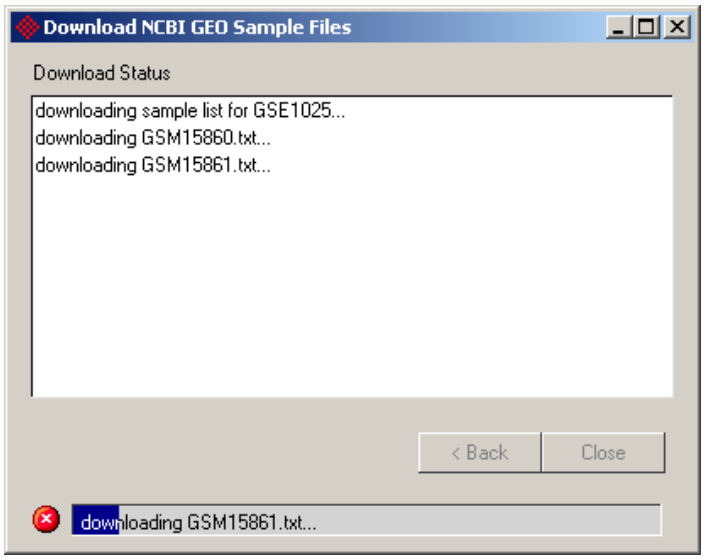


Figure 4. 58: Viewing the GEO Importer downloading the individual sample files

Merging the Files

Once all the files have downloaded to your local computer or network, you will be prompted to merge the files.

- Click **Yes** (Figure 4. 59)

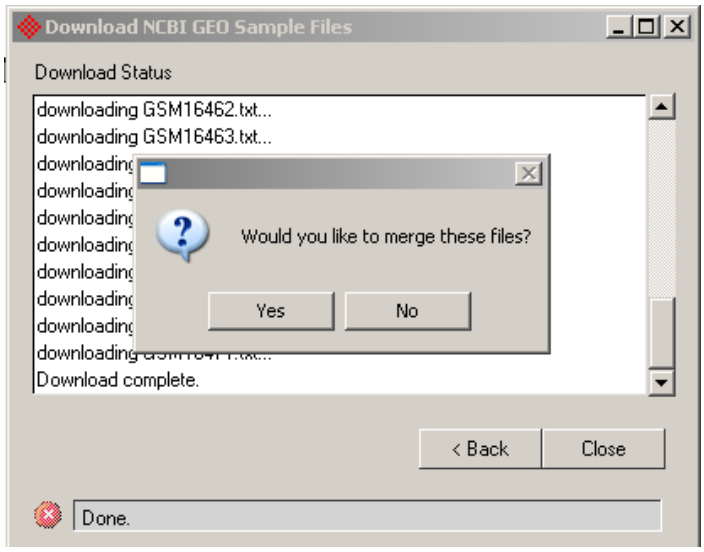


Figure 4. 59: Download complete

- Select the name of the folder where the files to be merged can be found
- Specify the name of the final merged data file

The name is automatically filled in when merging files are imported using the *GEO Downloader*. For most operations in Partek, you will want the genes in the columns and the samples in the rows.

- Configure the dialog to put the "keys" (genes) in the columns as shown in Figure 4. 60

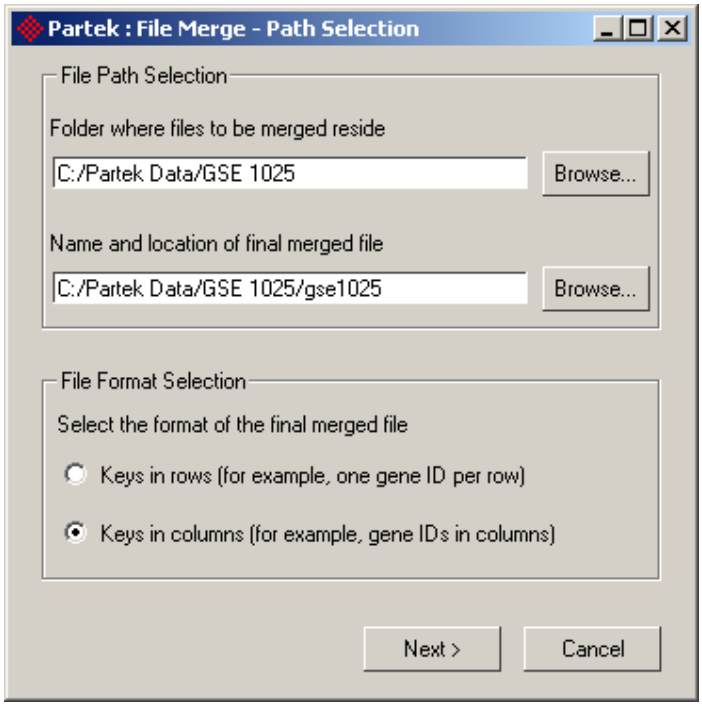


Figure 4. 60: Identifying the source, destination, and format of the data

Partek automatically selects the 36 files that were downloaded as the files to be merged (Figure 4. 61).

- Click **Next** > to proceed

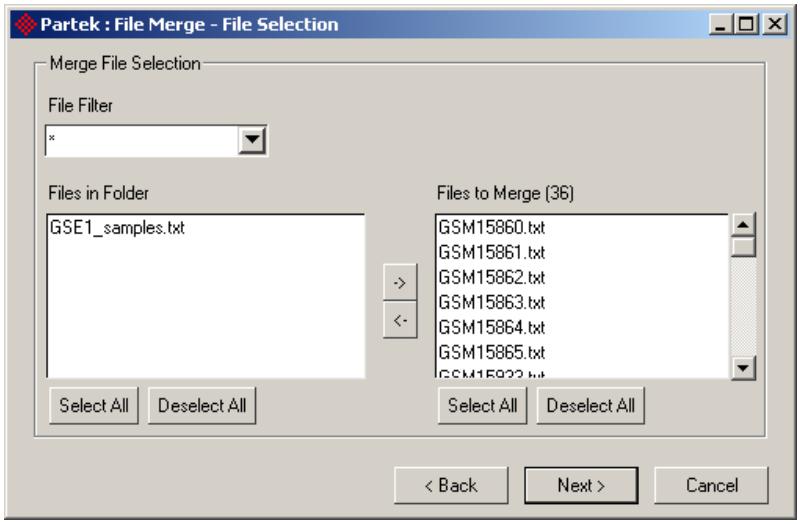


Figure 4. 61: Selecting the files to be merged

Selecting the Fields to be Extracted from each File

For Affymetrix data, the **VALUE** field is usually the only field of interest.

- Select that field and click **->** to move it into the **Columns to Include in Merge** box

The *Columns used to join files (key columns)* is automatically identified as the **ID_REF** column.

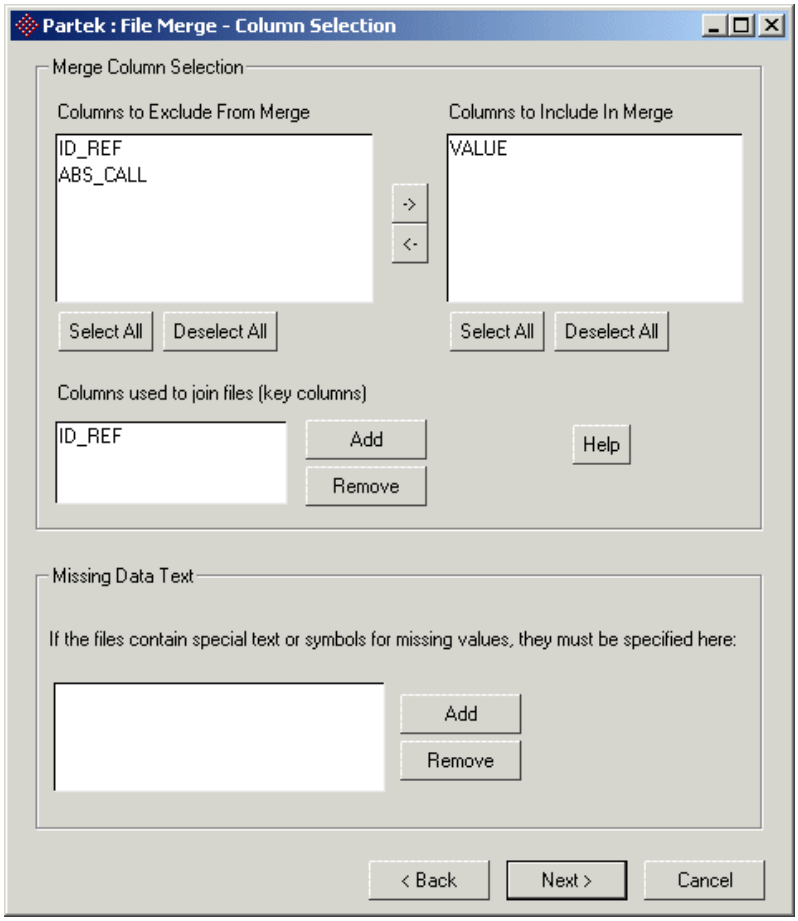


Figure 4. 62: Specifying the fields to extract and the key column

Identifying if there are Duplicate Values of a Gene on the Chip

For this Affymetrix chip, each probe appears only once in each file, therefore the ways to deal with duplicate data does not matter.

- Click **Next** > (Figure 4. 63)

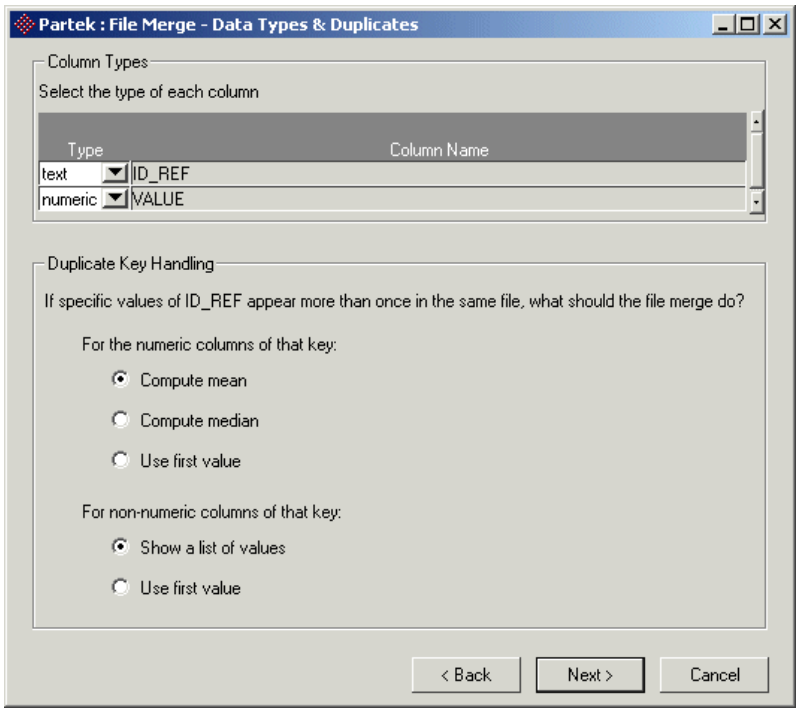


Figure 4. 63: Deciding how to deal with duplicate data

The sample information is not included in the data; it will be manually added later, so use the default settings as shown in Figure 4. 64.

- Click **Next** >

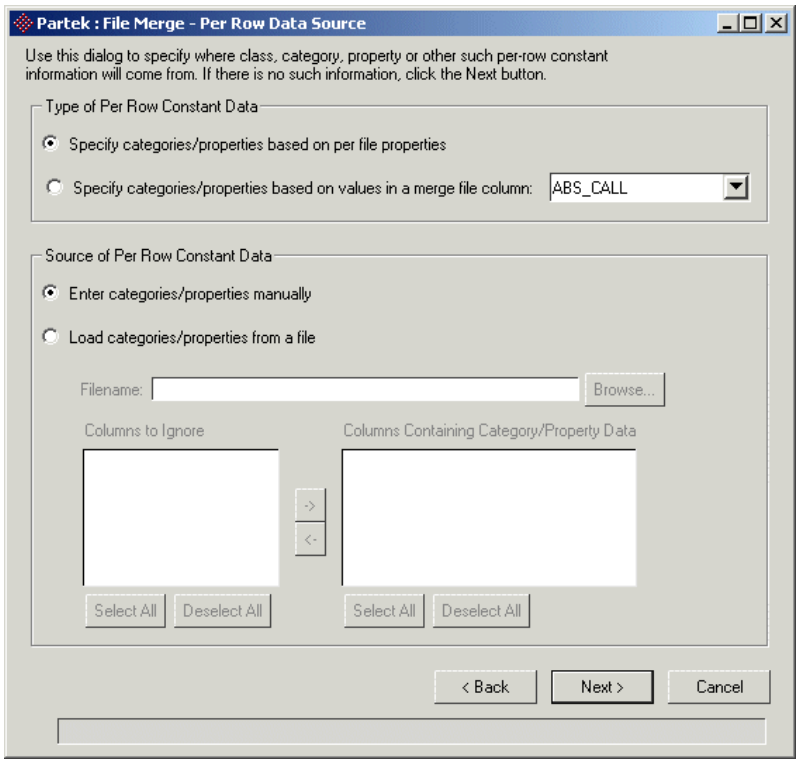


Figure 4. 64: Specifying sample properties

- Select the **Properties** column by clicking on it and click the **Delete Selected Property Columns** button since no sample information will be added at this time (Figure 4. 65)
- Click **Next >**

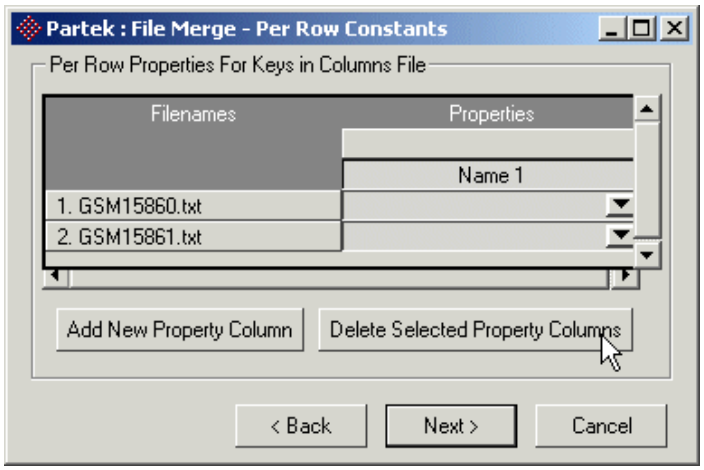


Figure 4. 65: Deleting selected property columns

Finally, the sample information to be inspected or edited is shown.

- When you are ready to extract and merge the files click **Next >**

After all the files have been merged successfully, you should see the dialog in Figure 4. 66. You are now ready to begin your analysis of this data using Partek software.

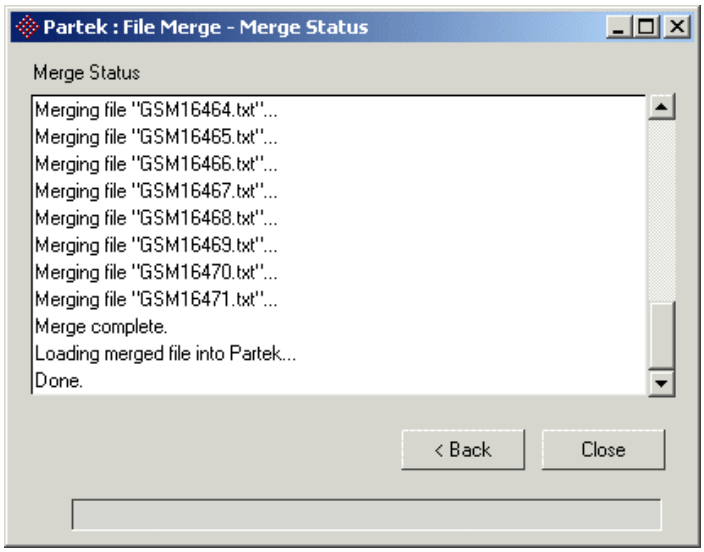


Figure 4. 66: Viewing the finished File Merge – Merge Status

Merging Files in Partek

Merging Text Files or Vendor-Specific File Formats Into One Spreadsheet

The *Partek File Merge* utility provides the ability to merge multiple text files into one file and then load that merged file into Partek. Partek File Merge recognizes several standard and proprietary file types, not just comma-separated (.csv) and tab-delimited (.txt or .tsv).

Before illustrating the steps of the file merge process, it is helpful to discuss terminology and concepts, as well as describe some general information about file merging.

Note: This section explains the terminology and concepts of merging file in Partek. Directions for merging files can be found in the *Importing from the NCBI GEO Database into Partek:* **Merging Files: Keys in Columns/Keys in Rows**.

File Merge Terminology and Concepts

The most common type of file to be merged contains information such as compound IDs, gene names, patient IDs, or protein IDs in the rows of a table, measurements, or other information in the columns (such as activation, intensity, expression level, concentration, etc.). For these types of files, the compound IDs, gene names, patient IDs, and protein IDs (or whatever the unique row-based identifiers are) are called keys because they are the key to matching data in one file with data in another file.

For example, if subject ID A12 appears in each of four files, then merging information from the files is possible by using subject IDs as the key. When merging files using Partek, you will be asked to specify the columns in the tables that are key columns, i.e. the columns that contain the keys that can be used to join the files together.

Consider two files to be merged represented by the following two tables:

Subject ID	Disease type	Measurement 1	Measurement 2	Measurement 3
A12	A	0.5	0.6	0.7
A20	B	0.3	0.2	0.1
B15	B	5.3	3.3	2.0

Table 4. 7: Example file to be merged

Subject ID	Disease type	Measurement 1	Measurement 2	Measurement 3
A12	A	1.5	1.0	2.2
A20	B	1.3	2.2	1.8
B15	B	3.3	4.3	3.0

Table 4. 8: Second example file to be merged

There are two ways in which you may want to merge these files together: keys in rows (e.g. subject IDs are in the rows of the final merged file), or keys in columns (e.g. subject IDs are in the columns of the final merged file). The final merged file might look something like those shown below.

If these two spreadsheets were merged with keys (subject IDs) in rows, the result might look like Table 4. 9 below.

Subject ID	Disease type	File 1 Measurement 1	File 1 Measurement 2	File 2 Measurement 1	File 2 Measurement 2
A12	A	0.5	0.6	1.5	1.0
A20	B	0.3	0.2	1.3	2.2
B15	B	5.3	3.3	4.3	3.0

Table 4. 9: Example of Keys in Rows merged file

If the two spreadsheets are merged with keys (subject IDs) in columns, the result might look like Table 4. 10 below.

Filename	Treatment	Time	A12 Measurement 1	A20 Measurement 1	B15 Measurement 1
TreatmentA30min	A	30	0.5	0.3	5.3
TreatmentB30min	B	30	1.5	1.3	3.3

Table 4. 10: Example of Keys in Columns merged file

Notice that the keys in columns spreadsheet is not just a simple transpose of the keys in rows spreadsheet. In the keys in columns spreadsheet, per-file-based constants, such as treatment and time, have been inserted. Also, note that not all of the columns of data have been merged into the two merged files. The File Merge utility allows you to select which columns to include in the final merged file.

Some files may only have one key column, while others may have a primary key column and one or more alternate key columns. Alternate keys are keys that uniquely identify a subject of interest (compound, gene, patient, protein) but are not used as the key to match data from one file to another. For example, a gene can have a gene name, but it can also have a GenBank® Accession number, a GI number, and possibly even a company-proprietary identifier. All of the “names” for a single gene are synonymous. Any of them can be used to uniquely identify a gene of interest. During the merge process, you will be asked to select which key to use as the primary key for merging files and which keys to use as alternate keys.

File Merge General Information

Column labels/names are important to the merging process. They are the only means by which the merging process can determine if a column in one file represents the same thing as a column in another file. If none of the column labels match among the files to be merged, then the files cannot be merged and Partek will display an error message. If most of the column labels match across the files except for a few, then the column labels that don't match will be ignored by Partek.

While column labels must match among the files to be merged, the order of the columns does not need to match. The Partek File Merge utility understands that a "Subject ID" column in one file is the same as a "Subject ID" column in another file, regardless of the order of the columns in their respective files.
Partek uses the union of all possible keys in the files to be merged, not the intersection. In other words, if subject ID A12 appears in only two of four files to be merged, the File Merge utility will recognize that and will insert missing data symbols ("?") into the final merged file where data is missing for subject ID A12.
If a particular primary key value appears more than once in the same file, it is considered a "duplicate key" (e.g. the same subject ID appears more than once in the same file). The File Merge utility will let you choose how you want to handle the measurement values of a duplicate key. You can choose to compute the mean or median of numeric values or take the first duplicate's value. For text values of a duplicate key, you can choose to keep all the values or just the first one.

Partek File Merge automatically recognizes the following file formats:

File Format	Description
Tab-separated (.tsv,.txt)	Tab-delimited records, one row per line
Comma-separated (.csv)	Comma-separated records, one row per line
NCBI GEO	See http://www.ncbi.nlm.nih.gov/geo/info/soft.cgi
Spotted Array (ATF 1.0)	GenePix export, ATF 1.0 format
Protein Mass Spec Hits	Columns: protein ID, hits, uniques
CloneTech Microarray	Tab-delimited CloneTech microarray data file
SMD	Stanford Microarray Database file
Affymetrix CHP	The ASCII text format of the Affymetrix CHP file. Note: you will have to use Affymetrix software to export the CHP file to a text file before using Partek File Merge.

Table 4. 11: File formats and descriptions

Saving Data

There are three options for saving your data in Partek, they are *Save…*, *Save As Text File*, and *Save As Web Page*. *Save* will save the file using the same name and overwrite the original version of the file. *Save As Text File* will save the file under a different name without overwriting the original version of the file. *Save As Web Page* saves the file in html format. You can also save the file by using the *Save As Text File* accelerator button ().

References

Bolstad, B.M., Irizarry R. A., Astrand, M., & Speed, T.P. (2003), A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Bias and Variance. *Bioinformatics* 19(2):185-193

Irizarry, R.A., Bolstad, B.M., Collin, F., Cope, L.M., Hobbs, B., Terence, P., & Speed, T.P. (2003), Summaries of Affymetrix GeneChip probe level data *Nucleic Acids Research* 31(4):e15

Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U., & Speed, T.P. (2002) *Exploration, Normalization, and Summaries of High Density Oligonucleotide Array Probe Level Data.*

Huang J, Wei W, Zhang J, et al. Whole genome DNA copy number changes identified by high density oligonucleotide arrays. *Human Genomics* 2004; 1:287–99.

Affymetrix® Copy Number Analysis Tool™ (CNAT) 3.0 Manual

Affymetrix® GeneChip Genotyping Analysis™ Software (GTYPE) 4.0 Manual

Managing the Analytical Spreadsheet®

Managing Rows and Columns in the Analytical Spreadsheet®

Rows and columns in the Analytical Spreadsheet® can be managed under the *Edit* menu.

Note: If you desire to edit individual cells in the spreadsheet only, and not whole rows or columns, they can be edited by clicking on the desired cell and editing the value as it appears in the *Current Selection* panel located above the Analytical Spreadsheet® and below the accelerator buttons (Figure 5. 1 To accept the change, press the **<Enter>** or **<Tab>** key on the keyboard. The selected entry will automatically advance along the current row or column (depending on the *Auto-Advance* setting for the spreadsheet).

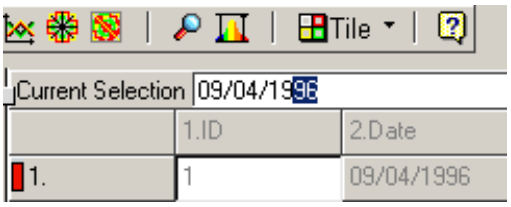


Figure 5. 1 : Viewing the Current Selection panel

Column Types

There are seven types of columns in Partek, but only four are generally used; they are bolded below.

text:	variable length string
categorical:	variable length nominal
double:	double precision floating point (8 bytes) (-1.7E308 to 1.7E308)
float:	single precision floating point (4 bytes) (-3.4E38 to 3.4E38)
integer:	integer (4 bytes) (-2147183648 to 2147483647)
short:	short integer, 2 bytes (-32768 to 32767)
byte:	1 byte (0 to 255)
snp:	genotype calls (AA, BB, AB, or NC)

Column Attributes

Three column attributes are used in Partek.

factor:	a variable that causes or influences another variable
response:	a variable that is caused by or influenced by another variable

Note: By default, numerical columns are automatically imported into Partek as response and double precision. Text columns are automatically imported as variable length nominal columns.

Cloning Spreadsheets

You can make a copy of a spreadsheet to maintain several states of filters or transformations by using the *Clone Spreadsheet* dialog (Figure 5. 2). The dialog is found at **Edit > Clone Spreadsheet.**

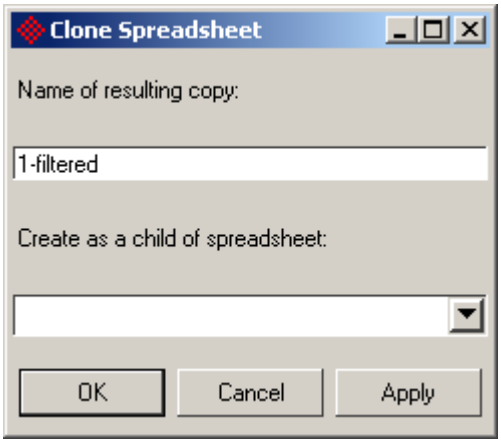


Figure 5. 2: Configuring the Clone Spreadsheet dialog

By using the *Create as a child* option, you can move a result spreadsheet from one parent to another.

Sorting Rows in the Analytical Spreadsheet®

Sorting Rows in the Analytical Spreadsheet® by Column

You can sort the spreadsheet based on the values of a specific column or by the similarity of the rows (Figure 5. 3). This dialog is found at **Edit > Sort Rows**.

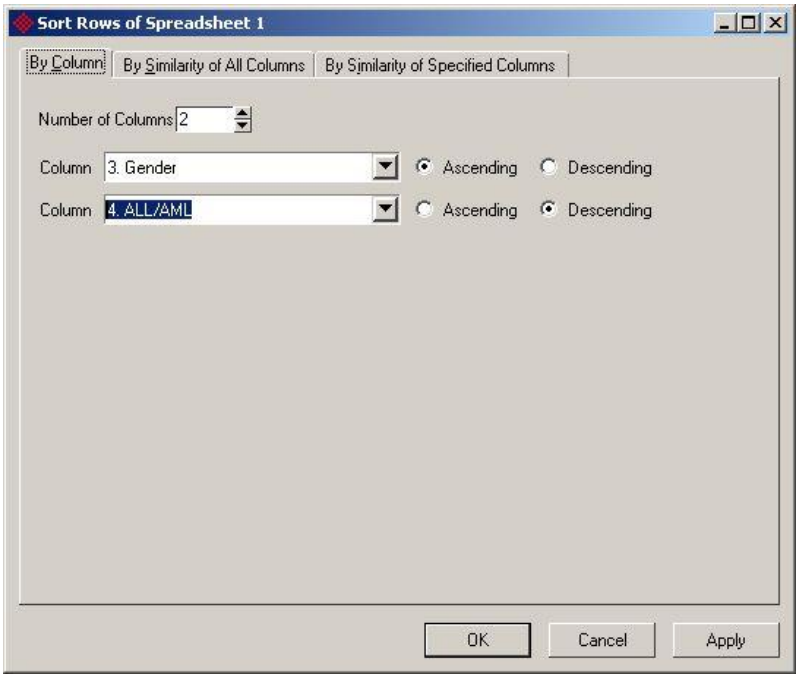


Figure 5. 3: Configuring to sort the rows by column

- Choose the column(s) that contain the values to be used to sort the spreadsheet from the drop-down list. The columns can be sorted either by *Ascending* or by *Descending*
- Select **Edit > Sort Rows** from the Partek main menu
- Select **OK** or **Apply** to invoke the sort

The row that has the smallest value in the specified column will be at the very top (ascending) or bottom (descending). This functionality can also be accomplished by right clicking the column header and choosing **Sort Ascending** or **Sort Descending** from the pop-up menu (Figure 5. 4).

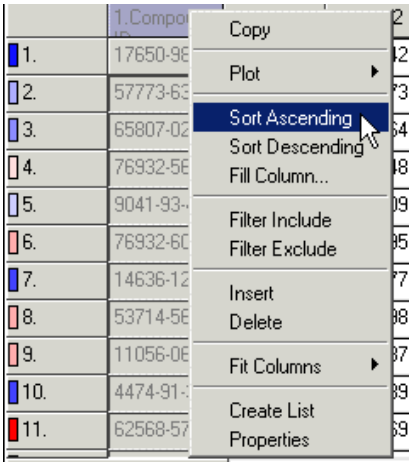


Figure 5. 4: Selecting to sort rows in ascending order from the pop-up menu

Sorting Rows in the Analytical Spreadsheet® by Similarity of All Columns

To sort by similarity of all columns, specify the method to use for calculating the distance between the rows from the drop-down list (Figure 5. 5).

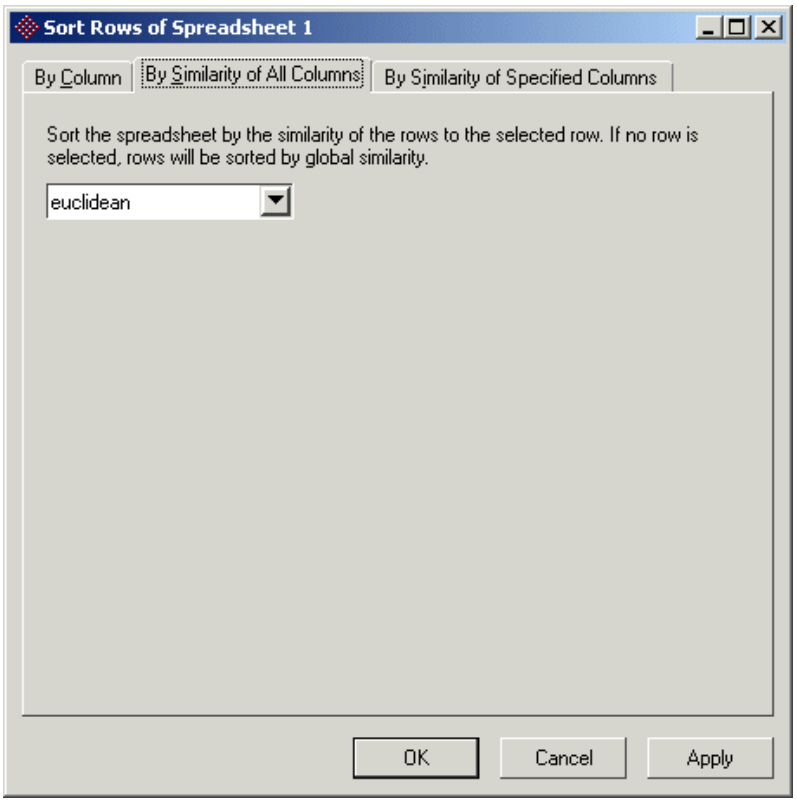


Figure 5. 5: Configuring to sort rows by Similarity of All Columns

Sorting Rows in the Analytical Spreadsheet® by Similarity of Specified Columns

Sorting rows by similarity of specified columns can be used to aid in finding rows (compounds) that are selectively active against certain assays (columns) (Figure 5. 6).

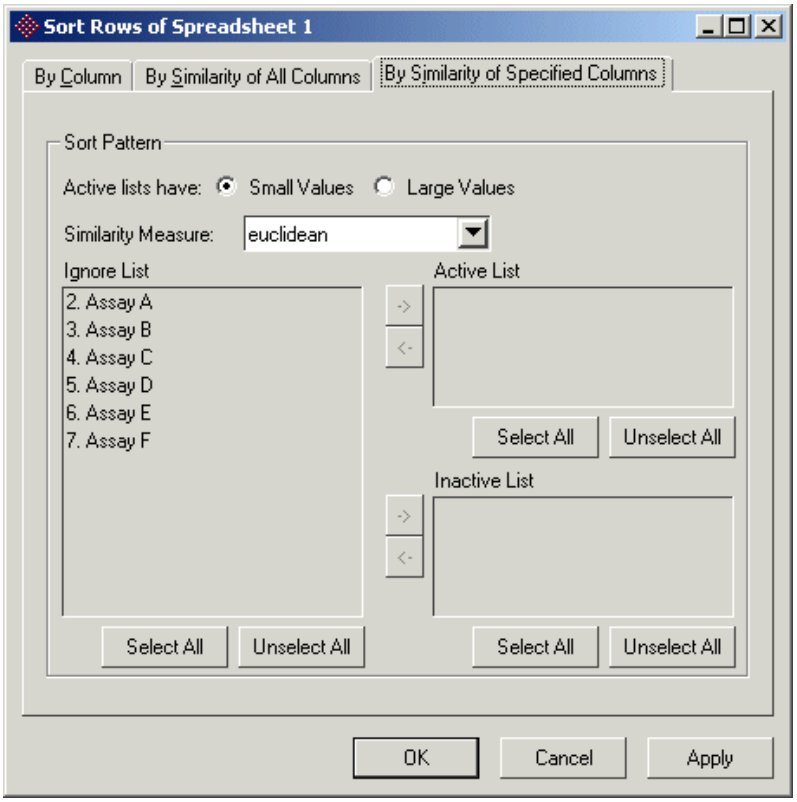


Figure 5. 6: Configuring to sort rows by Similarity of Specified Columns

A prototype will be created that consists of the selected active and inactive variables (assays). When the activity is specified as having small values, the prototype will be created by taking the minimum value(s) from the *Active List* and the maximum values from the *Inactive List*. Each row in the spreadsheet will be compared to the prototype, and the row (compound) most similar to the prototype will be placed in row 1. The row least similar to the prototype will be placed in the last row in the spreadsheet. Table 5. 1 shows the prototype.

Compound #	Assay A	Assay B	Assay C
1	3	7	9
2	10	3	11
3	12	8	4

Table 5. 1: Viewing an example prototype for selectively active compounds

For this example, activity is denoted by small values. If you want to find compounds that are selectively active on Assay B but not Assay A and Assay C, the prototype of Table 5. 2 will be created.

Assay A	Assay B	Assay C
12	3	11

Table 5. 2: Viewing an example prototype for selectively active compounds on Assay B

Using the Euclidean distance as the *Similarity Method*, the new order of this spreadsheet is shown in Table 5. 3.

Compound #	Assay A	Assay B	Assay C
2	10	3	11
3	12	8	4
1	3	7	9

Table 5. 3: Viewing an example prototype using the Euclidean distance method

Specify if the active compounds have *Small Values*, indicating inactivity, or *Large Values*, indicating activity. Also, select the *Similarity Measure*, which increases as the similarity between objects increases (Figure 5. 7).

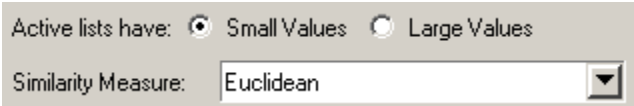


Figure 5. 7: Configuring the Sort panel

The *Ignore List* contains the values of the compounds that are included in the main spreadsheet, but are not necessary to the sort; however, the values for both the *Active List* and *Inactive List* are taken from this list. There must be at least one item in the *Active List* and the *Inactive List*. Select the corresponding -> <- buttons to move a selected item to the appropriate list.

Multiple items can be selected at one time in all three lists by pressing the **<Ctrl>** or **<Shift>** keyboard keys while left clicking on the items. To select all of the items in a list, click the **Select All** button, and click **Unselect All** to deselect.

Adding Rows and Columns to the Analytical Spreadsheet®

To add rows or columns to the spreadsheet, select **Edit > Add Rows/Columns** from the Partek main menu and specify the number of rows and columns to be added and where to add them in the spreadsheet.

In addition to specifying the number of columns to be added and where to add them, the column properties must be configured. If the *Label Prefix* is specified, the first column in the newly added column set will be labeled exactly as the text in the entry; all the following columns will be labeled with that as a prefix, followed by a number in parenthesis, e.g. Treatment (1), Treatment (2) … etc. After inserting a column, the number of the columns after the new column will shift (Figure 5. 8).

If a *Label Prefix* is not specified, the columns will be added without a column label.

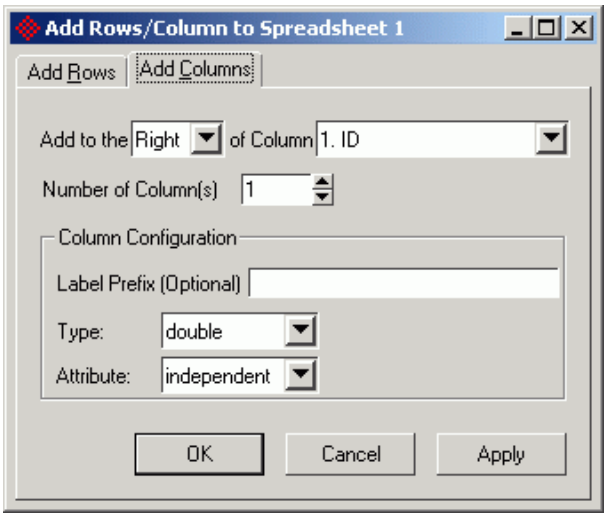


Figure 5. 8: Adding Rows/Columns dialog, Add Columns tab

If genomic information is on rows (as in an ANOVA result spreadsheet) then you will have the ability to add information from the annotation file in a new column (Figure 5. 9).

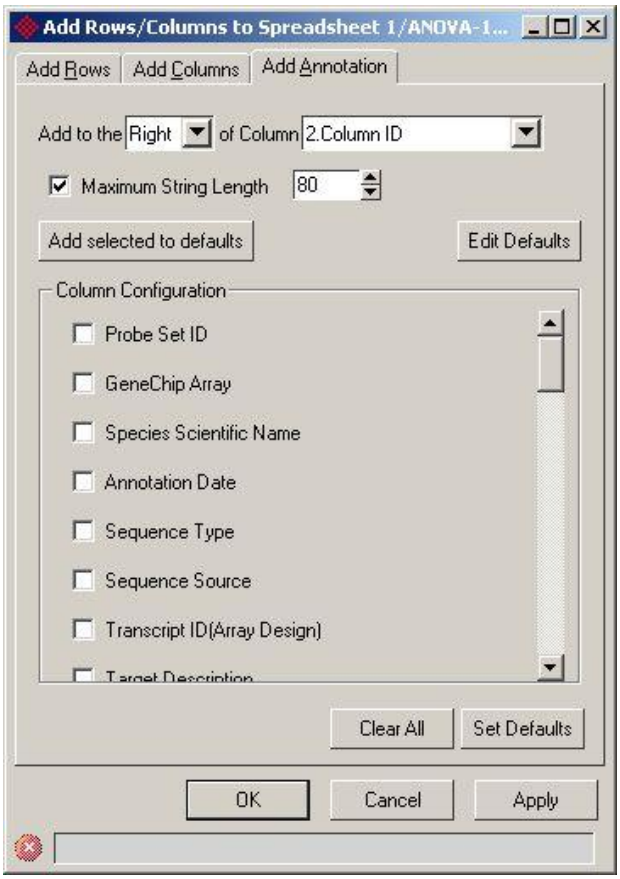


Figure 5. 9: Adding annotations

If the spreadsheet has the “region” property (**File > Properties**) specified, then you will have the option to add to each region the average value from another spreadsheet. Figure 5. 10 shows the *Add Average* dialog.

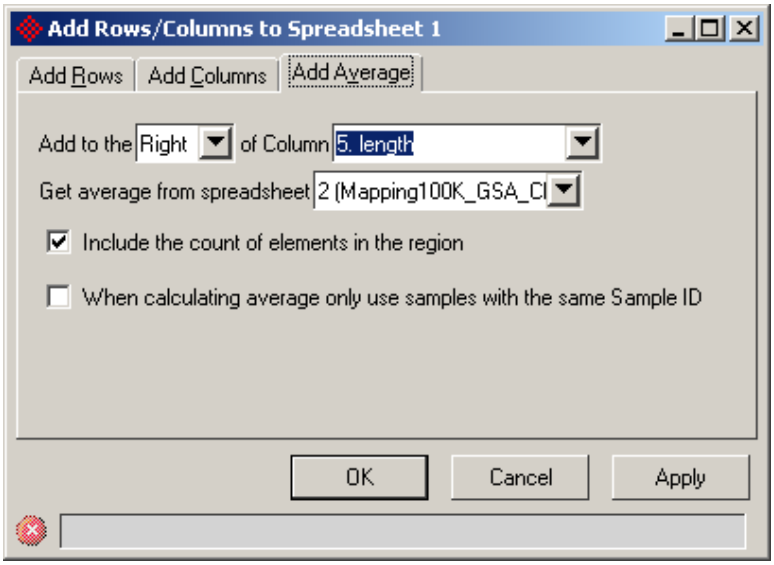


Figure 5. 10: Adding average

If the region spreadsheet has a sample ID column, then you will have the option to match the sample ID in the region spreadsheet with the sample ID in the other spreadsheet.

Filling Columns in the Analytical Spreadsheet®

To fill columns in the spreadsheet, select the column to be filled from the *Column to Fill* drop down list, and select **Edit > Fill Column** from the Partek main menu. If the column type is categorical, give the names of the categories (levels) and the range of the rows belonging to each level, respectively (Figure 5. 11).

Selecting the **Add Category** button will add a new *Category Name* and *Row Range* of the specified category. If the rows are continuous, use a dash to connect the first row and the last row in the section (e.g. 1-10); otherwise, use the space bar to separate different rows (e.g. 1 2 3) and different fields (e.g. 1 2 3-5).

If the spreadsheet does not contain a class variable, the *Make this the class variable* check button will be enabled to make the column a class variable. There can be only one class variable in the spreadsheet.

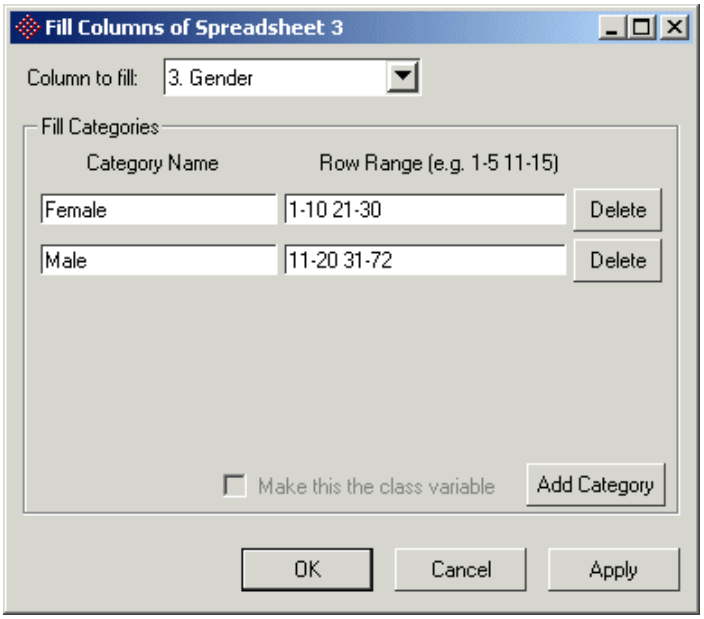


Figure 5. 11: Configuring the Fill a Categorical Column dialog

If the column type is not categorical, i.e. either text or numeric, Partek can auto fill a series in the column provided with *Start Value* and *Step Value* (Figure 5. 12). If a *Prefix* is specified for a text type column, it will be put in front of the serial numbers.

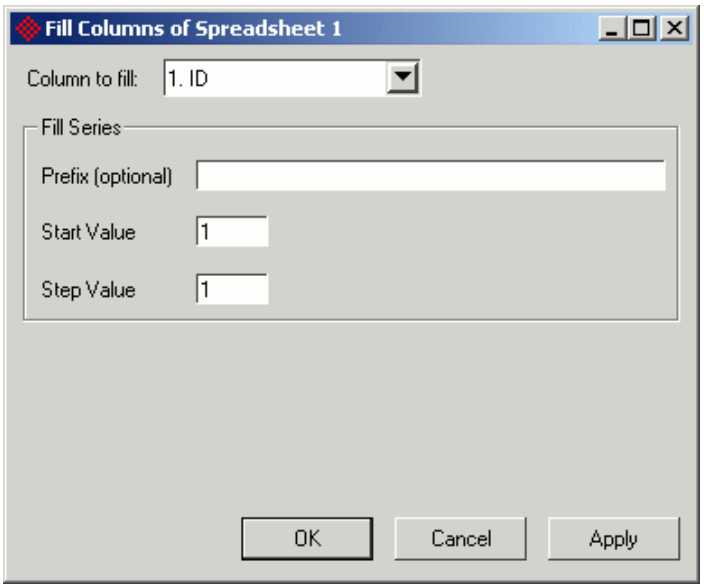


Figure 5. 12: Filling a Non-Categorical Column dialog

Merging Columns in the Analytical Spreadsheet®

By merging columns in the spreadsheet, you can combine the values of two or more columns on the same row using the defined string and put the new values in a new column (Figure 5. 13); to merge columns in the spreadsheet, select **Edit > Merge Columns** from the Partek main menu.

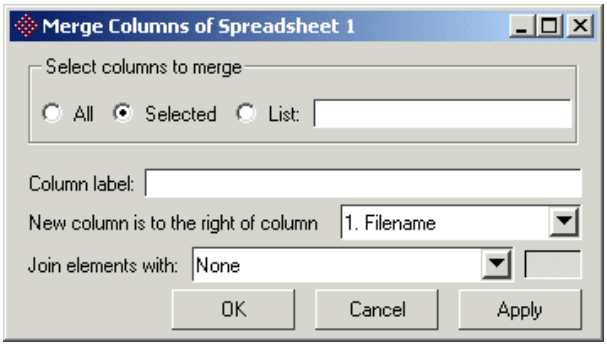


Figure 5. 13: Configuring the Merge Columns dialog

Choose the columns to merge and specify the label of the new column. The default is set to merge *Selected* columns. The order of the new string depends on the column order of the selected columns. Specify a list of column numbers by doing the following: if the columns are continuous, use a hyphen to connect the first column and the last column in the section (e.g. 1-10); otherwise, use the space bar to separate different columns (e.g. 1 2 3) and different fields (e.g. 1 2 3-5). The order of the new string depends on the order of the list provided (Figure 5. 14).

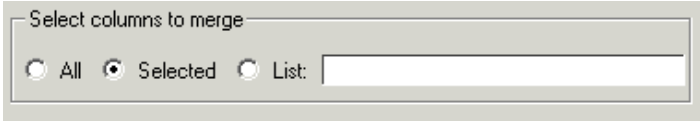


Figure 5. 14: Specifying which columns to merge

Select the text column from the drop-down list and specify where the new columns will be added (Figure 5. 15).

Figure 5. 15: Specifying where the new column is added

Specify a character or a string to join the elements from the drop-down list (Figure 5. 16).

Figure 5. 16: Configuring to split text by a delimiter

If *String* is selected, the string needs to be specified; it can be multiple characters, like *abc*.

Splitting Columns in the Analytical Spreadsheet®

If a column is a categorical or text type, you can divide the text column into multiple columns based on delimiter(s) or character width(s) specification (Figure 5. 17); to split columns, select **Edit > Split Columns** from the Partek main menu.

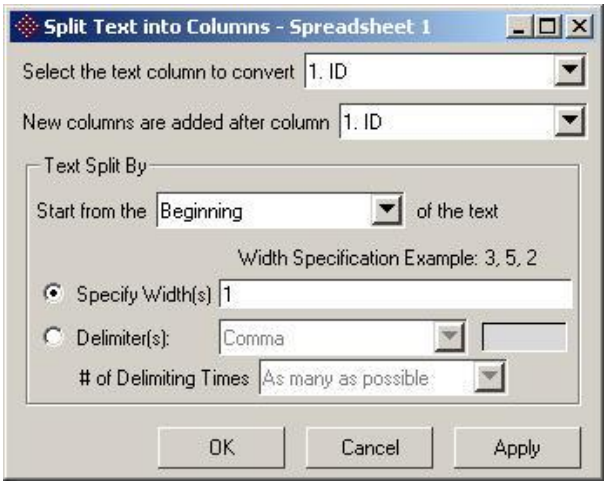


Figure 5. 17: Configuring the Split Text into Columns dialog

Select the text column from the drop-down list and specify where the new columns will be added (Figure 5. 18).

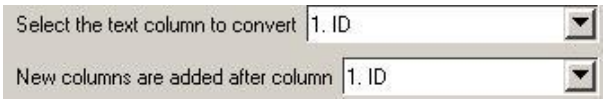


Figure 5. 18: Specifying which column to split and where new columns are to be added

There are two different methods to split the text:

- Split by specifying character width
- Split by delimiter

If the field you want to extract is aligned, for instance, you want to split a 9 digit ID number ‘xxxxxxxxx’ into 3 parts ‘xxx’, ‘xx’, and ‘xxxx’, specify the widths as shown in Figure 5. 19.

Figure 5. 19: Configuring to split text by specifying width(s) of the text

If the fields are separated by one or more delimiters, configure the dialog similar to Figure 5. 20. In Figure 5. 20, the upper panel uses *Hyphen (-)* as the delimiter; it can split ‘xxx-xx-xxxx’ into 3 parts ‘xxx’, ‘xx’, and ‘xxxx’. The middle panel uses a string 2005 as the delimiter; it can split ‘mm2005dd’ into 2 parts ‘mm’ and ‘dd’. The lower panel uses multiple delimiters; it can split a+{b*[c/(d-e)]}into 7 parts ‘a+’, ‘b*’, ‘c/’, ‘d-e’, and 3 trailing empty columns. To avoid empty columns, you can specify the *Number of Delimiting Times* as 4, then the result will be 5 parts ‘a+’, ‘b*’, ‘c/’, ‘d-e’, and a trailing part ‘]}’. Remember the number of inserted columns is the *Number of Delimiting Times* plus 1.

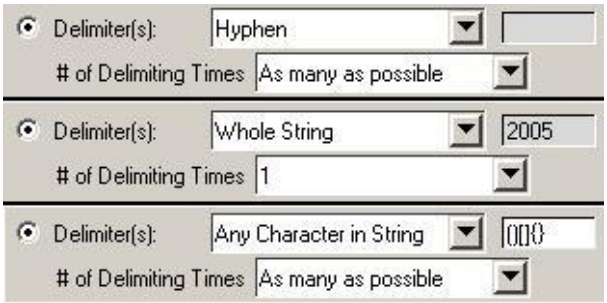


Figure 5. 20: Configuring the text split by delimiter(s)

In the *Split Text into Columns* dialog, you can also specify the splitting direction, either *Start from the Beginning* of the text or *Start from the End* of the text.

Selecting and Deselecting Cells in the Analytical Spreadsheet®

The Select/Deselect option can Deselect All cells in the spreadsheet, Deselect Rows, Deselect Columns, Select All Rows, Select All Columns, Invert Row Selection or Invert Column Selection; to use these options, select **Edit > Select/Deselect** from the Partek main menu.

Select Rows Based on a List Spreadsheet

When you have at least two spreadsheets open, there will be a *Select Rows Based on a List* option under the **Edit > Select/Deselect** menu (Figure 5. 21).

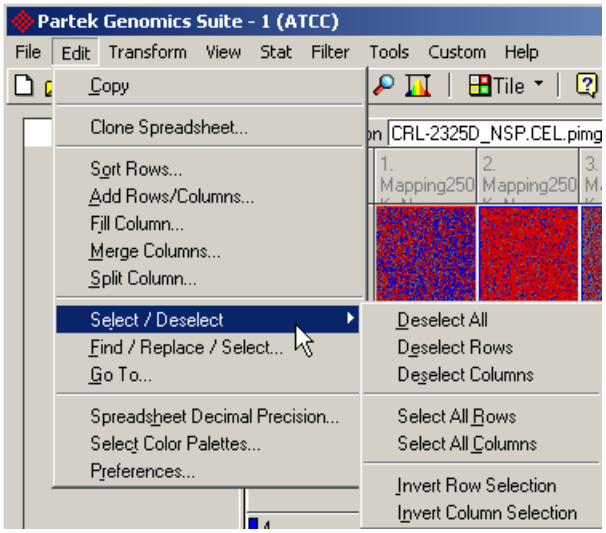


Figure 5. 21: Selecting rows/columns based on a list spreadsheet

This allows you to select rows of the current spreadsheet whose key column values match the key column values in another spreadsheet (Figure 5. 22).

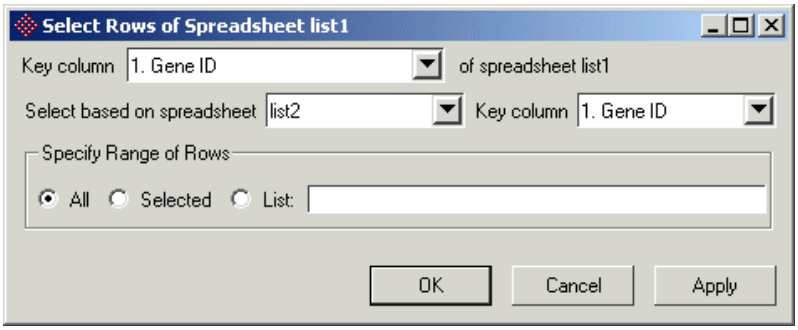


Figure 5. 22: Configuring the select rows dialog

Suppose the current spreadsheet has genes on the rows, in this example, the gene IDs are the values of column *1. Gene ID* (Figure 5. 23). Continuing with this example, if you wanted to select only the genes whose IDs are the values of column *1. Gene ID* of spreadsheet *list2* (Figure 5. 24), then the column, which contains the genes' IDs as the values in both spreadsheets, is the *Key column*.

Figure 5. 23: Configuring the key column of the current spreadsheet (Spreadsheet 1)

Figure 5. 24: Filtering based on a spreadsheet and its key column

If *All* is selected in the *Specify Range of Rows* panel (Figure 5. 25), all of the values in the column, *1. Gene ID* of *list2,* which also appear in column *1.Gene ID* of the current spreadsheet (spreadsheet *1*), will be selected; if you select *Selected* rows or you specify a list of row numbers, only the values of *1. Gene ID* in those rows will be selected in the current spreadsheet.

Figure 5. 25: Specifying a range of rows

Selecting Columns Based on a List Spreadsheet

When you have at least two spreadsheets open, there will be a *Select Columns Based on a List* option under the **Edit > Select/Deselect** menu (Figure 5. 21).

Selecting *Based on a List* allows you to select response numeric columns whose headers match the value of the key column of another spreadsheet (Figure 5. 26)

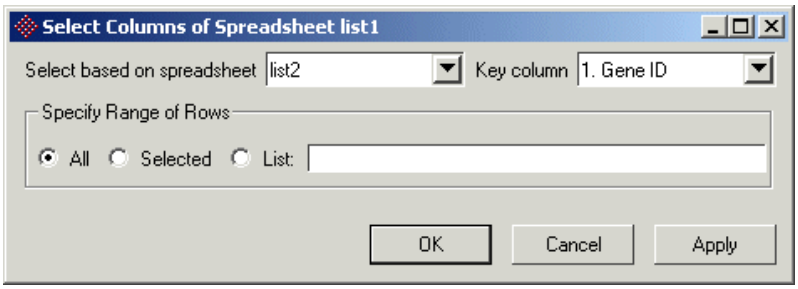


Figure 5. 26: Configuring the select columns dialog

Suppose the current spreadsheet has genes in the columns and the genes' IDs are column headers. If you want to select the genes that have IDs that are the values of column *1. Gene ID* of *list2* (Figure 5. 27), select *Key column* as the column that has the values of the genes' IDs.

Figure 5. 27: Filtering based on spreadsheet and its key column

If *All* is selected in the *Specify Range of Rows* section (Figure 5. 28), all of the values in column *1. Column ID* of *list2* that appear in column *1.GeneID* of the current spreadsheet (spreadsheet *1*) will be selected; if you select *Selected* rows or you specify a list of row numbers, only the values of *1.GeneID* will be selected the current spreadsheet.

Figure 5. 28: Specifying a range of rows

Finding and Replacing Data in the Analytical Spreadsheet®

To find and replace data in the spreadsheet, find cells containing specific text, and/or any text that needs to be replaced, select **Edit > Find/Replace/Select** from the Partek main menu. The search details can be specified (Figure 5. 29).

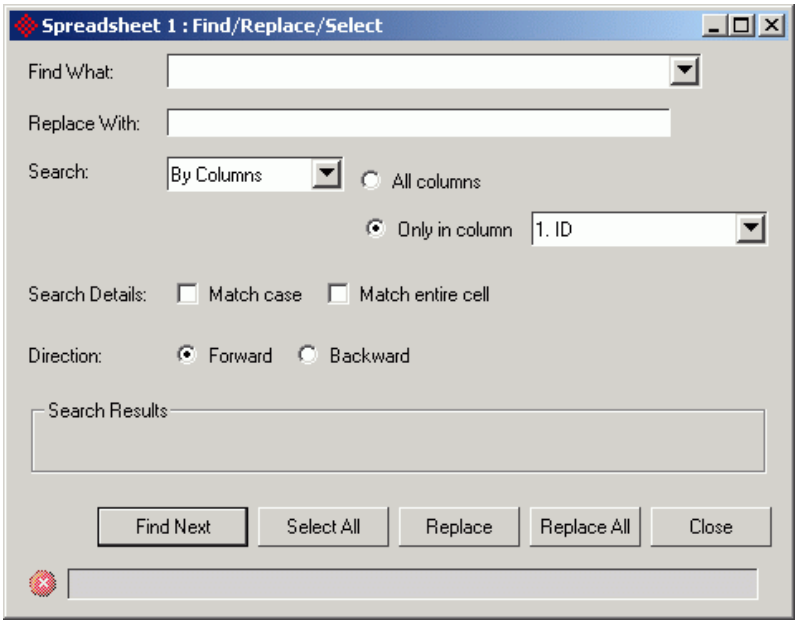


Figure 5. 29: Configuring the Find and Replace dialog

Navigating the Analytical Spreadsheet® (Go To)

To navigate to a specific cell in the spreadsheet select **Edit > Go To,** and the *Go To* dialog will appear (Figure 5. 30)

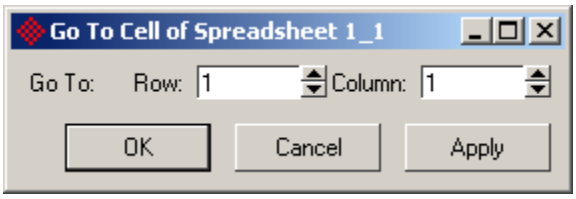


Figure 5. 30: Go To Cell dialog

Sorting Rows by Prototype

If there are no more that 200 variables in the spreadsheet, the rows can be sorted by prototype. To invoke the *Sort Rows by Prototype* dialog, select **Tools > Discover > Sort Rows by Prototype** from the Partek main menu (Figure 5. 31). A figure like the one pictured in Figure 5. 32 will appear.

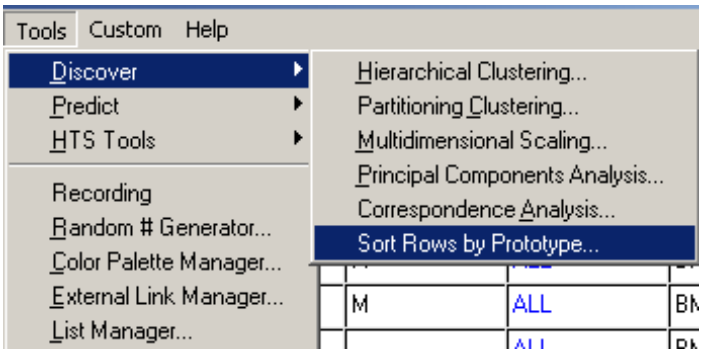


Figure 5. 31: Accessing the Sort Rows by Prototype viewer

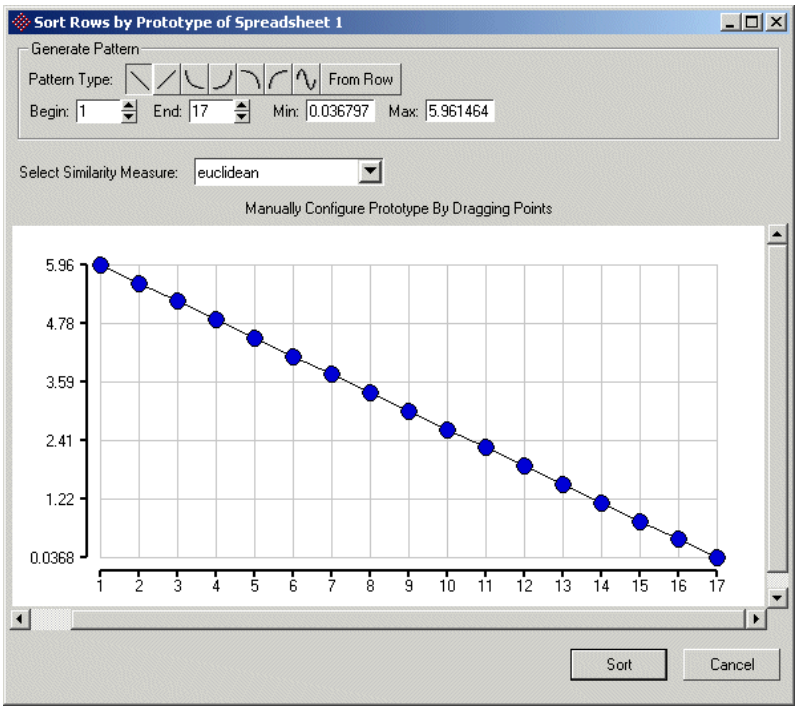


Figure 5. 32: Sorting rows by prototype—Pattern Type I

In the graph, the X-axis represents the numeric columns on the spreadsheet; the Y axis represents the value of a row. By default, the first *Pattern Type* is selected and a straight line is drawn -- the first numeric column has the maximum value of the spreadsheet and the last numeric column has the minimum value. Select the second *Patten Type* by clicking on it; another straight line will appear when the first column is *min* and the last column is *max*. You can also change the value of *min/max* and the *beginning/ending* of the columns to draw the line pattern.

When one of the 3rd to the 6th *Pattern Types* is selected, you can specify the *Decay Rate* from the first column to the last column (Figure 5. 33).

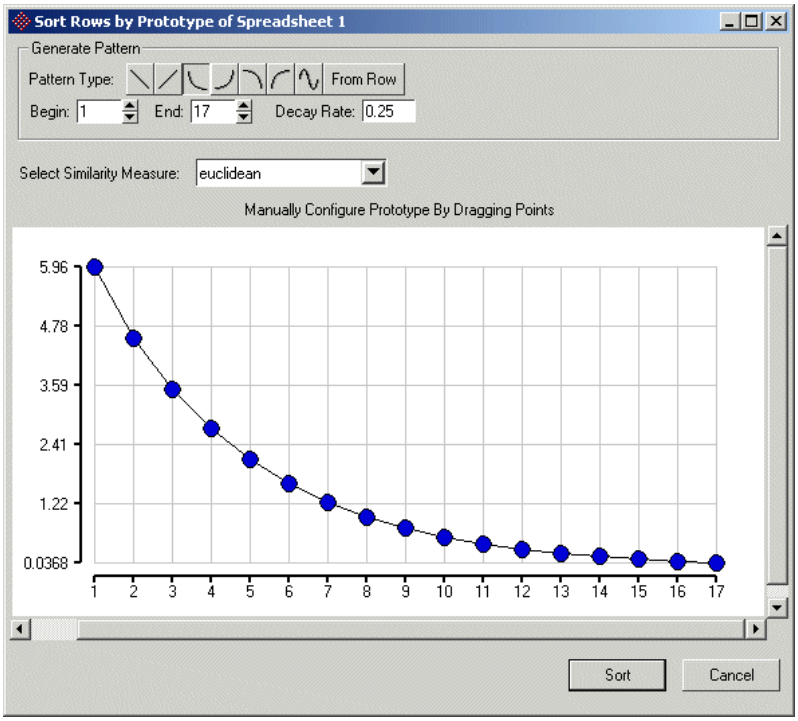


Figure 5. 33: Sorting rows by prototype – Pattern Type III

When the 7th pattern button is pressed, the prototype is a sine wave (Figure 5. 34).

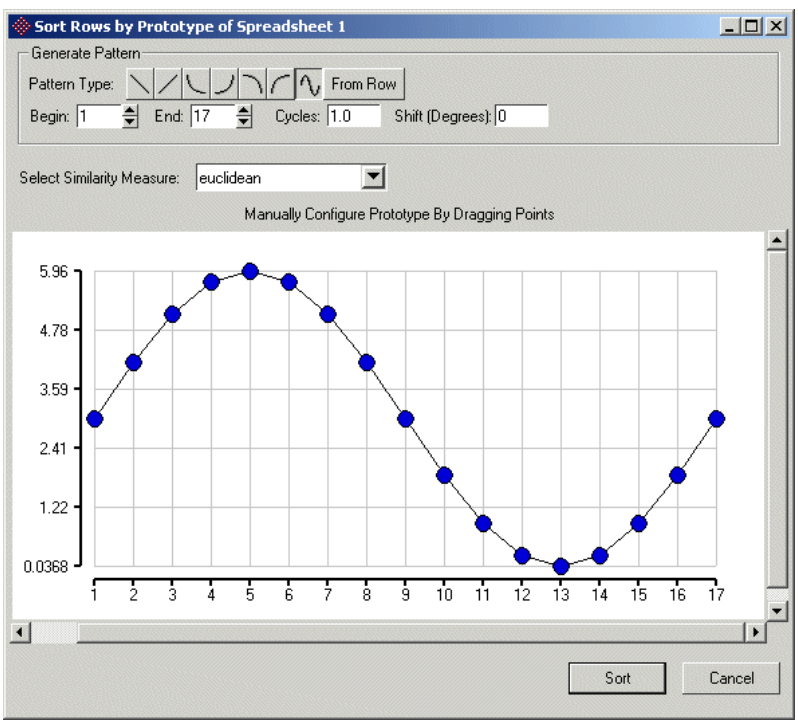


Figure 5. 34: Sorting rows by prototype – Pattern Type VII

By default, there is one cycle. *Shift* specifies how many degrees to shift the sine wave; 0 degrees is the default.

When you mouse over a point, the column number and column header of the point represented will pop-up. Click on the pop-up to manually specify the value of the point by dragging it up or down.

Select **Sort** to sort the spreadsheet so that the first row is the one that best matched the specified prototype using the similarity measurement selected from the drop-down list. The second row is the second closest match to the pattern, and the last row contains the patterns that are the most dissimilar to the prototype.

Filtering Rows and Columns in the Analytical Spreadsheet®

Data Filtering Process

Depending on the research objective, you may wish to focus on just a portion of the data. Partek's row filters and column filters were designed for this purpose. They are powerful and flexible and can be used to easily filter-in or filter-out portions of the data based on literally any criteria.

Row Filters

Row filters are used to determine which rows are retained in the spreadsheet for analysis. They can be easily configured to select observations of the data based on any criteria. Additionally, row filters can be used to randomize and resample data to easily enable such analyses as the *bootstrap*, *cross-validation*, *jackknife*, and more.

Column Filters

Column Filters are filters that determine which columns (variables) are used for analysis of the data. Column filters are most commonly used to analyze just a subset of variables.

The combination of row filters, column filters, and a scripting interface provide unlimited ways in which you can select to filter in or filter out portions of data for a particular analysis. This example showed a few usage scenarios for filtering data ranging from simple interactive filtering to more complicated data selection.

Filtering Rows

Invoking the Interactive Filter

The interactive filter is part of the graphical front end functionality that the row filter mechanism offers. To invoke the interactive filter, choose ***Filter > Filter Rows > Interactive Filter*** *or click on the accelerator button (*
Figure 5. 35).

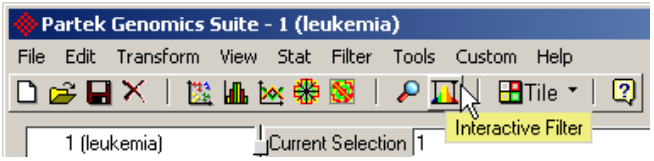


Figure 5. 35: Selecting the Interactive Filter accelerator button
When it is first opened, the interactive filter will, by default, display the class variable or the first categorical variable. If there are no class variables in the spreadsheet, any column can be selected by choosing it from the *Column* drop-down list. Below the *Column* entry is a small configuration menu (arrow button). Clicking on the button will allow you to **Clear All Filters** or control the type of filter being applied and change the appearance of the histogram.

If the column selected is a categorical variable, the bar chart will represent the distribution of each category, and the color of the bars are coded according to the category (Figure 5. 36). The category name and number of observations contained in a category will be displayed by using the mouseover function over each colored bar. Left clicking on a bar will toggle the filtering of the chosen category; right clicking on a bar will filter out all the other categories but the chosen one.

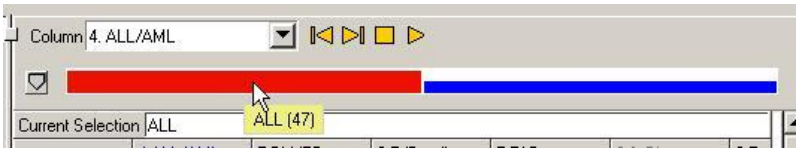


Figure 5. 36: Interactive Filter of Categorical Variables

If the column selected is a continuous variable, the *minimum value* and the *maximum value* of the specified column will be displayed to the right of the *Column* selector (Figure 5. 37). These values change to reflect the status of the range tabs and it can be directly typed in to set the precise *minimum* and *maximum* for the filter. The histogram of the distribution for a column with continuous values will contain two tabs to control the minimum and maximum values to be used by the filter. Drag the tabs using the left mouse button to adjust the *minimum* and *maximum* values, and the filter will update when the mouse button is released. This is useful for large data sets with the graphics visible because filtering and graphics updates will only occur one time when the mouse button is released. Using the right mouse button to drag the range tabs will cause the filter to update real-time, filtering the values in the spreadsheet and any graphics drawn from that spreadsheet as the mouse is moved. The **Configuration Menu** can configure the filter type (Figure 5. 37).

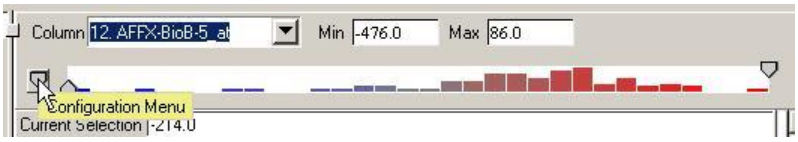


Figure 5. 37: Configuring the Interactive Filter of the continuous variables

The filter type can be **Pass** (include) or **Stop** (exclude). The default filter is the *Pass Filter*. When the filter type is *Pass*, the interactive filter will filter out all records with values <u>outside</u> the range tabs of the histogram (Figure 5. 38). When the filter type is *Stop*, it will filter out all records with values <u>inside</u> the specified range; however, in either case, filtered out portions of the data will be grayed-out to show the status of the filter.

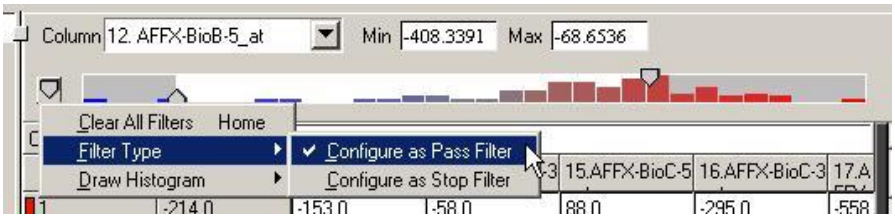


Figure 5. 38: Configuring the Filter Type

If the column is categorical then by default the number of rows in a given category is represented by the height of the bar. You also have the option to show all bars at the same height and display the number of rows as a label.

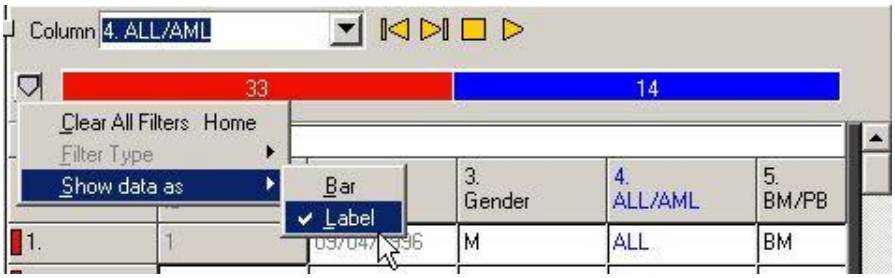


Figure 5. 39: Configuring the data display

When a row filter is applied to a spreadsheet, a gold bar will appear to the right of the spreadsheet. The length of the bar represents the portion of the rows that passed. To clear the row filter, right click on the gold bar; then left click on the **Clear Filter** pop-up (Figure 5. 40). Another way to clear all interactive filters applied to any and all columns of the spreadsheet is to select **Filter > Filter Rows > Clear Row Filters** from the Partek main menu. To clear the filter applied to a single variable hold down the <**Shift**> key while left clicking on the bar in the bar chart of the interactive filter.

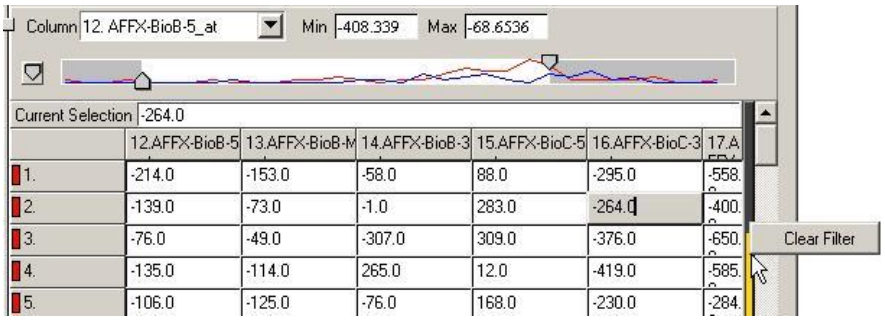


Figure 5. 40: Clearing Row Filters

To turn off the interactive filter, click on the **Interactive Filter** icon again.

Clearing Row Filters

To clear all row filters applied to the spreadsheet, select **Filter > Filter Rows > Clear Row Filters**.

Sampling Rows

To get a subset of the samples (rows) based on a regular interval in the current spreadsheet, click **Filter > Filter Rows > Sample Rows…**, specify the sampling interval in Figure 5. 41 and select **OK** or **Apply**; every 10th row will be included in the spreadsheet.

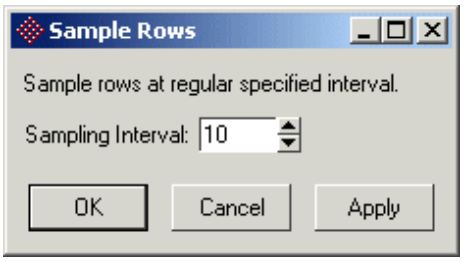


Figure 5. 41: Configuring the Sample Rows dialog

Shuffling Rows Randomly

To randomly shuffle rows click **Filter > Filter Rows > Randomly Shuffle**, and the order of the rows will change. This applies a filter to the spreadsheet; however, the gold bar will not appear to the right of the spreadsheet.

Note: To see what filters are applied to the spreadsheet, select **Filter > Row Filter Manager > Currently Applied**.

Managing the Row Filter

The *Row Filter Manager* creates filters, examines the current filters, and deletes filters.

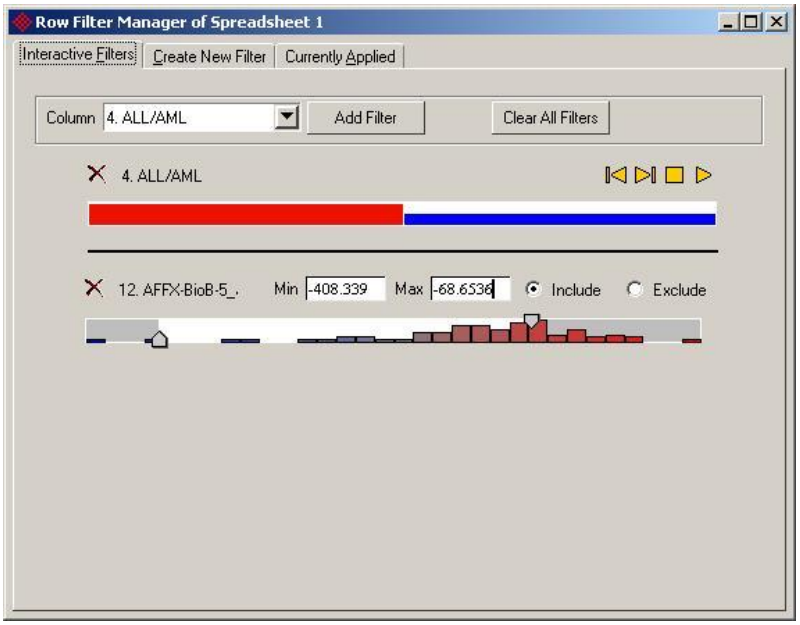


Figure 5. 42: Configuring the Interactive Filters page

In the *Create New Filter* page (Figure 5. 43), there are 4 filter type options:

- *Include*: Pass filter, includes the rows that meet the specified criteria in the *Filter Configuration* panel
- *Exclude*: Stop filter, excludes the rows that meet the specified criteria in the *Filter Configuration* panel
- *Randomize*: Randomly shuffle the rows
- *Resample*: Resample rows with replacement, the total number of rows remain the same, which means some rows may appear more than once and other rows may not be chosen. This is also called a bootstrap sample

Note: When *Randomizing* and *Resampling* the rows, the gold bar will not appear to the right of the spreadsheet even though row filters are applied.

The *Filter List* can specify a range with a dash ("–") in between the column numbers, e.g. "1-4" is the same as "1 2 3 4". In addition, a range can be specified *based on a value in a specified column* (Figure 5. 43).

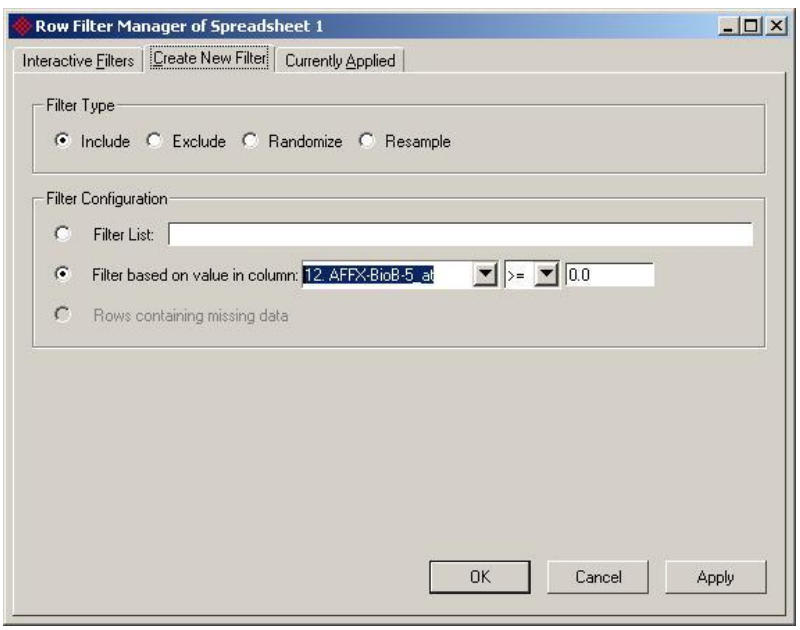


Figure 5. 43: Viewing the Create New Filter page of the Row Filter Manager

The *Currently Applied* page displays all the row filters used in the spreadsheet (Figure 5. 44). To check a specific filter, select a filter name from the *Filter List* panel on the left, and the configuration of the filter will be displayed on the right panel. To remove the selected filter, select **Delete**.

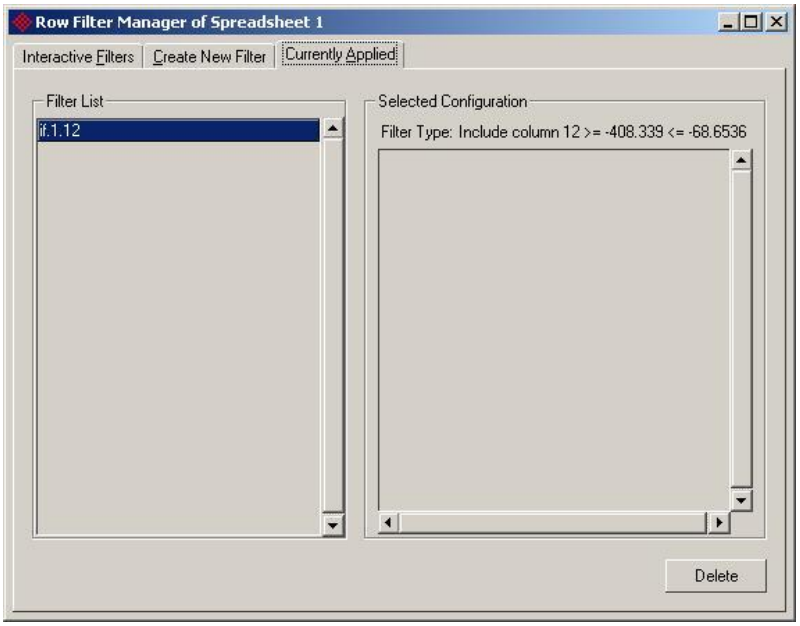


Figure 5. 44: Viewing the Currently Applied page of the Row Filter Manager

Filtering Rows Based on a List Spreadsheet

When you have at least two spreadsheets open, there will be a *Filter Rows Based on a List* option under the *Filter Rows* menu (Figure 5. 45).

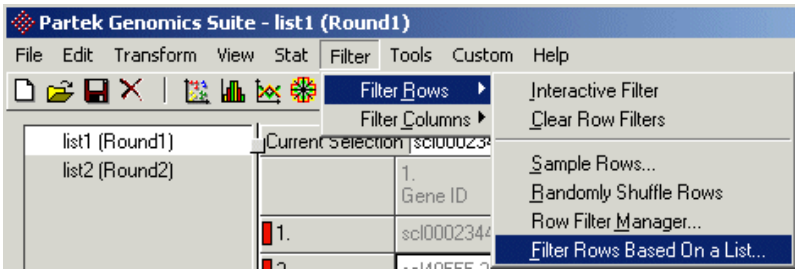


Figure 5. 45: Filtering rows based on a list spreadsheet

This allows you to filter to include rows of the current spreadsheet whose key column values match the key column values in another spreadsheet (Figure 5. 46).

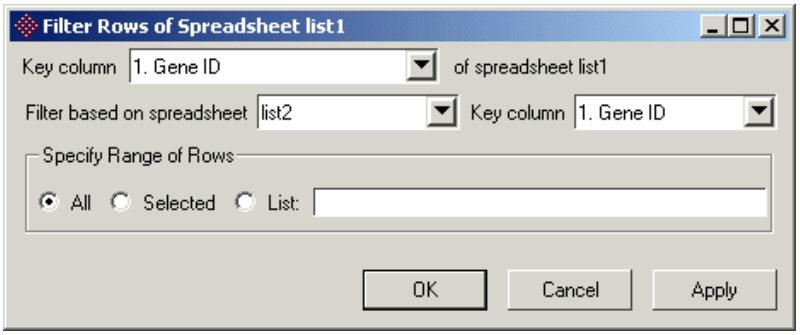


Figure 5. 46: Configuring the Filter Rows dialog

Suppose the current spreadsheet has genes on the rows, in this example, the genes IDs are the values of column *1. Gene ID* (Figure 5. 47). Continuing with this example, if you wanted to filter to include only the genes whose IDs are the values of column *1. Gene ID* of spreadsheet *list1* (Figure 5. 48), then the column, which contains the genes' IDs as the values in both spreadsheets, is the *Key column*.

Figure 5. 47: Configuring the key column of the current spreadsheet (Spreadsheet 1)

Figure 5. 48: Filtering based on a spreadsheet and its key column

If *All* is selected in the *Specify Range of Rows* panel (Figure 5. 49), all of the values in the column, *1. Gene ID* of *list1*, which also appear in column *1.Gene ID* of the current spreadsheet (spreadsheet *1*), will be filtered in; if you select *Selected* rows or you specify a list of rows, only the values of *1. Gene ID* in those rows will be filtered and included in the current spreadsheet.

Figure 5. 49: Specifying a range of rows

Filtering Columns

Sampling Columns

To get a subset of the numeric variables (columns) based on a regular interval of the current spreadsheet, click **Filter > Filter Columns > Sample Columns…** Specify the sampling interval as in Figure 5. 50 and click **OK**. A numeric variable can be either factor or response. If the *Variable Group to Sample* is specified as response, every 10th (example of Figure 5. 50) of the response numerical variables will be included in the sample; if the *Variable Group to Sample* is specified as factor, every 10th (example of Figure 5. 50) of the factor numerical variables will be included in the sample. In both cases, all the other variables will be included also.

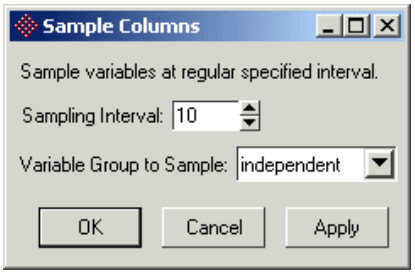


Figure 5. 50: Configuring the Sample Columns dialog

When there is a column filter applied to a spreadsheet, a gold bar will appear at the bottom of the spreadsheet. The length of the bar represents the portion of the columns that passed

through the filter. To clear the column filter, right click on the gold bar, and left click on the **Clear Filter** pop-up. Another way to clear the filter is to select **Filter > Filter Columns > Clear Column Filters** from the main menu.

Filtering Out Response Variables

Filtering Out Response Variables filters out only numeric variables whose attributes are response; to filter out response variables, select **Filter > Filter Columns > Filter Out Response Variables…** from the Partek main menu.

Filtering Out Factor Variables

Filtering Out Factor Variables filters out only numeric variables whose attributes are factor; to filter out factor variables, select **Filter > Filter Columns > Filter Out Factor Variables…** from the Partek main menu.

Clearing Column Filters

Clearing Column Filters deletes all the column filters that are applied to the spreadsheet; to clear column filters, select **Filter > Filter Columns > Clear Column Filters…** from the Partek main menu.

Managing the Column Filter

The *Column Filter Manager* allows you to create filters, examine the current filters, and delete filters; you can invoke the *Column Filter Manager* from **Filter > Filter Columns > Column Filter Manager** from the Partek main menu.

In the *Create New Filter* page (Figure 5. 51), there are 2 filter type options:

- *Include*: Pass filter, include the columns that are in the *Filter List* panel
- *Exclude*: Stop filter, exclude the columns that are in the *Filter List* panel

The *Filter List* can specify a range with a dash ("–") in between the column numbers, e.g. "1-4" is the same as "1 2 3 4". In addition, a column can be filtered based on its statistic value (min, max, mean, median, variance, or std. dev.) being less than, greater than, equal to, etc., than a numeric value (Figure 5. 51).

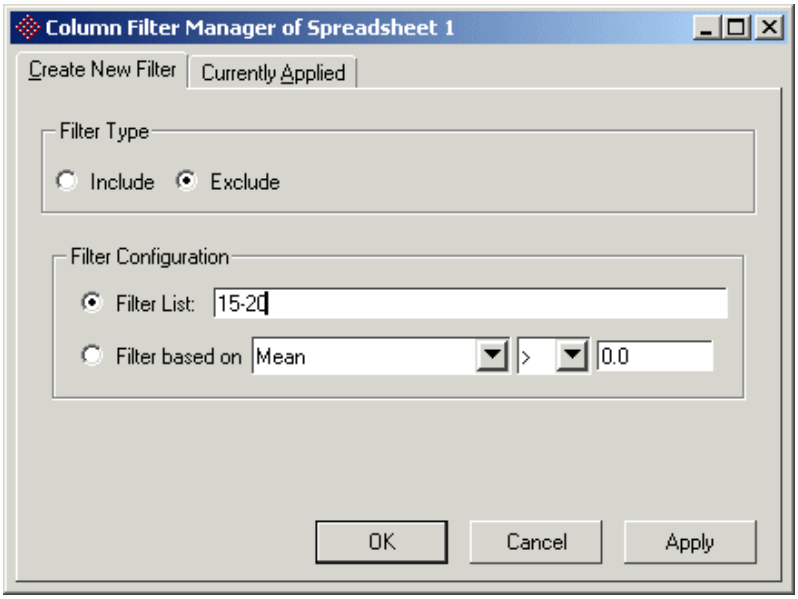


Figure 5. 51: Creating a new filter in the Column Filter Manager

The *Currently Applied* page displays all the column filters used in the spreadsheet (Figure 5. 52). To check a specific filter, select a filter name from the *Filter List* panel on the left, and the configuration of the filter will be displayed on the right panel. To remove the selected filter, select **Delete**.

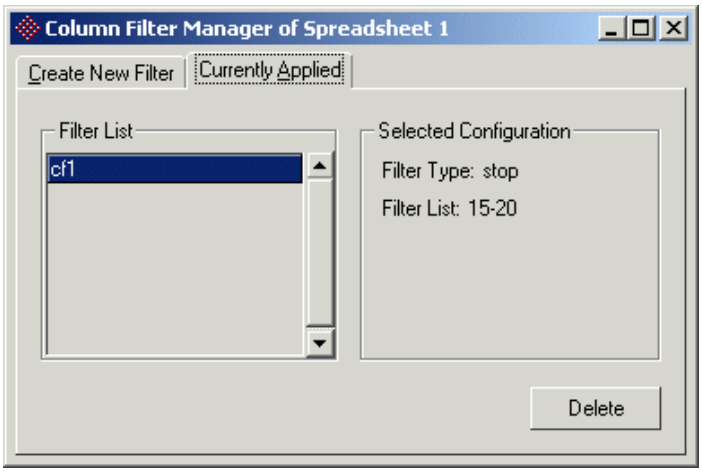


Figure 5. 52: Viewing the currently applied filters in the Column Filter Manager

Filtering on Test Results

If you performed inferential statistics in the spreadsheet (e.g. t-Test, ANOVA), and the results were stored in a child spreadsheet, there will be another item added on the *Filter Column* menu of the parent spreadsheet: *Filter on Test Results* (Figure 5. 53).

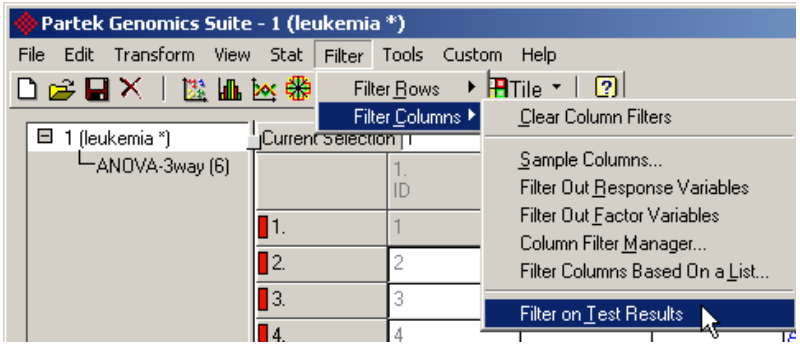


Figure 5. 53: Selecting to filter on test results

The column header of the numeric variables in the parent spreadsheet will be laid out on rows in the child result spreadsheet. This option is often used after sorting the child result spreadsheet based on a variable (e.g. p-value). To examine the variables that meet the filtering criteria go back to the parent spreadsheet (e.g. p-value <0.05) (Figure 5. 54).

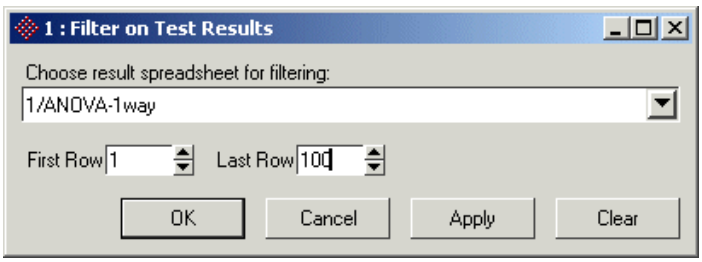


Figure 5. 54: Configuring the Filter on Test Results dialog

Filtering Columns Based on a List Spreadsheet

When you have at least two spreadsheets open, there will be a *Filter Columns Based on a List* option under the *Filter Columns* menu.

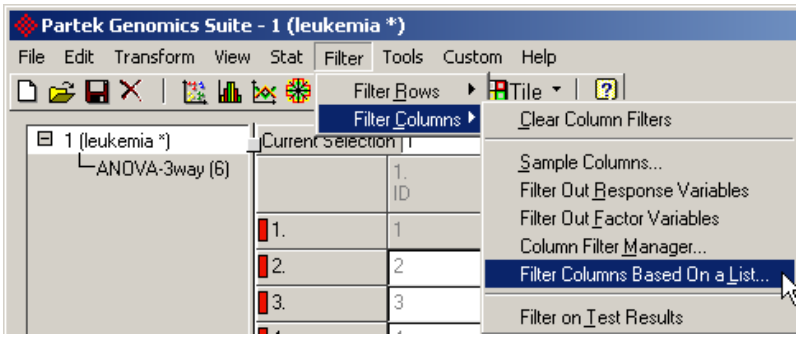


Figure 5. 55: Filtering columns based on a list

Filtering based on a list allows you to filter to include response numeric columns whose headers match the value of the key column of another spreadsheet (Figure 5. 56), in addition to all the sample information columns.

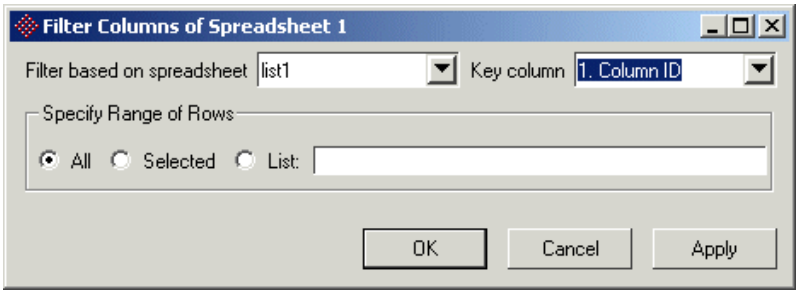


Figure 5. 56: Configuring the Filter Columns dialog

Suppose the current spreadsheet has genes in the columns and the genes' IDs are column headers. If you want to filter to include only the genes that have IDs that are the values of column *1. Column ID* of *list1* (Figure 5. 57), select *Key column* as the column that has the values of the genes' IDs.

Figure 5. 57: Filtering based on spreadsheet and its key column

If *All* is selected in the *Specify Range of Rows* section (Figure 5. 58), all of the values in column *1. Column ID* of *list1* that appear in column *1.ID* of the current spreadsheet (spreadsheet *1*) will be filtered in; if you select *Selected* rows or you specify a list of rows, only the values of *1.ColumnID* will be filtered and included in the current spreadsheet.

Figure 5. 58: Specifying a range of rows

Filtering based on Genomic Location

Filtering Based on Genomic Location is found in the Partek GS **Filter > Filter Based on Genomic Location** menu (Figure 5. 59).

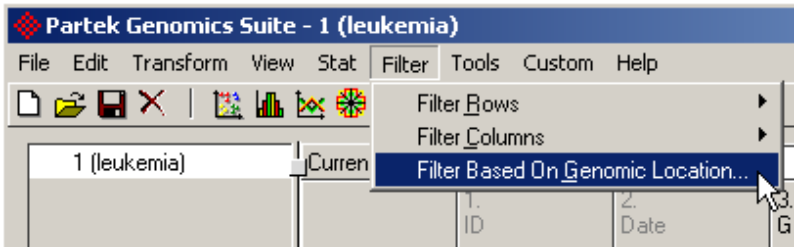


Figure 5. 59: Selecting the Filter Based on Genomic Location menu item

In the *Filter Spreadsheet Based on Genomic Location* dialog (Figure 5. 60), you can filter the spreadsheet based on chromosome, base pair locations, and any field in the annotation file.

You can also choose to create a list of column labels and use "Filter based on list" (see above).

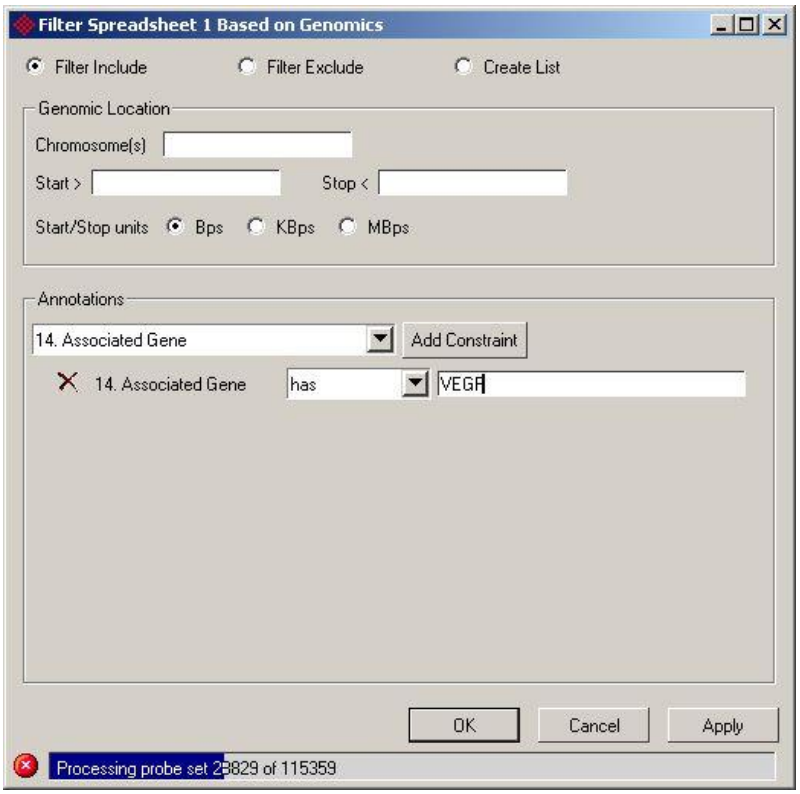


Figure 5. 60: Filtering based on genomic location

Summary

This concludes the chapter on data filtering. Row and Column filters are helpful in looking at specific data within the spreadsheet without the hindrances of unnecessary data. Please contact us at support@partek.com if you have any questions about using the row and column filters in Partek. You may close any graphics and spreadsheets that you have before continuing.

The Pattern Visualization System®

Introduction

The Pattern Visualization System® contains pertinent graphs and plots that produce quality interactive data visualizations to help you gain a better understanding of your data.

Each section in this chapter describes a different plot and its function. The graphs and plots are viewed within a "viewer", which have their own menu and mode buttons. The viewer menu and mode buttons are discussed in detail in **The Scatter Plot** section and in **Chapter 15 Quick Reference**; however, if there is a graph or plot specific function within the viewer, it will be discussed in the corresponding section. Generic visualizations are mentioned first; genomic specific visualizations follow after.

The Scatter Plot

A scatter plot is one method used to visually represent the contents of a spreadsheet where each point in the scatter plot corresponds to a specific row in the spreadsheet. The Partek scatter plot can be 2 or 3 dimensional and can plot individual columns or high dimensional projections using linear projections such as principal components analysis (PCA) or non-linear projections such as multidimensional scaling. The scatter plot objects are 3 dimensional with special effects such as lighting, perspective, and opacity.

Invoking a Scatter Plot

To invoke a scatter plot, select **View > Scatter Plot** from the Partek main window (Figure 6. 1), or click the accelerator button on the tool bar (Figure 6. 2).

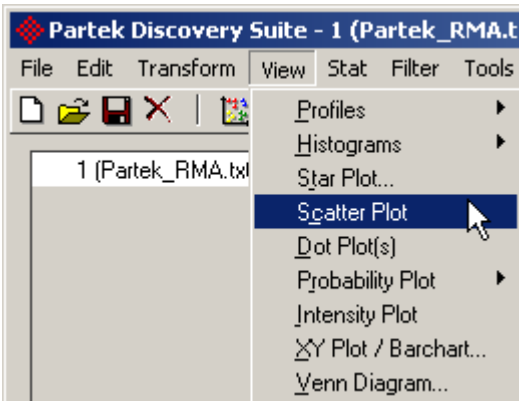


Figure 6. 1: Selecting the Scatter Plot menu option

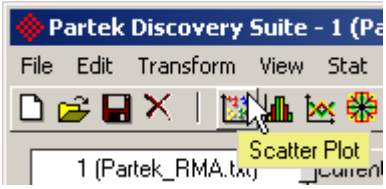


Figure 6. 2: Selecting the Scatter Plot accelerator button

To invoke a principal components analysis (PCA) scatter plot, select **Tools > Discover > Principal Components Analysis**. After clicking on **Compute,** click on **Bi-plot**.

Visualization of Multivariate Data

The most common visualization in the scatter plot is a principal components analysis (PCA) projection of the numerical data. The PCA projection maps high dimensional data to 3 dimensions for visualization (see **Chapter 7 Advanced Dimensional Reduction** for more information). The X, Y, and Z axes are PC #1, PC #2, and PC #3, respectively (Figure 6. 3).

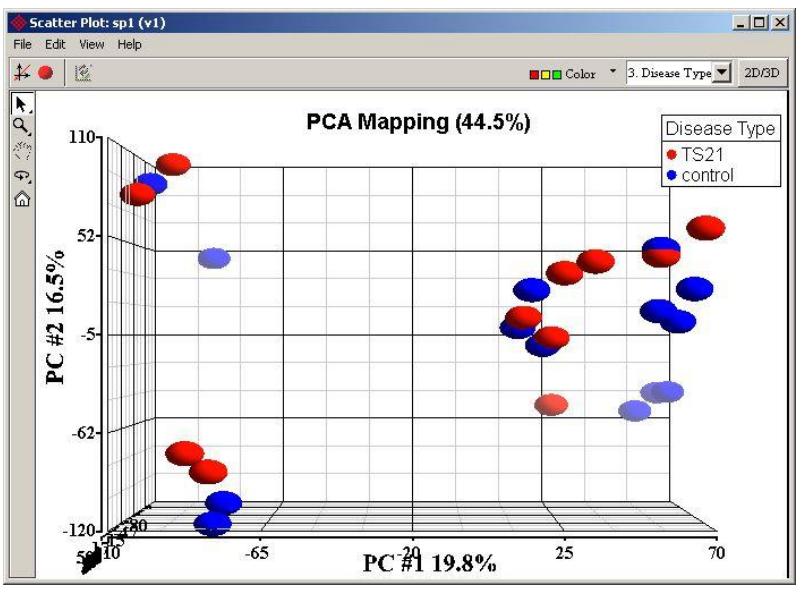


Figure 6. 3: Viewing a PCA scatter plot

In the scatter plot, each point represents one row in the spreadsheet, which usually is a sample or an experiment.

Behavior with Selected Columns

If 1, 2, or 3 columns are selected when the scatter plot is invoked, the scatter plot will be drawn with the selected columns. If there is only one column selected, the values of the column will be on the X-axis, and the Y-axis values will represent the row number of each sample. If no columns are selected, or more than three columns are selected, then a PCA scatter plot will be drawn provided there are more than three numeric response variables in the spreadsheet.

Refresh

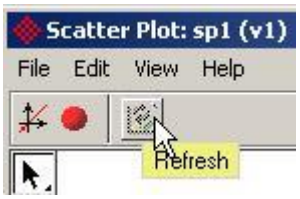


Figure 6. 4: Selecting the Refresh accelerator button in the scatter plot viewer

Applying a row filter to the spreadsheet will not cause the PCs to be recomputed, but it will cause the indicator in the *Standard Toolbar* to become active (Figure 6. 4). Clicking the *Refresh* accelerator button ( ) within the scatter plot viewer or applying the *Configure Plot* dialog will recompute the PCs. Filtering or deleting columns or deleting rows will cause the PCs to be recomputed. If variables rather than PCs are plotted, then clicking the **Refresh** button will update the axis minimums and maximums.

Viewing the Scatter Plot Results

Figure 6. 5: Viewing the file menu items

Saving the Image in the Viewer

The scatter plot can be saved as any one of the following image files:
JPEG – JPEG Files
GIF – CompuServe GIF
PNG – Portable Network Graphic
PPM – Portable Pixel Map
SVG – Scaleable Vector Graphic (publication ready format)
TIF – Tagged Image File Format

Importing Clusters

If you have clustered the rows of this data set and saved the results in the *Partek Cluster Format*, you can view the clusters in the scatter plot. This is the same as selecting **View > Scatter Plot** from the cluster set viewer (see *Chapter 8 Hierarchical and Partitioning Clustering* for more information on clustering).

Configuring the Scatter Plot

The *Configure Plot* dialog can be invoked from the *Edit* menu, or it can be invoked from the accelerator button on the viewer tool bar (Figure 6. 6).

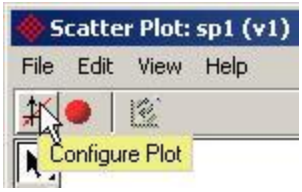


Figure 6. 6: Selecting the Configure Plot accelerator button

The *Configure Plot* dialog configures the values that are plotted as well as the range and scaling of all the axes.

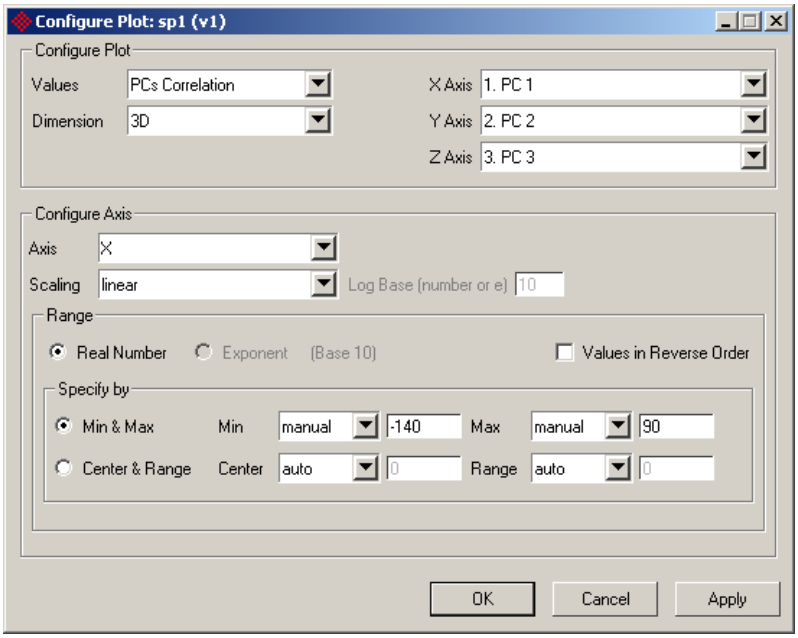


Figure 6. 7: Configuring the Configure Plot dialog

Configuring the Scatter Plot Values, Dimensions, and Axes

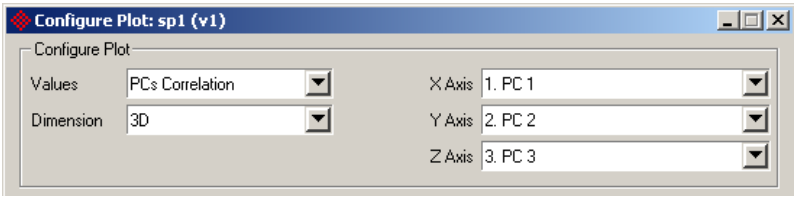


Figure 6. 8: Configuring the Configure Plot panel

In this *Configure Plot* panel, the content of the plot is configured. Some scatter plots will not have a *Configure Plot* panel (Multidimensional Scaling, for example).

Values

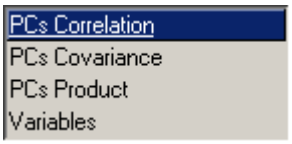


Figure 6. 9: Viewing the values drop-down list

The default PCA projection method is the *correlation matrix*.

Dimension

When the dimension is set as *3D*, all X, Y, and Z axes will contain a column or PC from their respective drop-down list. For 2D plots only, X and Y will be drawn and the viewer perspective will be turned off. Clicking the *2D/3D* button on the scatter plot tool bar will toggle this value.

Configuring the Scatter Plot Axis

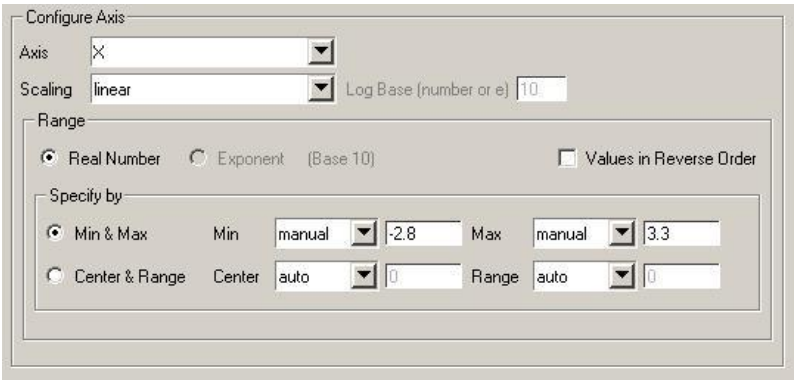


Figure 6. 10: Configuring the Configure Axis panel

The scaling and range of axes can be set one axis at a time, or they can be set as the same when *All* is selected (Figure 6. 11).

Figure 6. 11: Viewing the Configure Axis drop-down list

By default, the scaling on all the axes is linear. Fold scale is designed for columns that hold ratios. Axis labels that would be between 0 and 1 (exclusive) are shown as the negative inverse. The axis label that would be 1 is shown as “N/C” (no change). When the scaling is log, the log base can be set as either a *number* or *e*. If the scaling is non-linear and all values for the given axis are negative, then an error will be generated and the plot will remain linearly scaled, otherwise the values less than or equal to zero will simply not be shown. If scaling is set to *fold*, then the points will remain in the same place, but axis labels that would be negative are replaced with a dash. If the axis is log scaled then the points with negative values will not be shown.

When the scaling is *log*, the range of the axis can be specified as either *Exponent* or *Real Number*.

Configuring the Scatter Plot Range

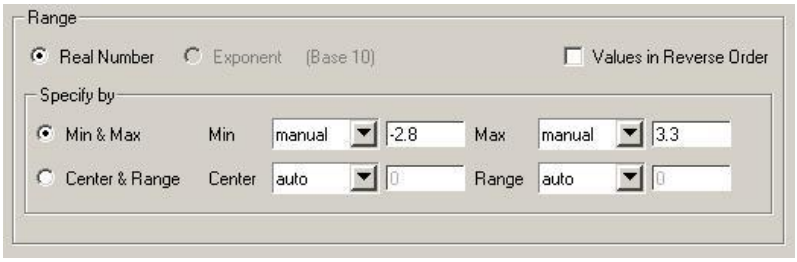


Figure 6. 12: Configuring the Range panel

The range of the axis can be specified by *min* and *max* or *center* and *range.* Set the parameters to *manual* when first editing them. The axis can be drawn in reverse order by checking the **Values in Reverse Order** button.

Scatter Plot Properties

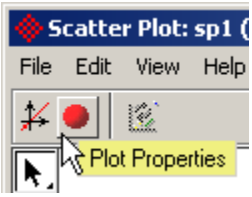


Figure 6. 13: Selecting the Plot Properties accelerator button

The *Plot Properties* dialog can be invoked from the *Edit* menu or it can be invoked from the accelerator button on the tool bar (Figure 6. 13).

Note: Properties that are changed within the scatter plot viewer only change the properties within that particular viewer. If you wish to change the properties of all the viewers within Partek globally, you can do that under **Edit > Preferences.** If you wish to create a new global color palette, you can do that under **Tools > Color Palette Manager**.

Style

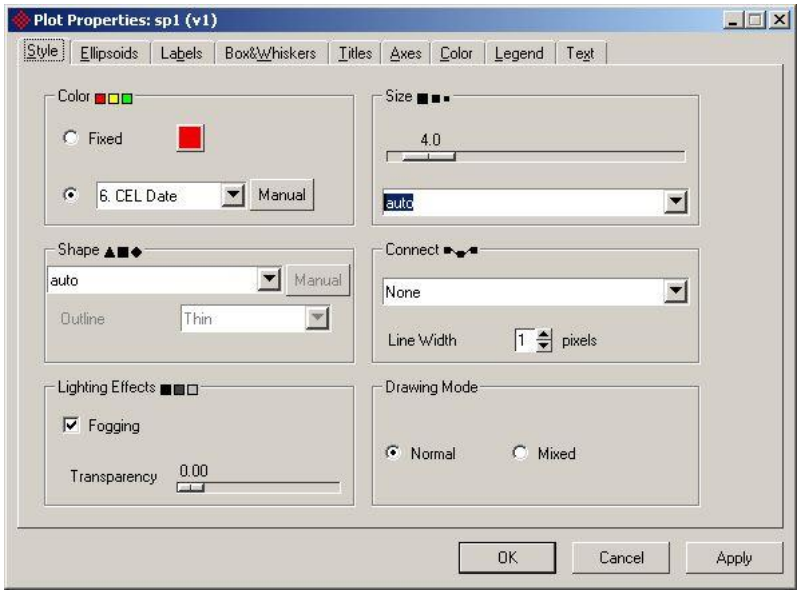


Figure 6. 14: Configuring the Style page

Configure the appearance of the points on the *Style* page.

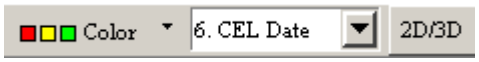


Figure 6. 15: Viewing the Style options on the viewer toolbar

Color, size, shape, and connecting lines can be configured on the toolbar within the scatter plot viewer.

Color

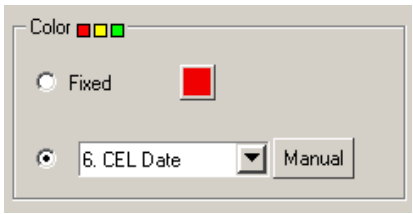


Figure 6. 16: Configuring the Color panel

By default, the color of the points is determined by the class variable. If the spreadsheet does not have a class variable then the points will be colored by the first column. The points can be colored by any column or by all the same color. If the column is a categorical variable, each category has a distinct color based on the categorical color palette. If the column is a numeric variable, the color is based on the continuous color palette. The color palette for the plot is configured on the *Color* page of the *Plot Properties* dialog.

To choose the colors of the plot select the **Manual** button as shown in Figure 6. 16.

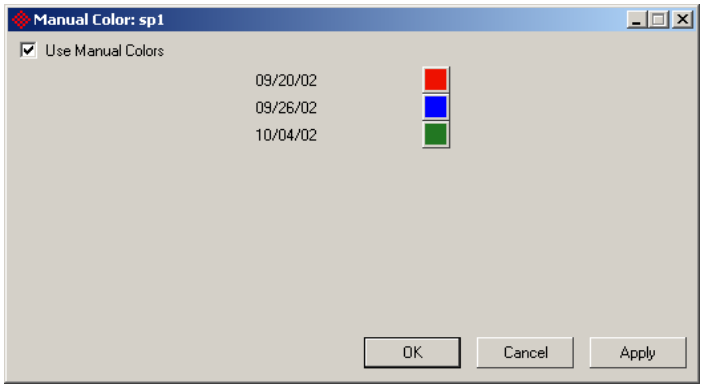


Figure 6. 17: Choosing to manually color by Type

Size

Figure 6. 18: Viewing the Size panel

By default, all the points in the plot are the same size. The slider bar determines the size of the points. If *Size* is set to **auto**, then the size of the points will be based on the number of points in the scatter plot. The size of a point can also be determined by the value in a specified column. The slider bar determines the size of the middle point.

When sizing by a numeric column, the legend lists the minimum, middle, and maximum values. When sizing by a categorical column, the legend lists the distinct values in the column and the relative size of each.

Shape

Figure 6. 19: Viewing the Shape panel

The shape of all the points will be the same when *auto*, *point*, *marker*, *tetrahedron*, *cube*, *octahedron*, *icosahedron*, or *sphere* is selected from the *Shape* drop-down list. The columns of the spreadsheet are below those options. When a column is selected, the shape of a point is determined by the value in the specified column. There are five shapes from which to choose.

If the specified column is a numeric variable, the range of the column is divided into four groups of equal range. The points will be shaped according to the group into which they fall. The sample with the smallest value will be drawn as *tetrahedron*, and the sample with the largest value will be drawn as *icosahedron*. The legend will show the range for each shape.

If the specified column is a categorical variable and has four or fewer shapes, each category will have its own shape. If the number of categories exceeds 5, the shapes will be reused. If there are 10 or fewer categories, the legend will explain the shape of each category. If there are more than 10 categories, ellipses will be used to indicate that a given shape has more categories than are listed.

Outlines can be applied to 2-D shapes only, such as the *triangle, square, diamond, hexagon*, or *circle*.

Categorical columns can be manually shaped. First, change the *Shape* combo box to a categorical column, then press the **Manual** button.

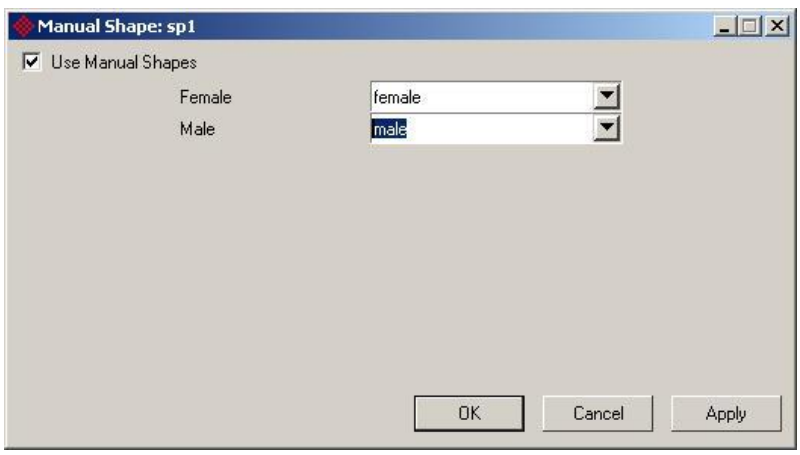


Figure 6. 20: Selecting shapes to use

From this dialog, you can choose the shape for each group. In addition to the fixed shapes available from the *Plot Properties* dialog, gender shapes are available.

Connecting Points

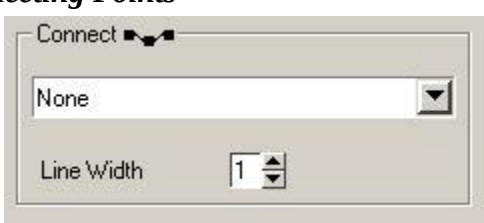


Figure 6. 21: Viewing the Connect panel

A line can be drawn among points that have the same value in the specified column. This is useful when looking at samples from the same subject.

Drawing Mode

Figure 6. 22: Selecting Normal mode in the Drawing Mode panel

In *Normal* mode, each point is drawn as the defined shape. In *Mixed* mode, only selected points are drawn as the defined shape, all other points are drawn as small dots. *Mixed* mode is faster.

Lighting Effects

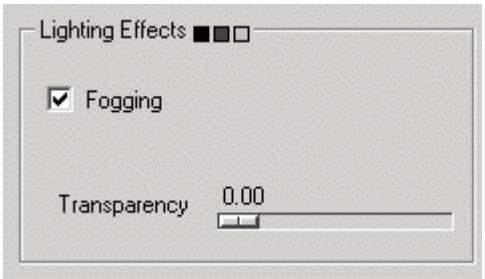


Figure 6. 23: Selecting Fogging in the Lighting Effects panel

Fogging can be turned on and off by the check box. If fogging is on, the points far away look as if they are disappearing into the fog. The slide bar determines the *Transparency* of the color applied. It is completely opaque when it is set to 0.0 and completely transparent when it is set to 1.0. This is useful for viewing a few selected points out of a large number. The selected points are always drawn opaque, which makes them visible through the large cloud of transparent points.

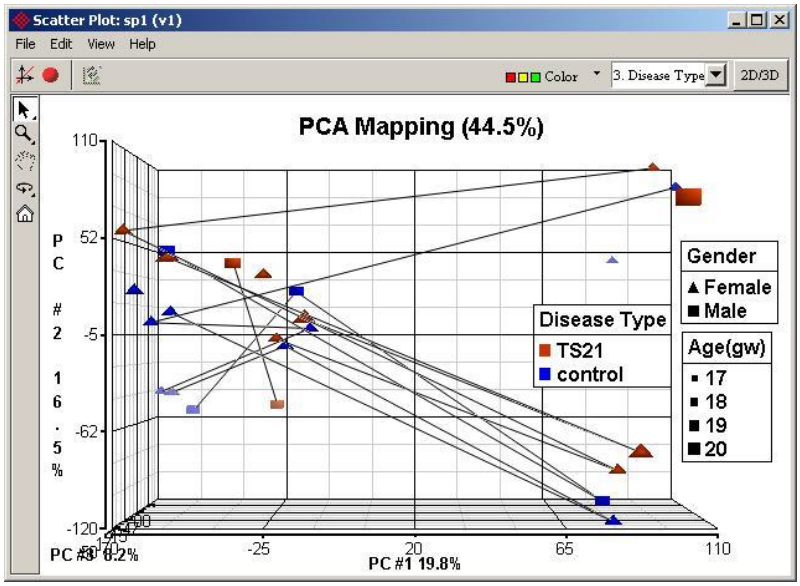


Figure 6. 24: Viewing a PCA Plot Rendering Result (Color represents Type — Normal or TS 21, Shape represents Gender — Female or Male, Size represents Age — from 17 to 20, and the line connects samples from the same subject).

Viewing Error Plots

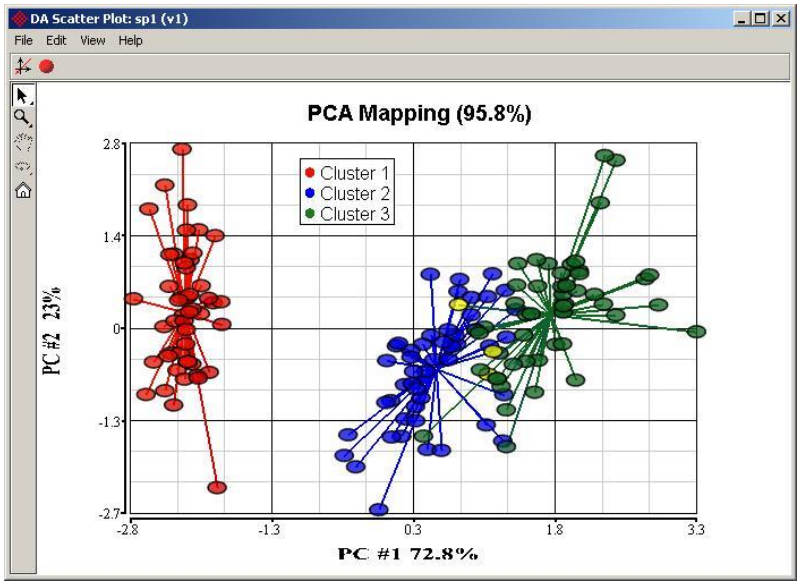


Figure 6. 25: Viewing a Discriminant Analysis Scatter Plot colored by Posterior Probability with Errors colored yellow

The Discriminant Analysis and Multiprototype Classifier scatter plots have additional color options. *Coloring By Cluster* assigns a color from the categorical color map for each cluster. Coloring using *Posterior Probabilities (DA)* or *Fuzzy Membership (MPC)* will color the points by cluster, but will interpolate the colors, blending the appropriate clusters.

Also, instead of a *Connect* panel, the Discriminant Analysis and Multiprototype Classifier scatter plots have an *Error Coloring* panel. Each category of the class column is assigned a cluster. Any row for which the class and cluster do not match is considered an error. If *Error Coloring* is set to *None*, the points will be colored according to the method in the *Color* panel. If *Error Coloring* is set to *Fixed,* the points will be drawn using the selected color.

Adding Ellipsoids and Centroids to the Scatter Plot

Adding ellipsoids to the plot is a way to look at the distribution of categorical variables.

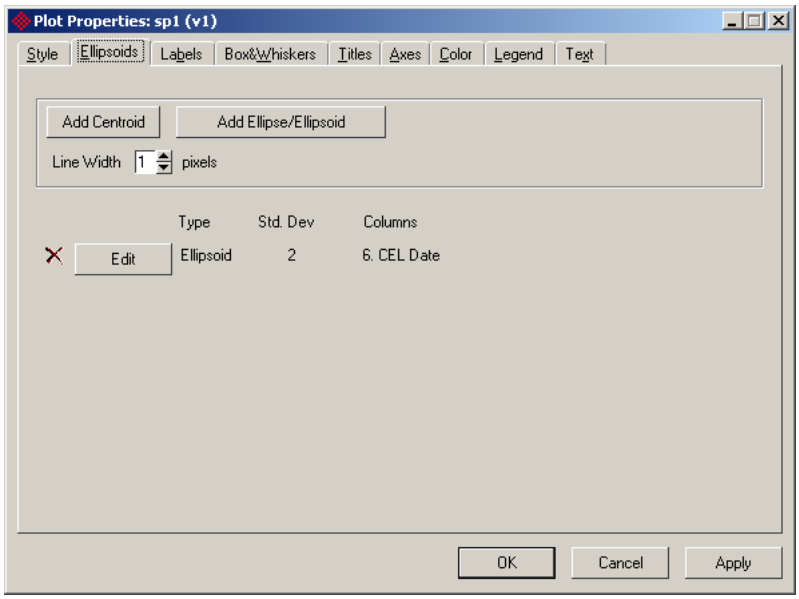


Figure 6. 26: Configuring the Ellipsoids page

Selecting the **Add Centroid** button (Figure 6. 26) will invoke the *Add Centroid* dialog (Figure 6. 27).

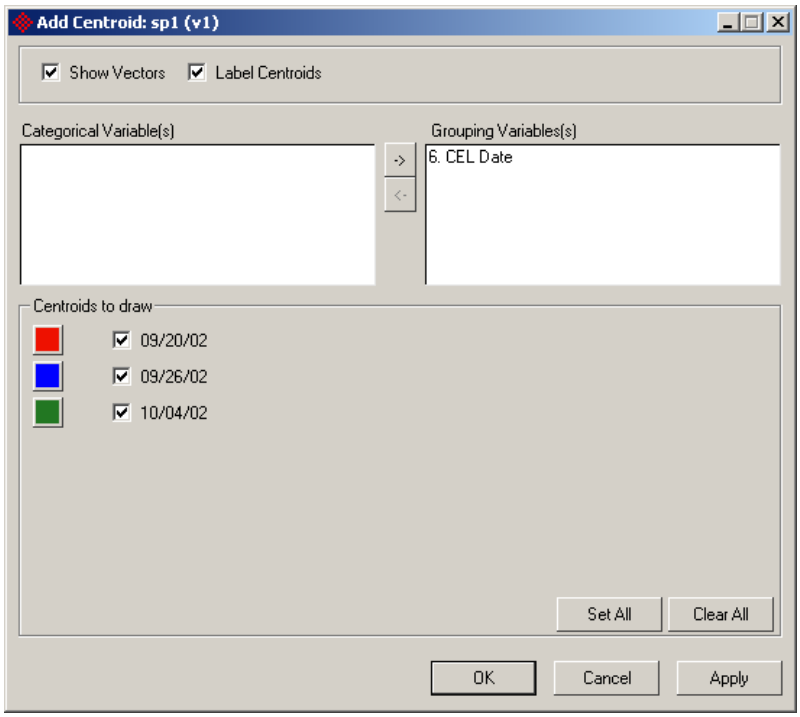


Figure 6. 27: Configuring the Add Centroid dialog

Vectors can be shown connecting each centroid to all points from which that centroid was derived. It is also possible to label each centroid.

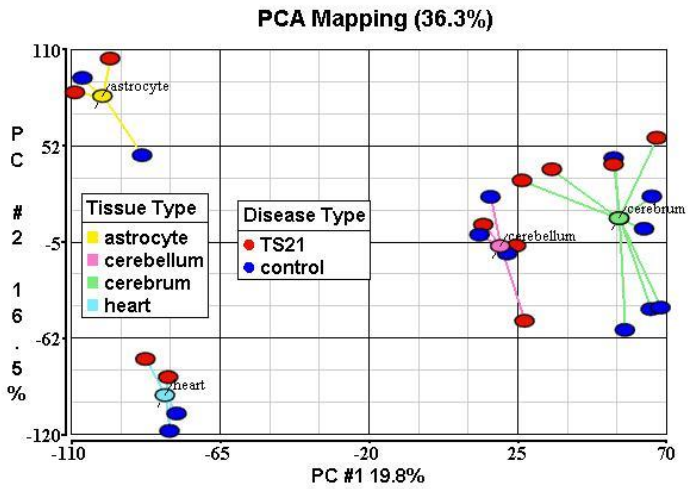


Figure 6. 28: Viewing Centroids on one factor

Selecting the **Add Ellipse/Ellipsoid** button from the *Plot Properties* dialog will invoke the *Add Ellipse/Ellipsoid* dialog (Figure 6. 29).

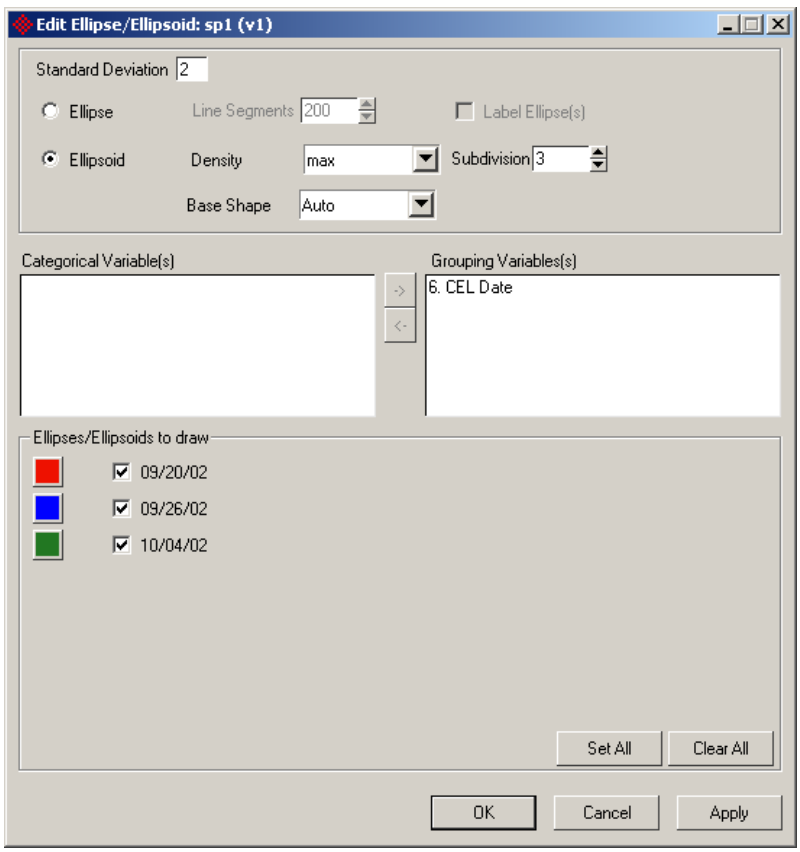


Figure 6. 29: Configuring the Add Ellipse/Ellipsoid dialog

Standard Deviations determine how far the ellipsoid will go along the axes. If the data is normally distributed, about 99% of the data points will fall in to the ellipsoid with the standard deviation as 3.0.

The font for *Centroid* and *Ellipse* labels is set on the *Point Labels* tab.

If no grouping variables are chosen, the ellipse/ellipsoid will be drawn using all samples.

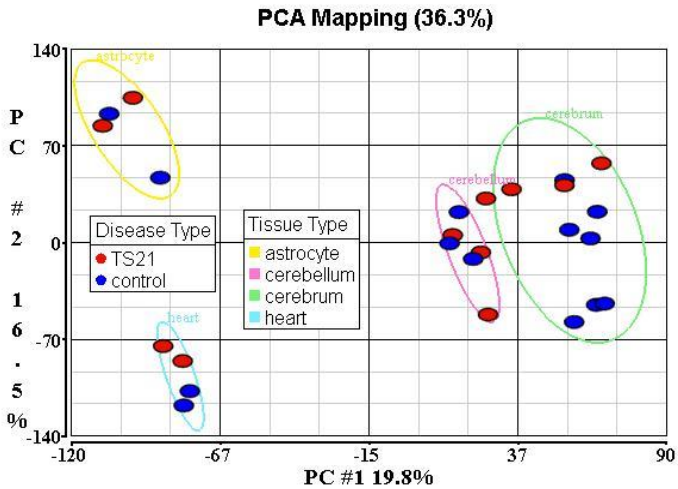


Figure 6. 30: Viewing an Ellipse for one factor

The number of *Line Segments* determines the quality of the ellipse. More lines results in better quality, but will render more slowly.

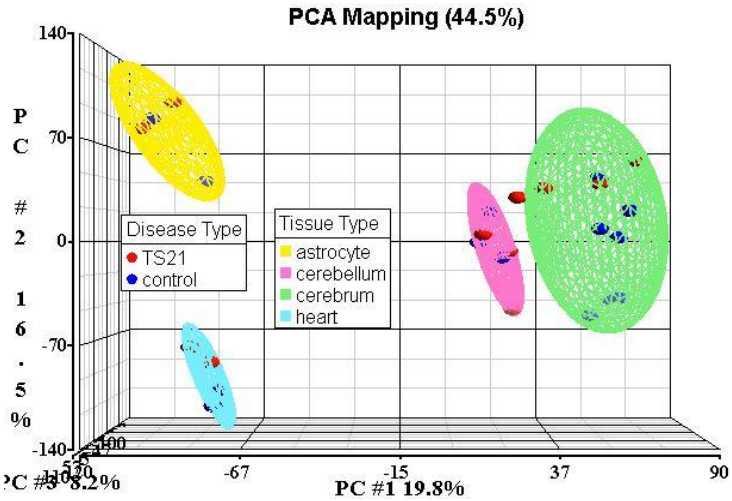


Figure 6. 31: Viewing an Ellipsoid for One Factor (each ellipsoid represents one level of Tissue factor and so does the color of the ellipsoid)

The ellipsoid is the 3-D version of the ellipse. Each ellipsoid is composed of tetrahedrons. *Subdivision* is the number of times each triangle is subdivided into 4 sub-triangles and *Density* is how many triangles will be drawn in each round of subdivision.

With *Base Shape* set to *Auto*, the base shape of the ellipsoid will be determined by the variation in the group that the ellipsoid surrounds. This means that the largest

ellipsoid will use the densest shape (icosahedron) as the base shape and the smallest ellipsoid will use the sparsest shape (tetrahedron).

To choose variables, select them from the *Categorical Variable(s)* list and click the -> button to move the selected items into the *Grouping Variable(s)* list. To remove a variable, select it in the *Grouping Variable(s)* list and click on the <- button. When the ellipsoid wire mesh represents only one categorical variable, each represents a level of the factor, and so will the color.

If the ellipsoid is drawn on the same variable that the plot points are colored by, then the ellipsoids will be drawn using the same colors. If the variables are different, the colors for the ellipsoids will be taken from the end of the color map (moving towards the start). Changing the point color column/method will update the ellipsoids color as appropriate unless you change the ellipsoid colors.

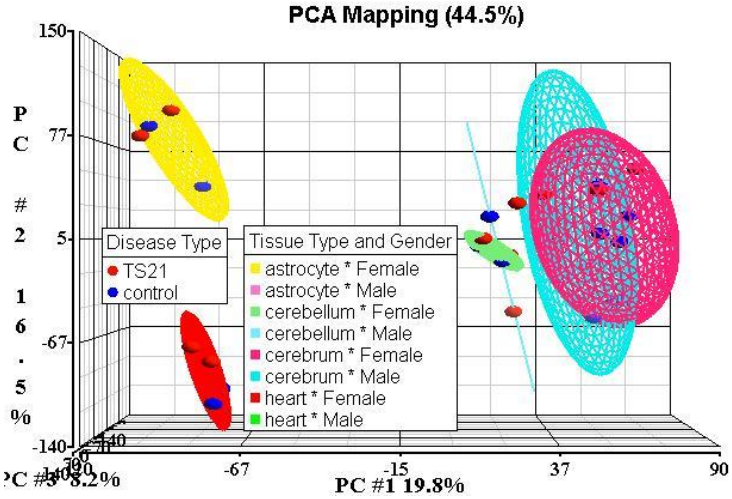


Figure 6. 32: Viewing an Ellipsoid for Two Factor Interactions (each ellipsoid represents an interaction of Tissue vs. Gender levels, so does the color of the ellipsoid).

Clusters

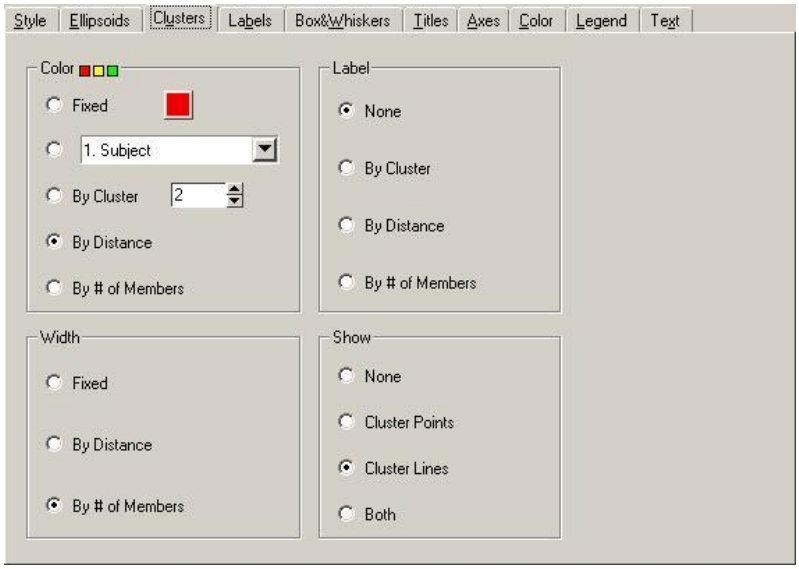


Figure 6. 33: Configuring the Clusters page

If the scatter plot contains cluster information, then the *Clusters* tab configures how the clusters are shown. The location of the cluster is determined by projecting the centroid of the cluster into PCA space.

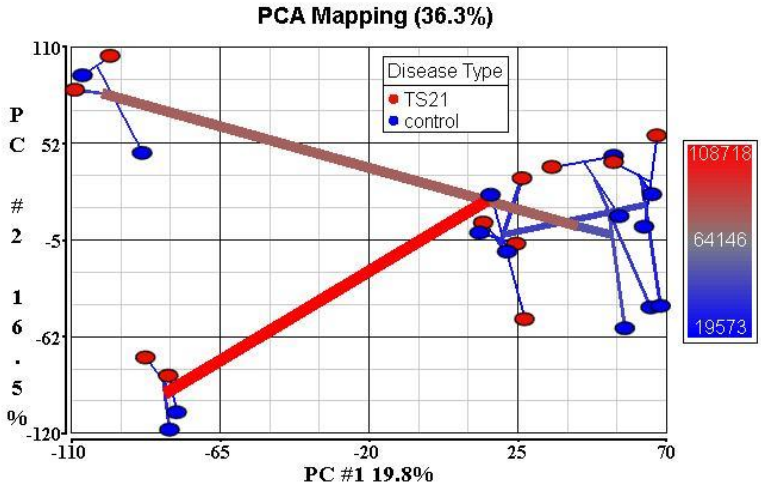


Figure 6. 34: Viewing a Cluster scatter plot

Labels

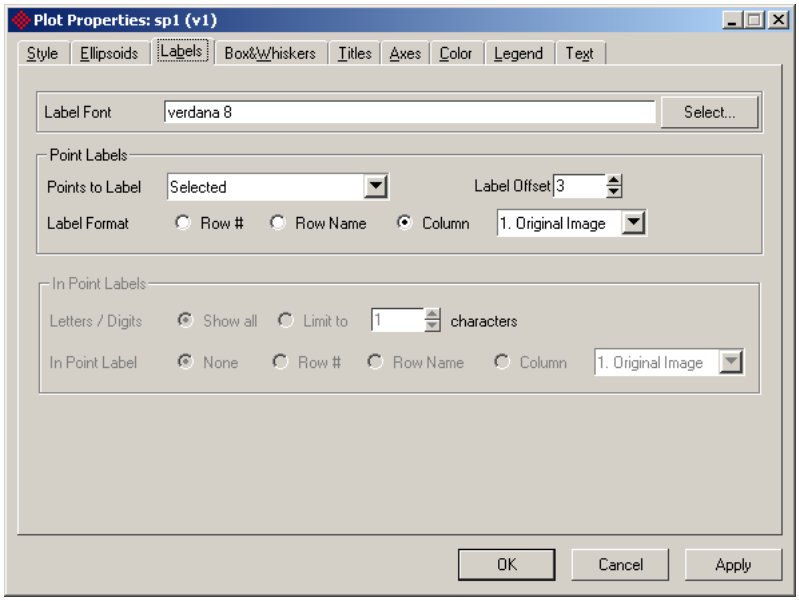


Figure 6. 35: Configuring the Labels page

The default behavior is to label the selected points only; however, all points can be labeled or none at all.

The in-point label can only be drawn if the plot is two-dimensional.

Label Format

Row # – Shows the row number
Row Name – If the row is labeled then the label will be shown, otherwise "row #" will be shown instead
Column – Shows the value in the cell associated with given row and column

Box & Whiskers

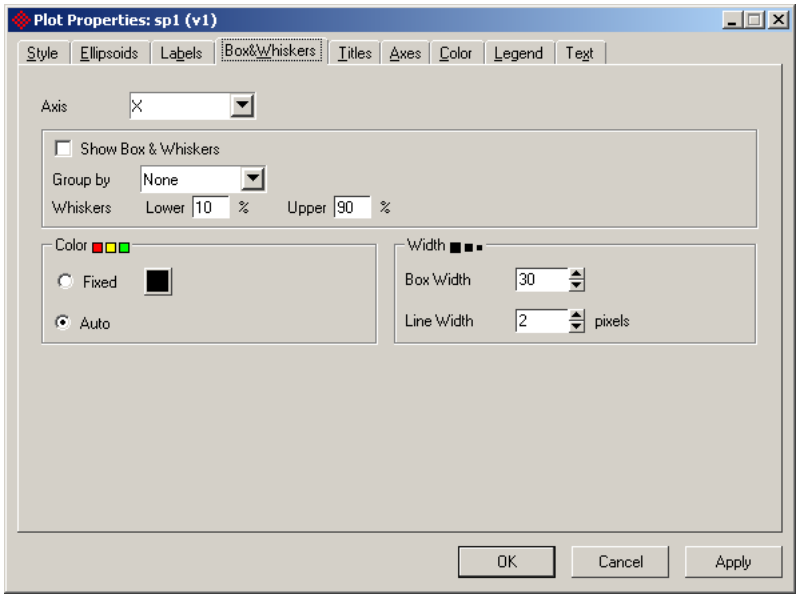


Figure 6. 36: Configuring the Box & Whiskers page

Clicking on *Box and Whiskers* will select all points of the appropriate class and within the appropriate range.

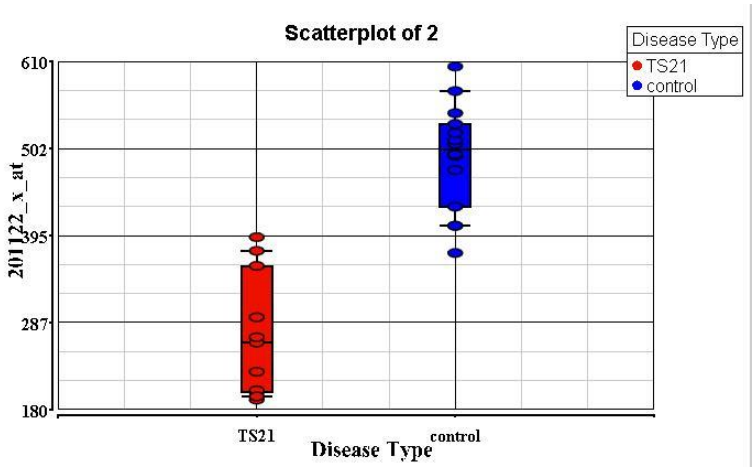


Figure 6. 37: Viewing a Box and Whiskers on a scatter plot

Titles

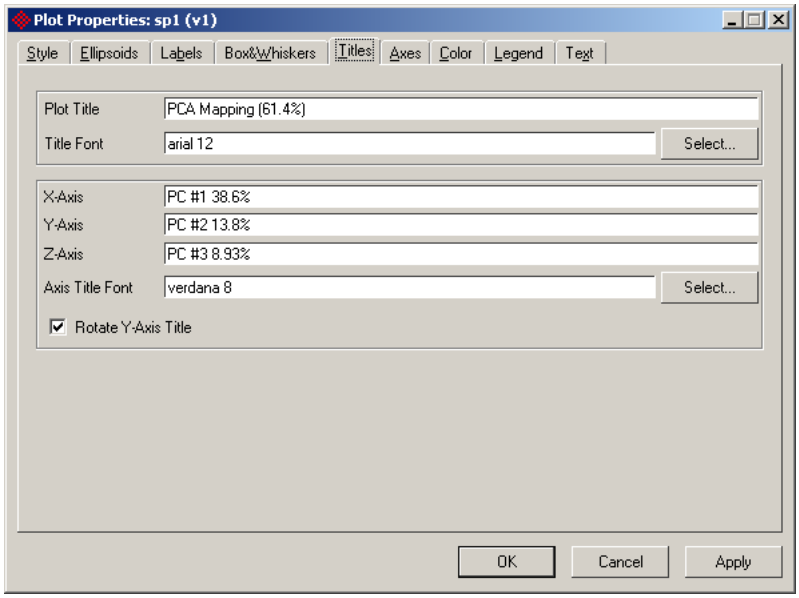


Figure 6. 38: Configuring the Titles page

The plot and axis titles are edited in the *Titles* page. If the content of the plot changes, the title will be updated, overwriting the entry.

Title fonts can also be changed here. All axes use the same axis title font.

Axes

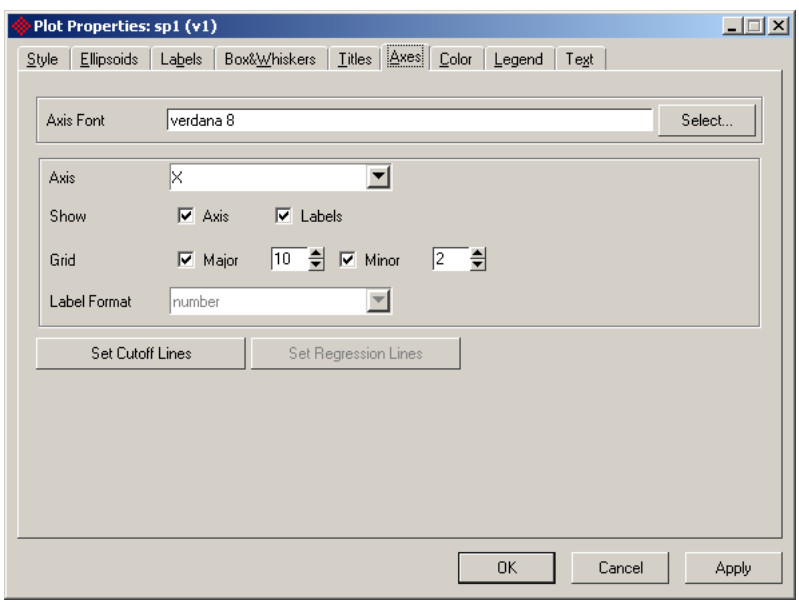


Figure 6. 39: Configuring the Axes page

The *Axis Font* controls the appearance of the numbers on all the axes.

When the plot is log scaled, the number of major ticks cannot be set manually unless the axis range spans less than one exponent. When the major grid is not shown, the minor grid will be turned off automatically.

If the scaling of the selected axis is set to *log*, the axis labels can be configured. If the label is shown as *number*, the numeric value of the tick will be shown (e.g. 100). If the label is shown as *exponent*, the value of the tick will be shown using the base and exponent (e.g. 10^2). If the label is shown as *number and exponent*, both will be shown (e.g. 10^2(100)).

Cutoff Lines

Selecting the **Set Cutoff Lines** button opens the *Set Cutoff Lines* dialog.

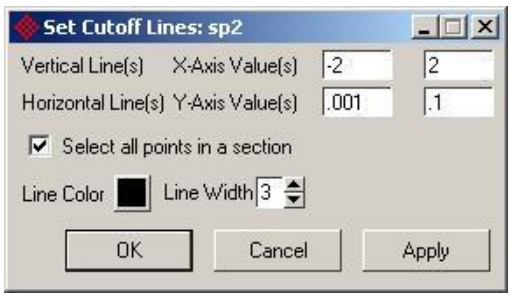


Figure 6. 40: Setting the cutoff lines

Using this dialog, you can add cutoff lines to the plot at significant values. If *Select all points in a section* is checked, then selection and the mouse-over will be determined by the values entered.

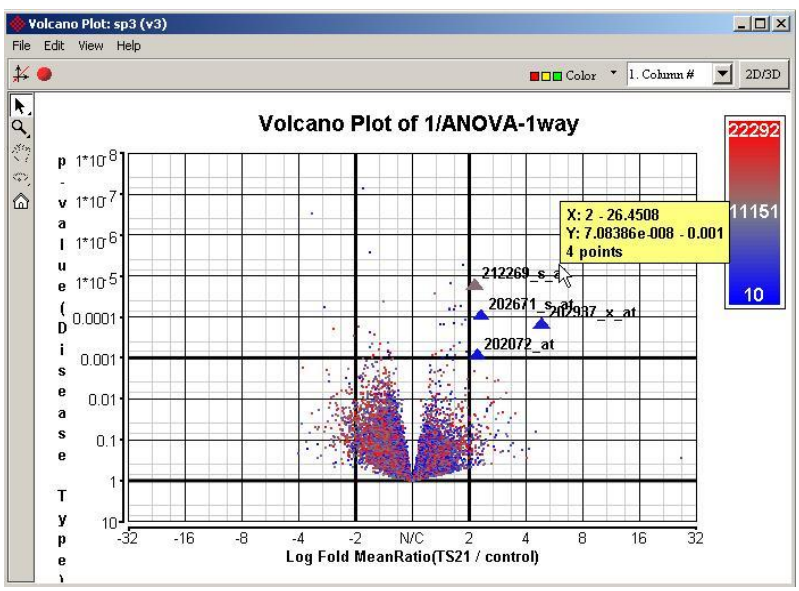


Figure 6. 41: Viewing a Volcano plot with cutoff lines set

Regression Lines

Selecting the **Set Regression Lines** button opens the *Set Regression Lines* dialog.

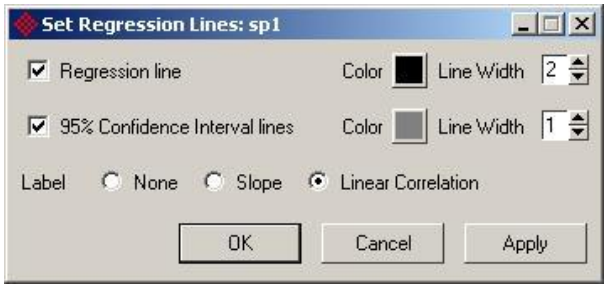


Figure 6. 42: Configuring the Set Regression Lines dialog

If the plot contains two columns, you can use the *Set Regression Lines* dialog to add regression lines to the plot. You can also add lines for the 95% confidence interval and configure the color and thickness of the lines.

The font of the label is determined by the *Label Font* on the *Point Labels* tab. The width of the regression line is the same as the width of the connecting lines on the *Style* tab.

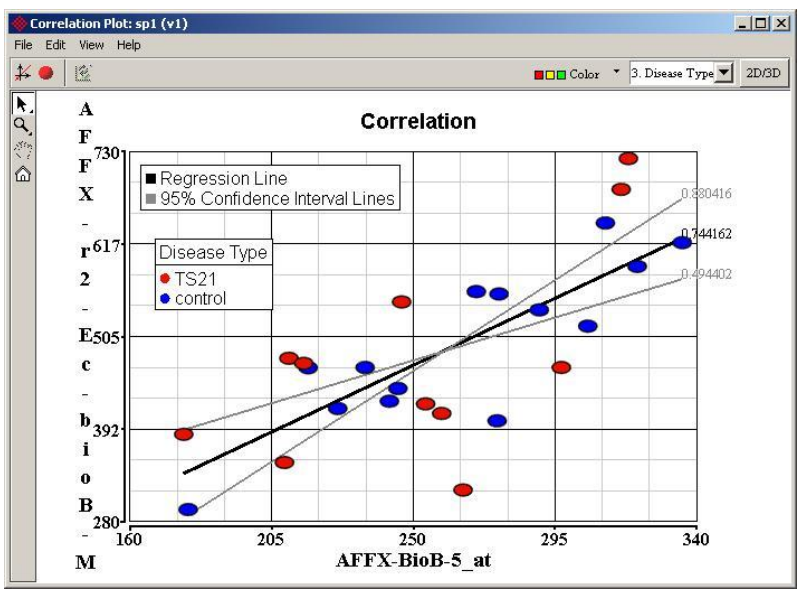


Figure 6. 43: Viewing Regression Lines

Color

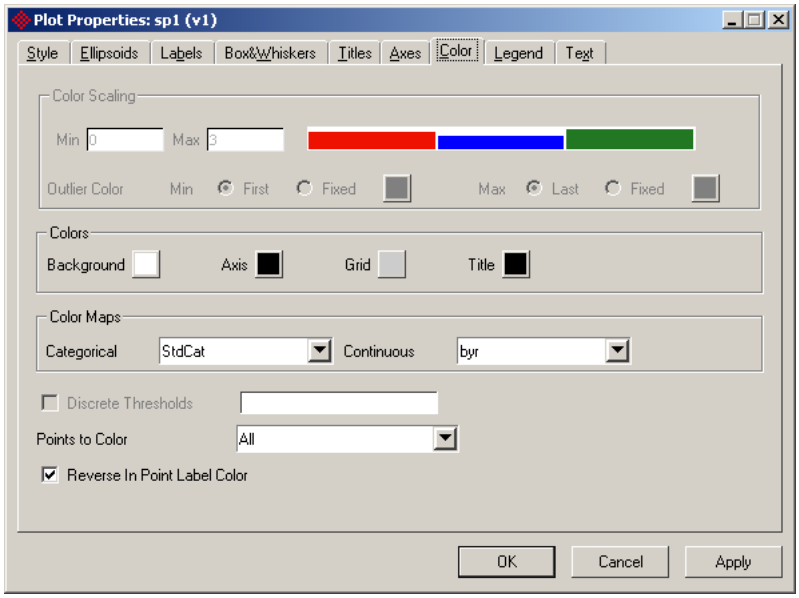


Figure 6. 44: Configuring the Color page

Color Scaling

Figure 6. 45: Configuring the Color Scaling panel

If outliers are colored using the first/last color then they will be the same color as values equal to the threshold. If outliers are colored as *Fixed*, then they will be drawn using the specified color.

Color scaling is only available if the colors are derived from the continuous color map.

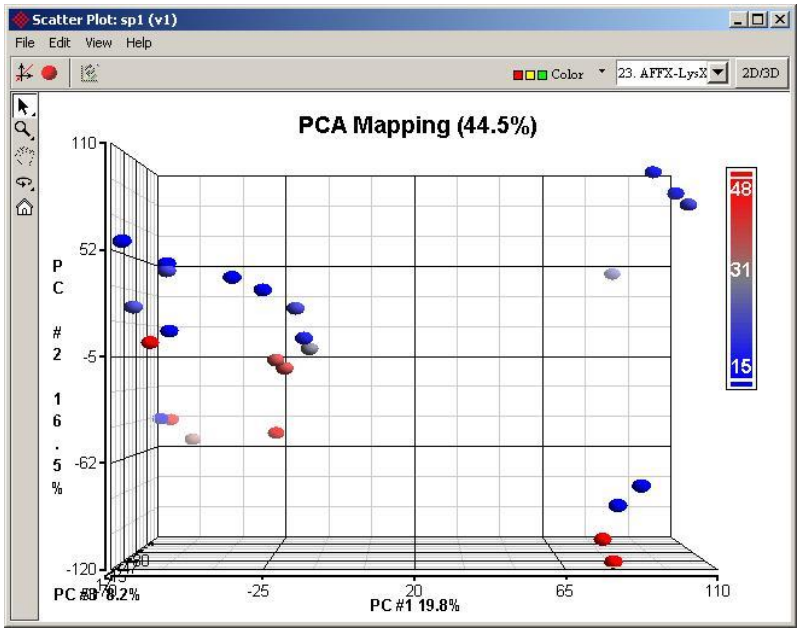


Figure 6. 46: Viewing the Color scaling

Colors

Figure 6. 47: Configuring the Colors panel

The major gridlines, ticks, point labels, axis titles, and axis labels are colored using the *Axis* color. The plot title, axis titles, point labels, and legends are colored using the *Title* color.

The colors used in the viewer can be changed by clicking on the color indicator button next to the color you want to change; a color selector palette will pop up to allow you to select a color.

The default colors can be changed by selecting **Edit > Preferences...** from the Partek main window.

Color Maps

Figure 6. 48: Configuring the Color Maps panel

The *Categorical Color Map* is used to color the categorical variables. The *Continuous Color Map* is used to color numeric variables. Color maps can be configured by selecting **Tools > Color Palette Manager...** from the Partek main window.

Advanced Options

Figure 6. 49: Configuring the advanced scatter plot color options

Discrete Thresholds are only available when coloring by a numeric column.

To activate color thresholds, click the check box and then enter one or more numbers separated by a space. The colors are derived from the *Categorical Color Map*.

By default, all points will be colored. But it is possible to color selected points or no points at all. If a point is not colored, it will have the same color as the background.

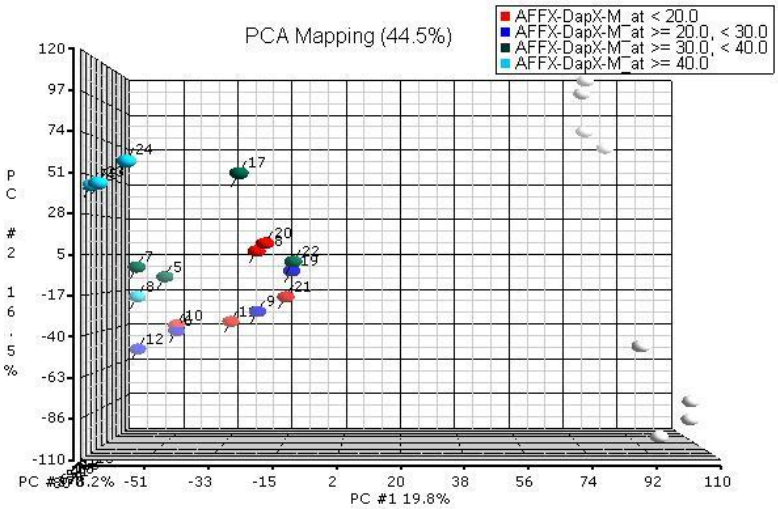


Figure 6. 50: Viewing PCA mapping with color thresholds on and points to color set to selected

If *Reverse In Point Label Color* is checked, then the in-point label is drawn in the background color, otherwise the label is drawn in the point color.

Legend

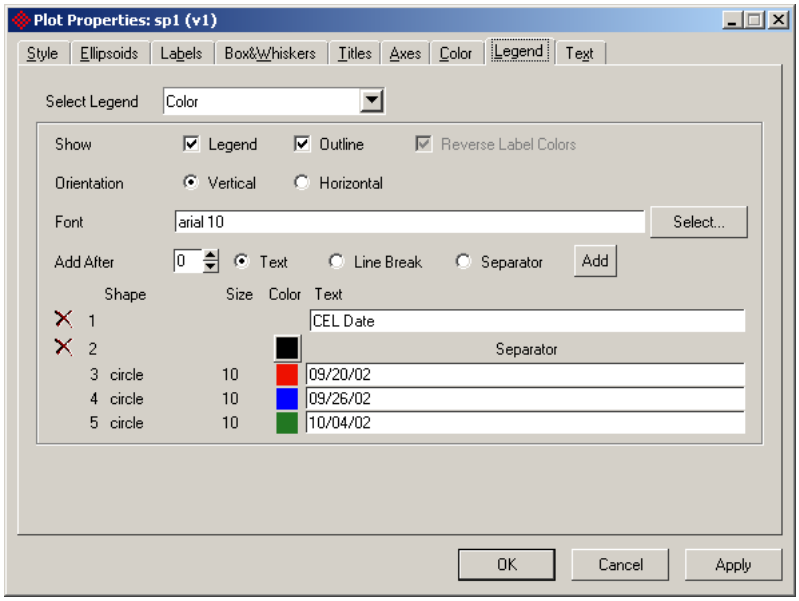


Figure 6. 51: Configuring the Legend page

The legend(s) of the plot can be edited in this dialog. First, choose the appropriate legend from the drop-down list. Lines can be added to the legend by specifying a line number and a line type (*Text*, *Line Break*, or *Separator*) then clicking **Add**.

The *Separator* draws a line extending the width (if vertical) or height (if horizontal) of the line. The *line break* moves right if orientation is *Vertical* or down if orientation is *Horizontal*.

Text

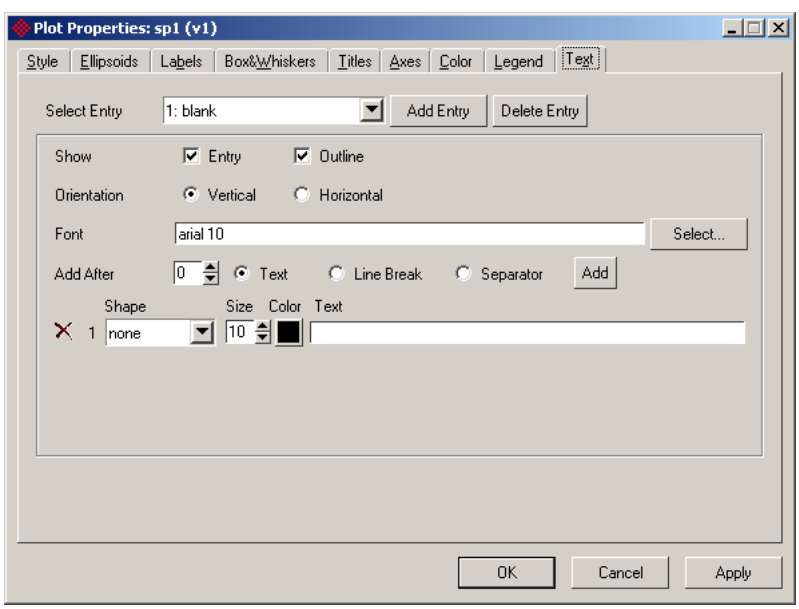


Figure 6. 52: Configuring the Text page

This page operates like the *Legend* page. On this page, text can be added to the plot. The *Select Entry* drop-down list contains the entry number and the first line.

Miscellaneous Viewer Options

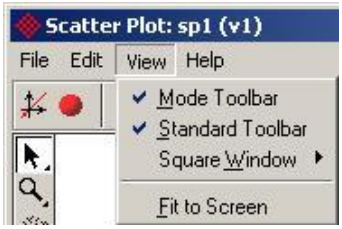


Figure 6. 53: Viewing the View menu items with the viewer

The *Mode Toolbar* is the vertical toolbar at the left side of the viewer. This option controls whether or not the *Mode Toolbar* is shown.

The *Standard Toolbar* is the horizontal toolbar at the top of the viewer. This option controls whether or not the *Standard Toolbar* is shown.

Square Window resizes the viewer either vertically of horizontally.

When the range is manually set, clicking **Fit to Screen** will adjust the range of the axes to fit the range of the points in the plot.

Figure 6. 54: Viewing the Help menu items within the viewer

Help > On-line Help will direct you to the Partek documentation.

Help > On Modes provides documentation on how to use the mode bar in the viewers. The Mode button functions are discussed below.

Changing Modes in the Scatter Plot Viewer

This section will explain how to use the *Mode Toolbar* in the viewer of the Scatter Plot, the Histogram, the Profile Plot, the Star Plot, the Intensity Plot, and the HTS Navigator (Figure 6. 55).

Figure 6. 55: Viewing the Mode toolbar in the Partek Visualization System

Changing Modes

Most of the icons in the vertical mode toolbar have multiple options that can be accessed by clicking and briefly holding down the left mouse button on the mode icon, upon doing that, a mode option menu will pop-up to the right. To select an option from the menu, drag the mouse cursor over to the desired mode option and then release the mouse button.

Common in All Modes

- The **<Home>** key resets *Rotation*, *Zoom* and *Pan Back* to their default values
- Holding down the middle mouse button (if there is one) and dragging the mouse over the visualization provides interactive rotation.
- Place the mouse cursor over a data item (without clicking) to see information about the item.

Selection Modes

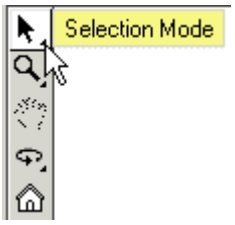


Figure 6. 56: Viewing the Selection Mode icon on the Mode toolbar

There are two selection modes, *Standard Selection Mode* (arrow) and *User-defined Selection Mode* (arrow with plus sign) (Figure 6. 57).

Figure 6. 57: Viewing the Standard Selection Mode and User-defined Selection Mode

Standard Selection Mode

- Clicking the left mouse button selects an individual item
- Holding down the left mouse button and dragging the mouse creates a box around the selected items
- **<Ctrl>** + left clicking adds the item under the mouse cursor (or in the box) to the list of selections

User-defined Selection Mode

In addition to the *Standard Selection* behavior, the *User-Defined Selection Mode* will send the selections to a user-defined process, such as the Partek Compound Viewer™ if the spreadsheet has a compound *External Link* defined.

Zoom Mode

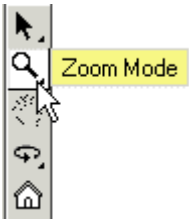


Figure 6. 58: Viewing the Zoom Mode icon on the Mode toolbar

In *Zoom Mode* left-click to incrementally zoom in, **<Ctrl>**-left-click to zoom out.

There are three options to zoom (Figure 6. 59): *Zoom in both X and Y*, *Zoom Horizontal*, and *Zoom Vertical*

Figure 6. 59: Viewing Zoom directions in the Zoom toolbar

Hold down and drag the left mouse button to create a bounding box around the items to zoom in on, and release the button to have a close-up picture of those selected items.

Pan Mode

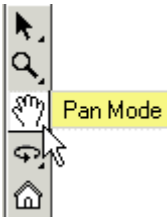


Figure 6. 60: Viewing the Pan Mode icon on the Mode toolbar

This icon is enabled only when the data is zoomed in on. Hold down the left mouse button while dragging the mouse to interactively move the data (pan).

Rotation Mode

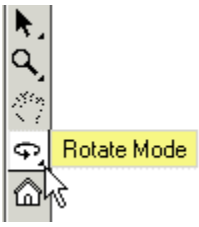


Figure 6. 61: Viewing the Rotation Mode icon on the Mode toolbar

There are two rotation modes: *Manual Rotation Mode* (one circle) and *Continuous Rotation Mode* (two circles, one is on top of another) (Figure 6. 62)

Figure 6. 62: Viewing the Manual Rotation Mode and the Continuous Rotation Mode

Manual Rotation Mode

The *Manual Rotation Mode* has the same functionality of the middle-mouse button. Hold down the left mouse button while dragging the mouse to interactively rotate the view.

Continuous Rotation Mode

Click the left mouse button to start and stop rotation. Selecting any other mode also stops continuous rotation.

Reset

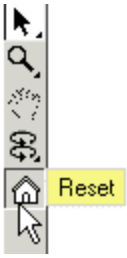


Figure 6. 63: Viewing the Reset icon on the Mode toolbar

Reset has the same functionality as the **<Home>** key. Selecting the *Reset* button will set the *Zoom, Pan,* and *Rotation Mode* back to their default values.

The Dot Plot

The Dot Plot is a 2-D view of the distributions of data. The rows of a given column are split into equal-sized bins with the number of counts in each bin represented by the number of points. By default, the counts are separated by the class column (Figure 6. 64).

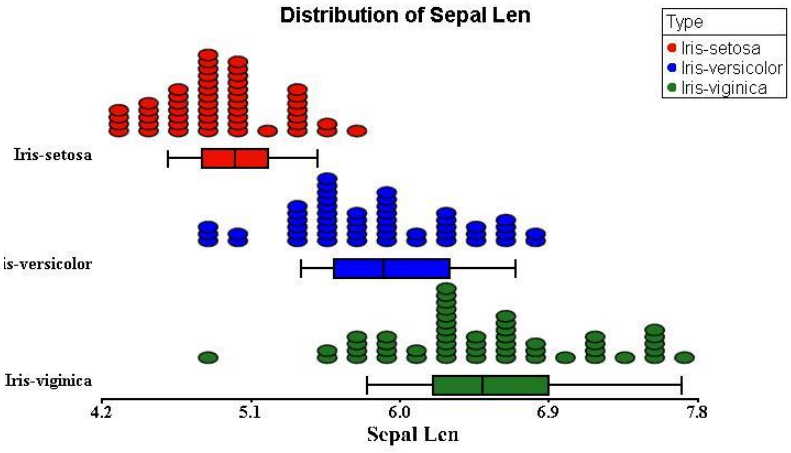


Figure 6. 64: Viewing a Dot Plot

Invoking a Dot Plot

To invoke a dot plot for each selected numeric column, click **View > Dot Plot(s)** from the Partek main window, or choose **Plot > Dot Plot(s)** from the column pop-up (Figure 6. 65).

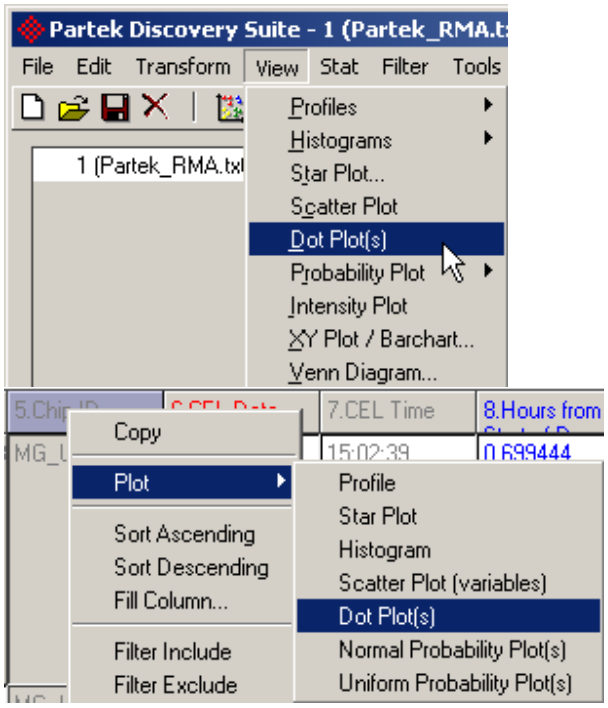


Figure 6. 65: Viewing Dot Plot menu options

Dot Plot Specific Menu Items

The File, Edit, View, and Help menus in the Dot Plot viewer behave the same as the menus in the Scatter Plot viewer. Any differences will be notated below, otherwise see the **Viewing the Scatter Plot Results** section above.

The **Edit > Plot Properties > Style, Labels, Box & Whiskers, Titles, Axes, Color,** and **Labels** in the Dot Plot behave the same as in the Scatter Plot - Plot Properties. Any differences will be notated below, otherwise see the **Scatter Plot Properties** section above.

The Mode buttons within the Dot Plot Viewer behave the same as in the Mode buttons in the Scatter Plot. Any differences will be notated below, otherwise see the **Miscellaneous Viewer Options** section above.

Configuring the Number of Bins in the Dot Plot Viewer

Figure 6. 66: Configuring the Bin spin box

The spin box determines the number of bins (Figure 6. 66). The range of each bin is equal to the range of the data divided by the number of bins. The X value of each point is determined by the beginning of the bin that the point falls in. The Y value of each point is determined by the class and the number of points with the same class in the same bin.

Dot Plot Properties

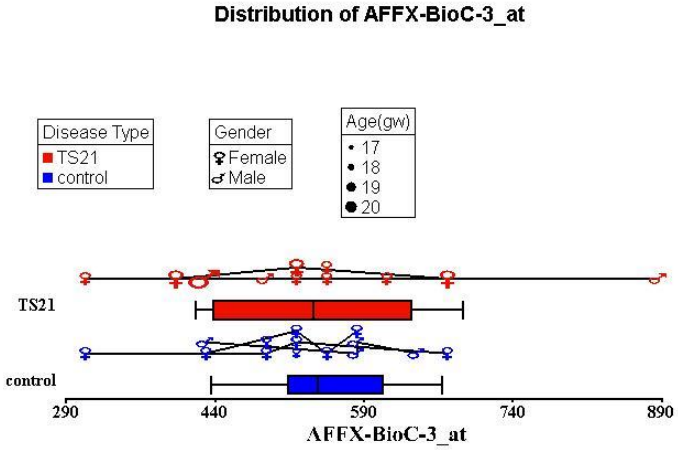


Figure 6. 67: Viewing the Dot Plot rendering result (Color represents Type — Normal or TS 21, Shape represents Gender — Female or Male, Size represents the Age — from 17 to 20 and the line connects samples from the same subject).

Color

Color scaling is only available if the colors are derived from the continuous color map.

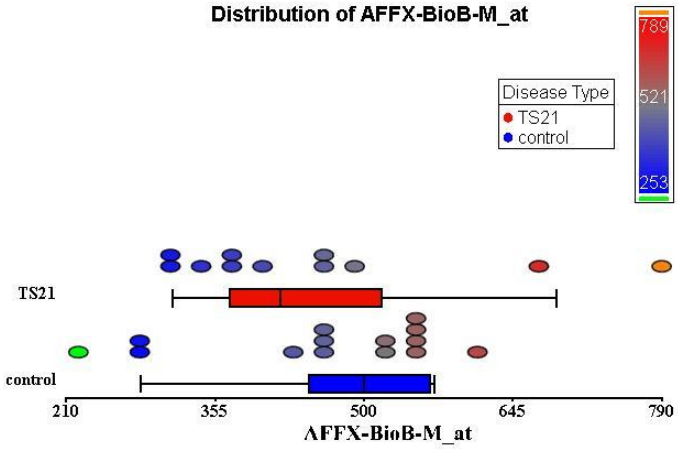


Figure 6. 68: Showing color scaling in effect

Discrete Thresholds and Points to Color

Figure 6. 69: Configuring the Discrete Thresholds and Points to Color

Discrete Thresholds are only available when coloring by a numeric column.

To activate color thresholds, click the check box then enter one or more numbers (separated by space) in ascending order. The colors are derived from the categorical color map.

By default, all points will be colored. It is also possible to color selected points or no points. If a point is not colored then it will have the same color as the background.

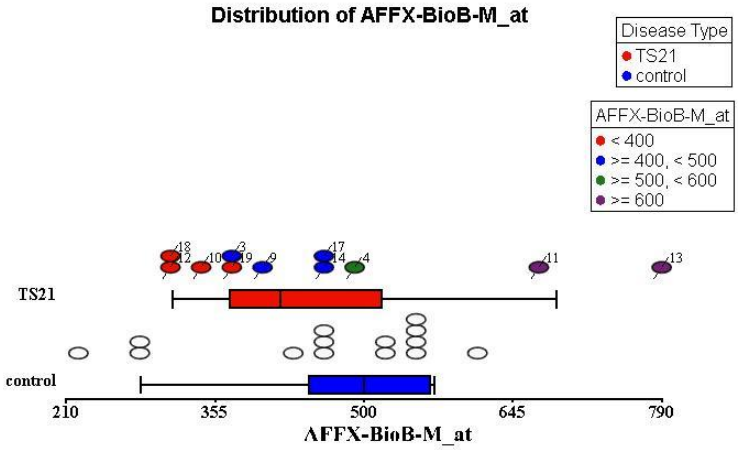


Figure 6. 70: Viewing a Dot Plot with Color Thresholds on and Points to Color set to Selected

The Histogram

The Histogram is a 2-D view of the distributions of data. The values of each variable are split into equal-size bins, and the number of counts in each bin is represented by the height of the bar. Histogram bars may be separated to represent categories of data.

Opening the Histogram

Histograms are available via **View > Histogram** in the Partek main menu (Figure 6. 71). The column histogram is also available from the accelerator button on the tool

bar (Figure 6. 72). Additionally, row and column histograms are available via the pop-up menu on rows or columns.

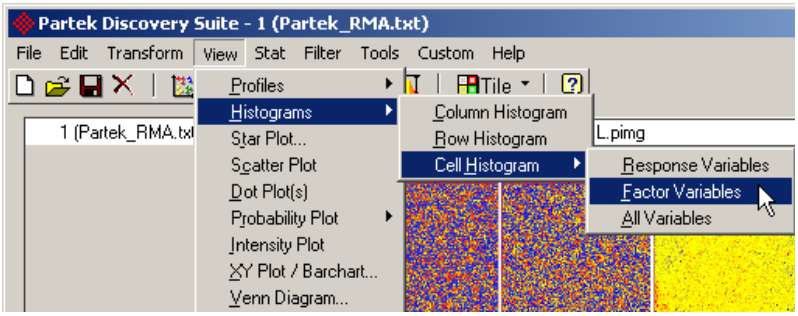


Figure 6. 71: Viewing the Histogram menu buttons

Figure 6. 72: Selecting the Histogram accelerator button

Visualizing the Distribution of Data

By default, the histogram draws the distributions of the data on the first non-string column.

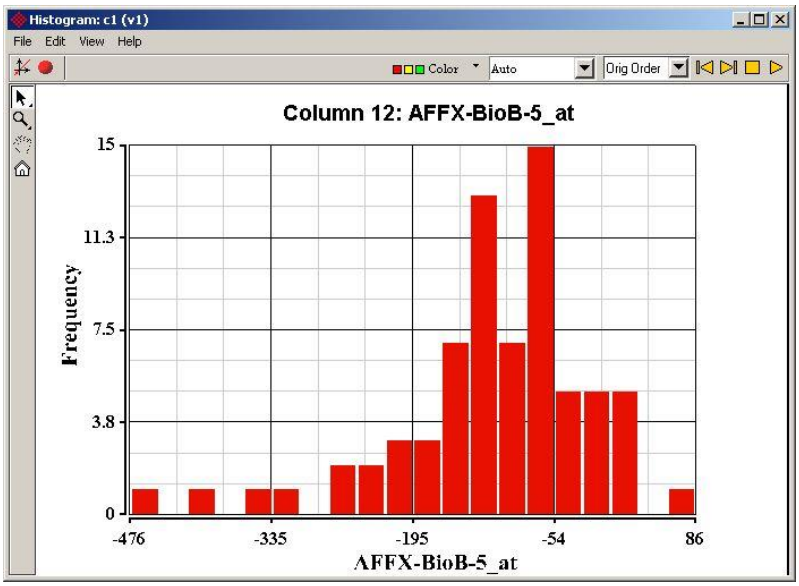


Figure 6. 73: Viewing a Histogram on a column

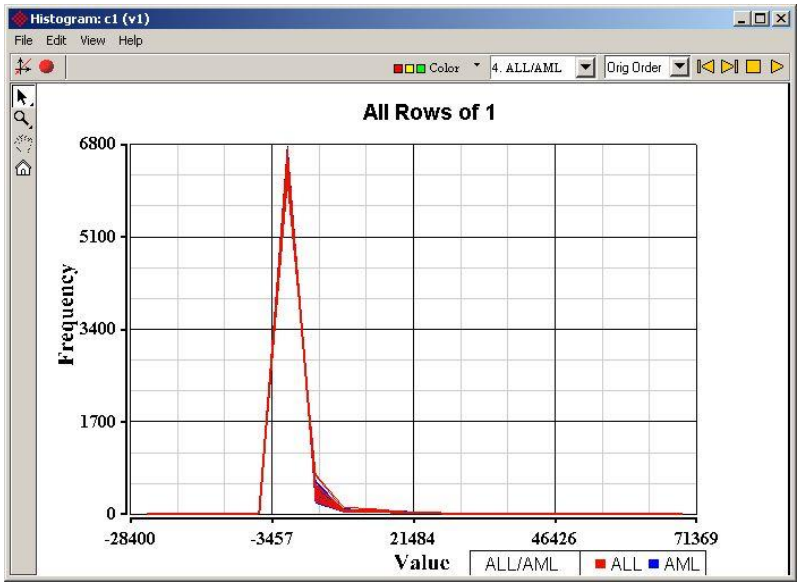


Figure 6. 74: Viewing a Histogram on rows

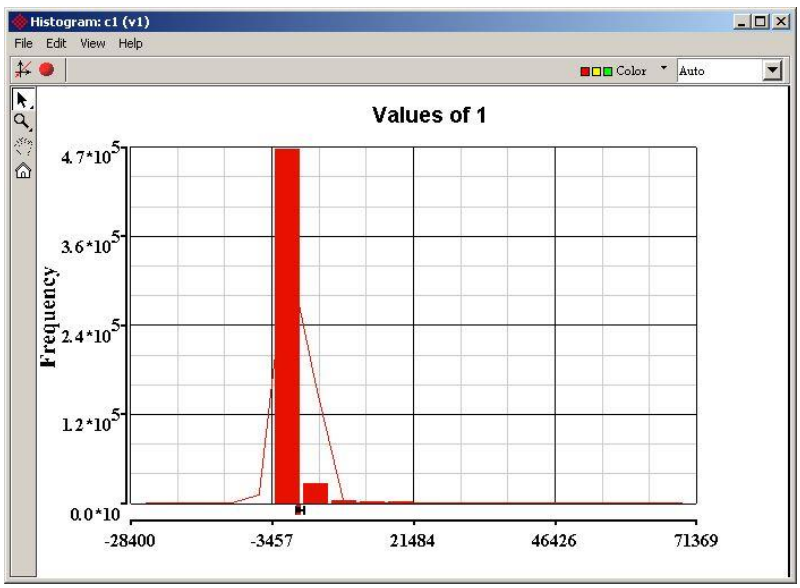


Figure 6. 75: Viewing a Histogram on response variables

By default, the values are divided into 20 bins. The values are shown on the X-axis, and the count of each value is represented on the Y-axis.

The mouseover of a bar reveals the range of the bin. The square brackets [] indicate inclusive and the parentheses () indicate exclusive. For the last bin both numbers are inclusive, for the other bins the start is inclusive and the end is exclusive.

Histogram Specific Menu Items

The File, Edit, View, and Help menus in the Histogram viewer behave the same as the menus in the Scatter Plot viewer. Any differences will be notated below, otherwise see the **Viewing the Scatter Plot Results** section above.

The **Edit > Plot Properties > Style, Labels, Box & Whiskers, Titles, Axes, Color,** and **Labels** in the Histogram behave the same as in the Scatter Plot - Plot Properties. Any differences will be notated below, otherwise see the **Scatter Plot Properties** section above.

The Mode buttons within the Histogram Viewer behave the same as in the Mode buttons in the Scatter Plot. Any differences will be notated below, otherwise see the **Miscellaneous Viewer Options** section above.

Configuring the Histogram

Range - Snip Range

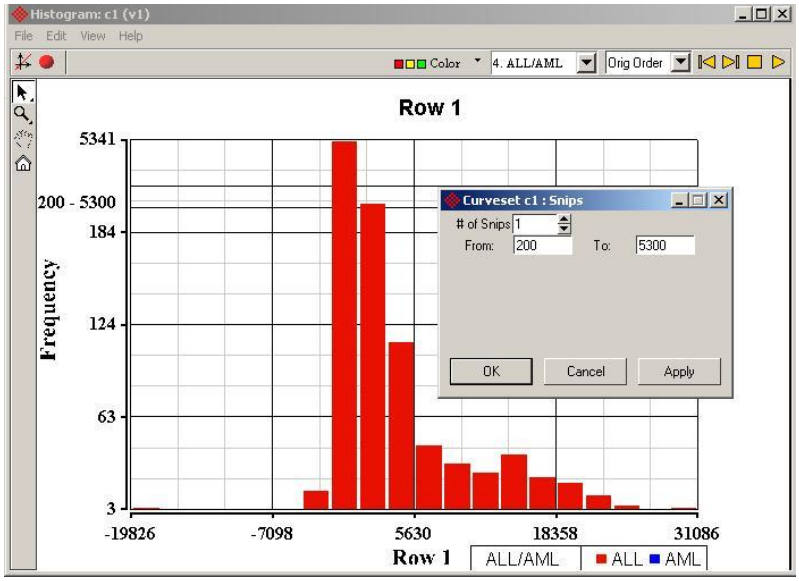


Figure 6. 76: Viewing the Snips dialog and its effect

The **Snip Range** button is available only for the Y-axis. First, enter the desired number of snips. Then, for each snip, enter the minimum and the maximum of the values that should be folded together. Any values that fall within the range will fall between the two lines near the label.

Histogram Plot Properties

Style

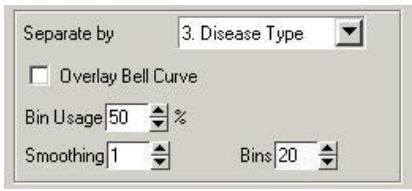


Figure 6. 77: Configuring the Histogram panel

A bell curve can be overlaid on any row histogram or any numeric column to show the normal distribution of the data. Configure **Bin Usage** to change the amount of padding between bins.

Labels

Checking **Show value label** will cause the exact frequency to appear for each point in the histogram. The x and y location of each label (relative to its point) can be configured using the two spin boxes and combo boxes.

Accumulation

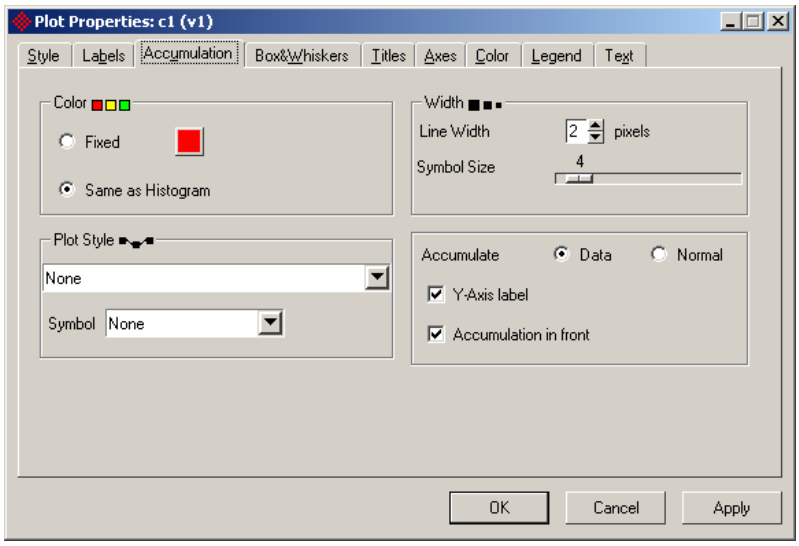


Figure 6. 78: Configuring the Accumulation page

An *Accumulation Histogram* shows the sum of all the bins before the current bin in addition to the current bin. The accumulation lines are turned off by default. To turn them on, change the line style or the shape. The accumulation lines have much of the same properties found on the *Style* page.

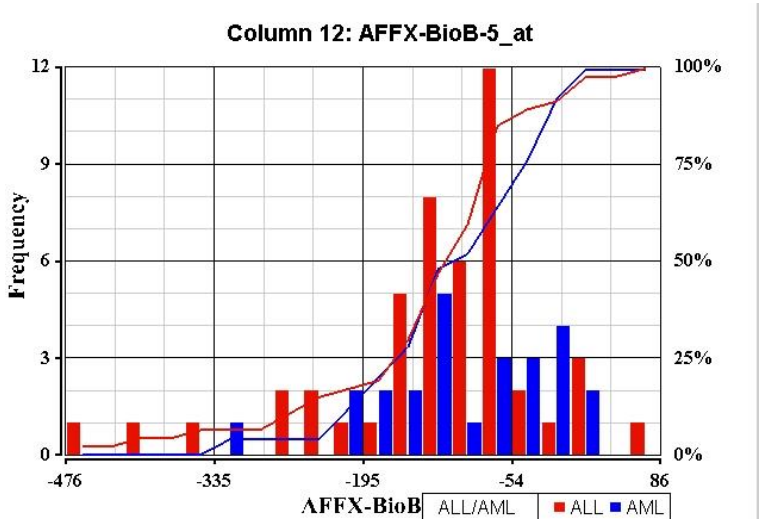


Figure 6. 79: Viewing histogram bars with Accumulation lines

Box & Whiskers

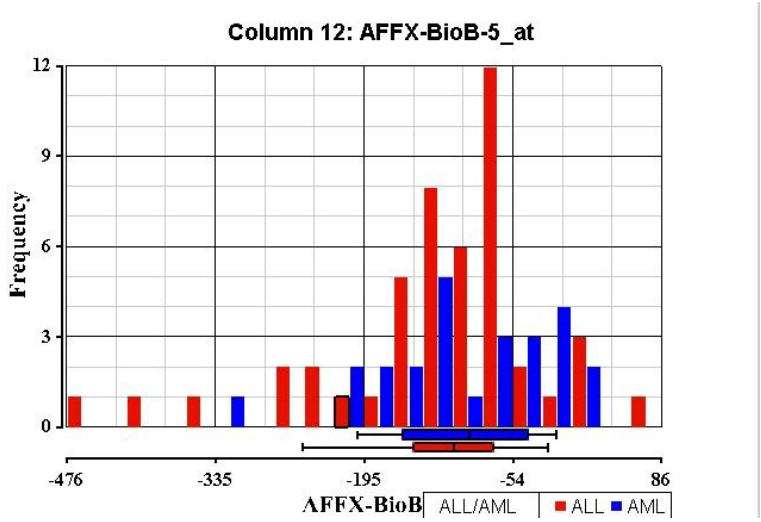


Figure 6. 80: Viewing a histogram with Box and whiskers

The Profiles Plots

Profiles show values arranged by row or column. The label for each row/column is on the X-axis, and the value of the corresponding row/column is on the Y-axis. The *Group Profile* summarizes values based on a categorical column. The *Box & Whiskers Profile* summarizes values across the entire row/column.

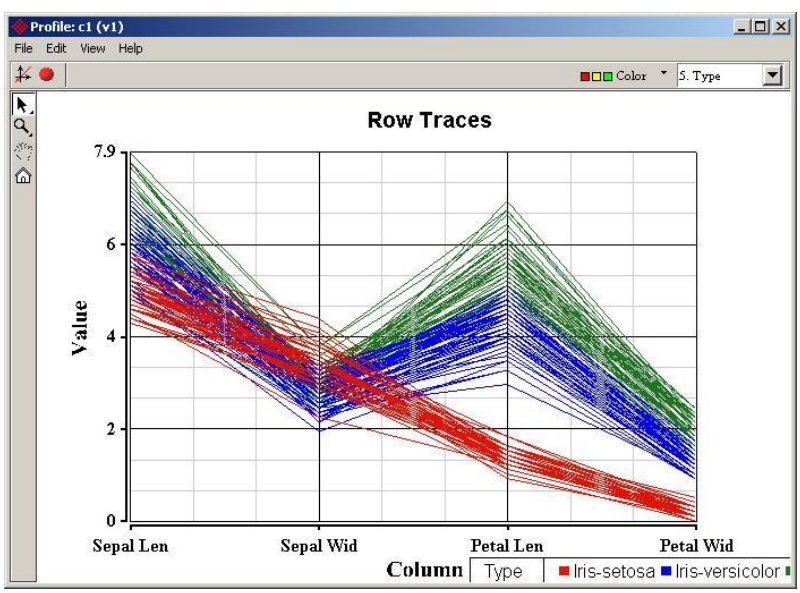


Figure 6. 81: Viewing a Profile on rows

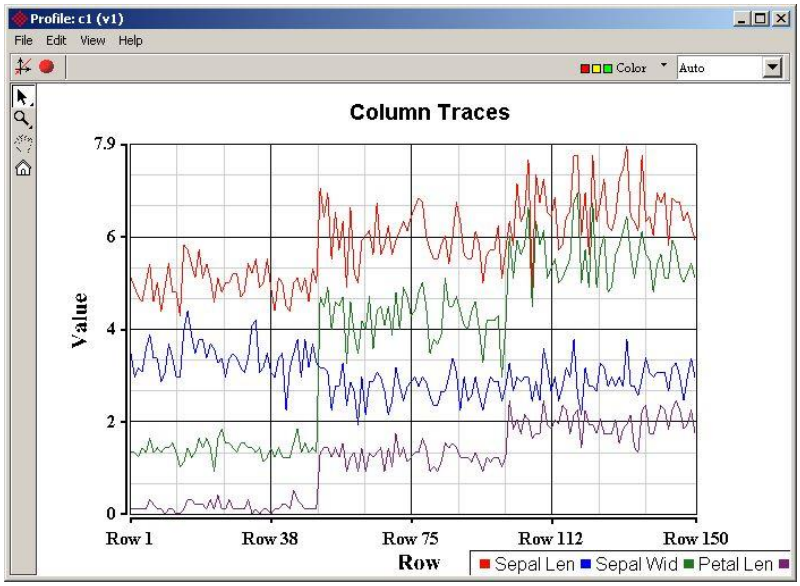


Figure 6. 82: Viewing a Profile on columns

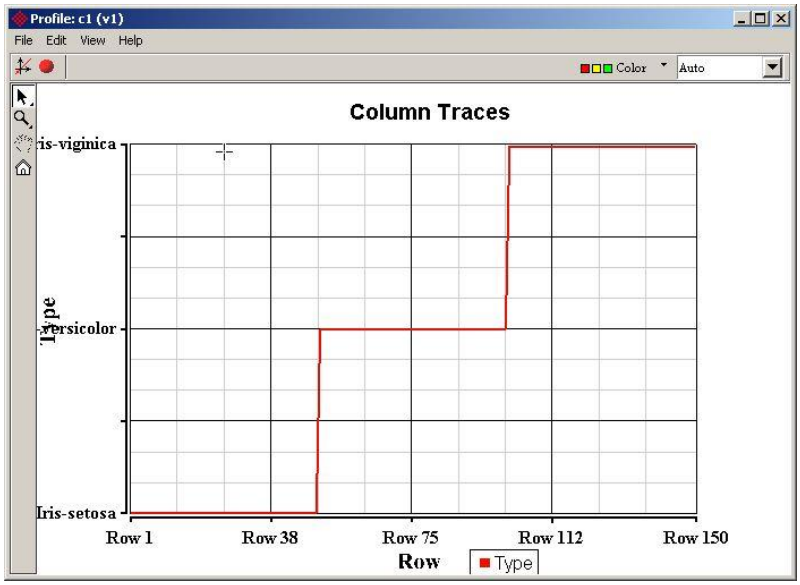


Figure 6. 83: Viewing a Profile on a nominal column

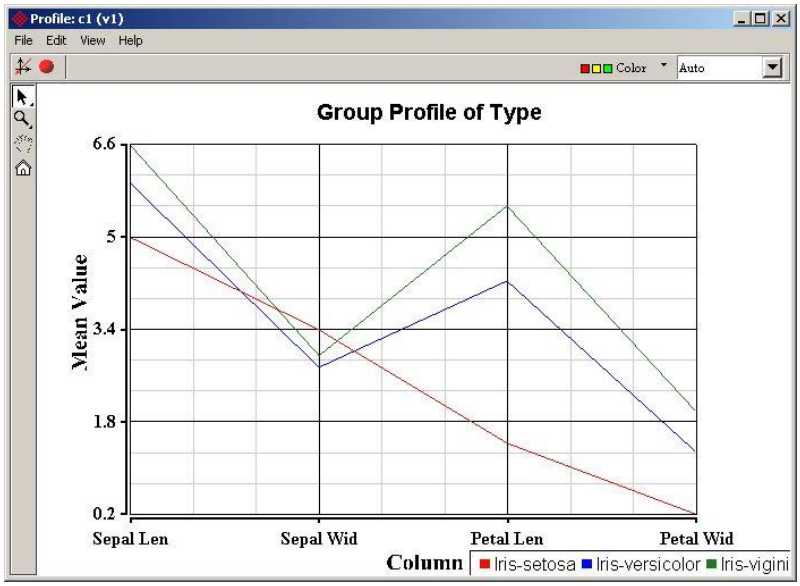


Figure 6. 84: Viewing a Group Profile

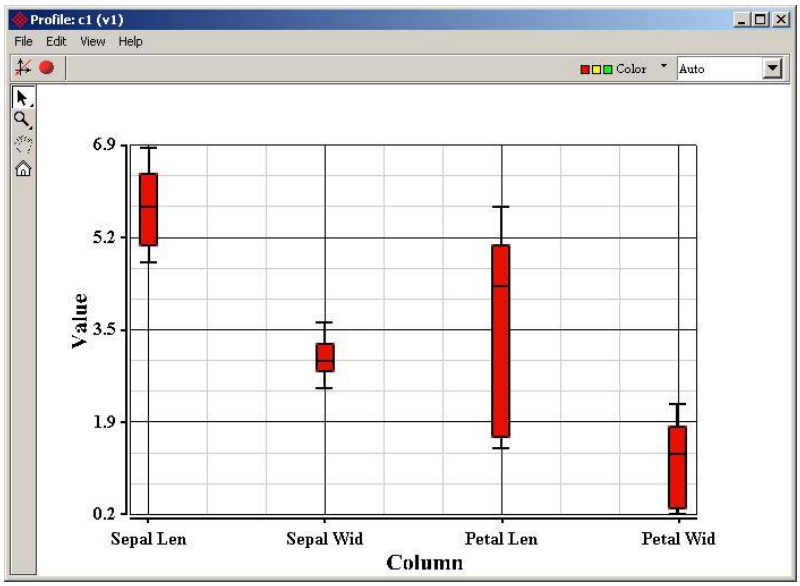


Figure 6. 85: Viewing a Box & Whiskers profile on columns

Opening a Profile Plot

To open a profile plot, click **View > Profiles > Row / Column Profiles...** (Figure 6. 86) from the Partek main window, or click the *Profiles* accelerator button on the tool bar (Figure 6. 87). Additionally, the profile plot is available via the pop-up menu on rows or columns.

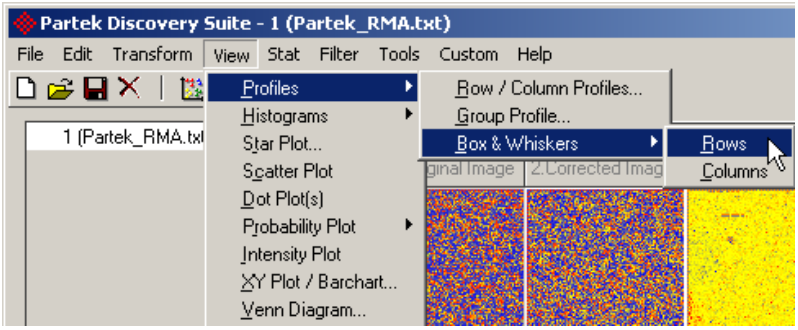


Figure 6. 86: Showing the Profiles menu option

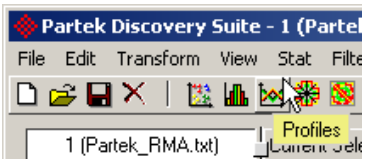


Figure 6. 87: Showing the Profiles accelerator button

Creating a Profile Plot

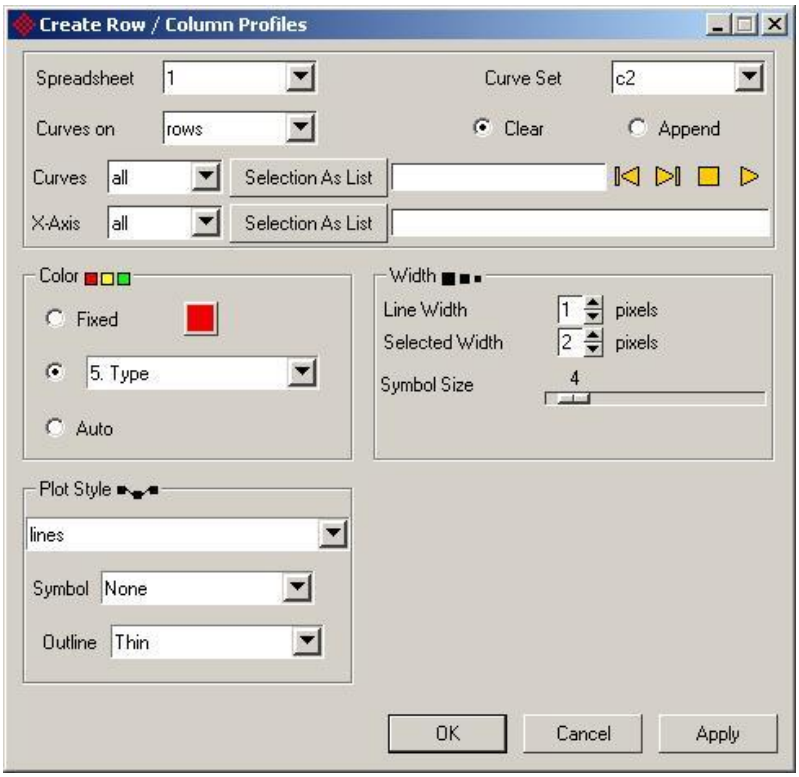


Figure 6. 88: Configuring the Create Profile dialog

The *Create Row /Column Profiles* dialog allows you to create and configure profile plots (Figure 6. 88).

What to Plot

Figure 6. 89: Configuring what to plot

To add lines to an existing Curve Set, specify the existing *Curve Set* and *Append*.

With the *Curve*s row/column entry, there are four buttons (**Prev**, **Next**, **Stop**, and **Fly Through**). Pressing **Prev** or **Next** will put the appropriate number in the entry and plot it. Click **Fly Through** to cycle through each row/column in the spreadsheet, starting at the number in the entry (the first row/ numeric column, if the entry is blank). Pressing the **Stop** button will end a fly through.

Nominal columns cannot be plotted with numeric columns. If both types are specified, then only numeric columns will be shown. Only one nominal column can be shown at a time.

Creating Group Profiles

To access a *Group Profile* click **View > Profiles > Group Profile...** from the Partek main window.

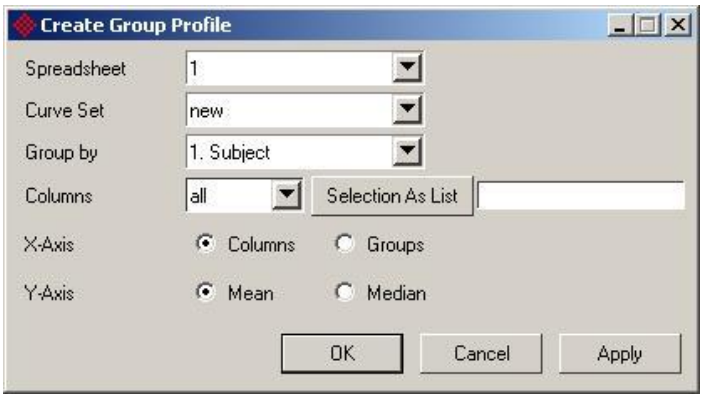


Figure 6. 90: Configuring the Group Profile dialog

The group profile is a plot of the means of the rows grouped by a given categorical column.

If the *X-Axis* is set to **Columns** then *Group by* can be set to **None**. This will plot one line giving the mean/median of all values in each column.

Profile Plot Specific Menus

The File, Edit, View, and Help menus in the Profile Plot viewer behave the same as the menus in the Scatter Plot viewer. Any differences will be notated below, otherwise see the **Viewing the Scatter Plot Results** section above.

The **Edit > Plot Properties > Style, Labels, Box & Whiskers, Titles, Axes, Color,** and **Labels** in the Profile Plot behave the same as in the Scatter Plot - Plot Properties. Any differences will be notated below, otherwise see the **Scatter Plot Properties** section above.

The Mode buttons within the Profile Plot Viewer behave the same as in the Mode buttons in the Scatter Plot. Any differences will be notated below, otherwise see the **Miscellaneous Viewer Options** section above.

Configuring the Profile Plot

Configure Axis

Fold scale is designed for columns that hold ratios. Axis labels that would be between 0 and 1 (exclusive) are shown as the negative inverse. The axis label that would be 1 is shown as "N/C" (no change). When the scaling is log, the log base can be set as either a *number* or *e*. If the scaling is non-linear and all values for the given axis are negative, then an error will be generated and the plot will remain linearly scaled, otherwise the values less than or equal to zero will simply not be shown. If scaling is set to *fold*, then the points will remain in the same place, but axis labels that would be negative are replaced with a dash. If the axis is log scaled then the points with negative values will not be shown.

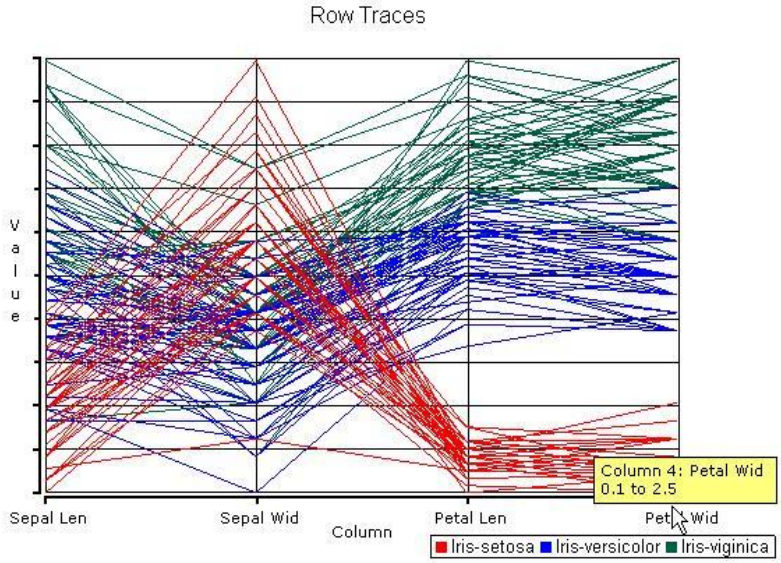


Figure 6. 91: Viewing an Independent scaling profile plot

If Y scaling is *independent (linear)* or *independent (log)*, then the Y position of each point will be based on the minimum and maximum of the row/column associated with the X position. Y min/max will be fixed to *auto*. Holding the mouse over the X-axis reveals the minimum and maximum of that row/column.

The label format for the X-axis is set in the *Annotation* dialog, available from the view menu.

The axis or axis label can be turned on and off to show or hide the axis or axis label, respectively.

When the scaling is *log*, the range of the axis can be specified as either *Exponent* or *Real Number*.

The range of the axis can be specified by *min* and *max*. Set parameters as *manual* first when editing. The axis can be drawn in reverse order by checking the **Values in Reverse Order** button.

Click **OK** or **Apply** to apply the changes only to the specified axis (or *All* axes).

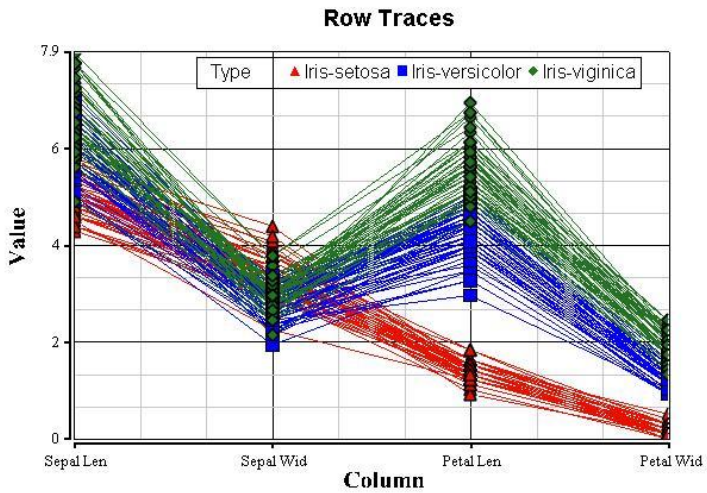


Figure 6. 92: Viewing a Profile plot shaped by column

Profile Plot Properties

Box & Whiskers

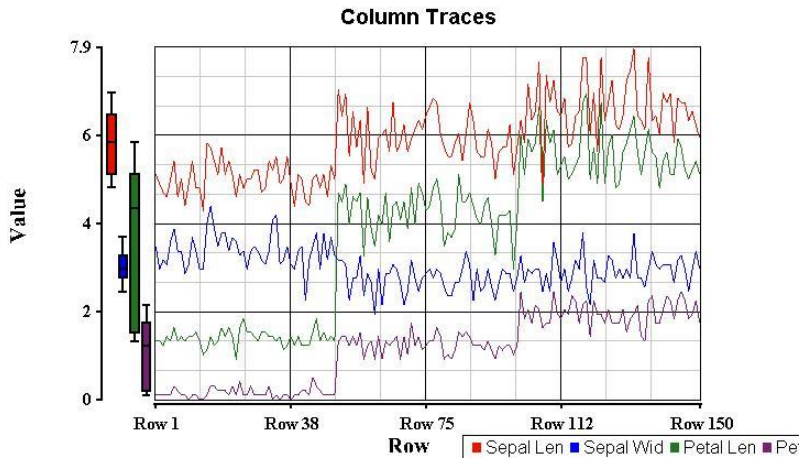


Figure 6. 93: Viewing a profile plot with Box and Whiskers

Error Bars

Error Bars can be drawn on group profiles. By default, the error bars are on standard error, but they can be set to standard deviation.

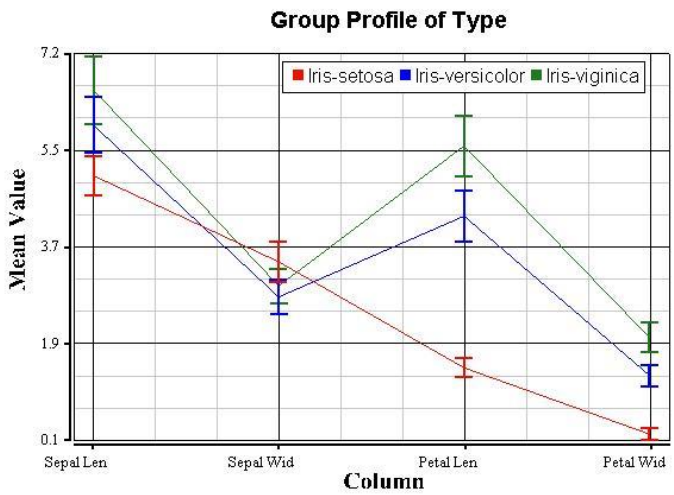


Figure 6. 94: Viewing a Group profile with error bars set to standard deviation

Color

Color scaling is only applied if coloring by a numeric column.

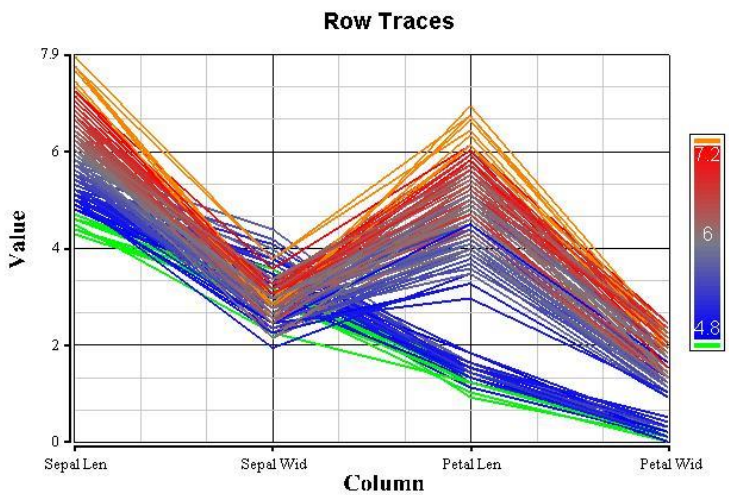


Figure 6. 95: Viewing a profile plot with Color scaling applied

The Intensity Plot

The Intensity Plot is a view of the numerical values in the spreadsheet. The columns of the spreadsheet are on the X-axis; the rows of the spreadsheet are on the Y-axis (Figure 6. 96).

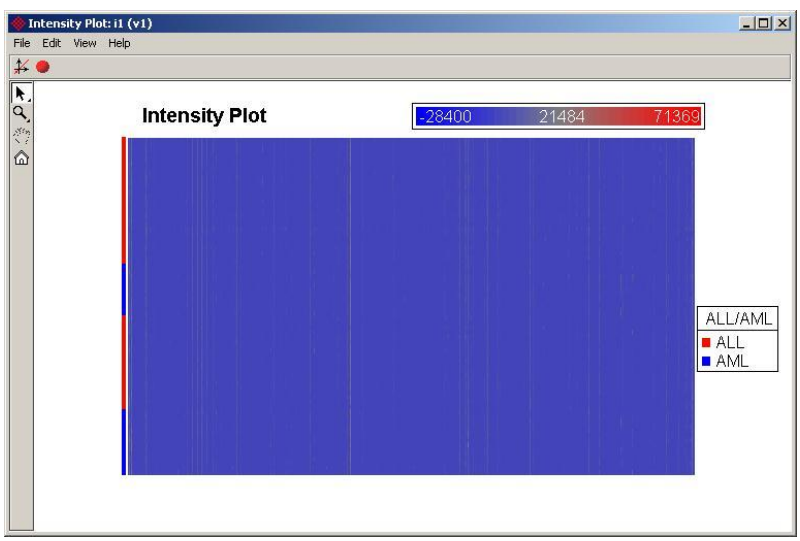


Figure 6. 96: Viewing an Intensity Plot

Opening an Intensity Plot

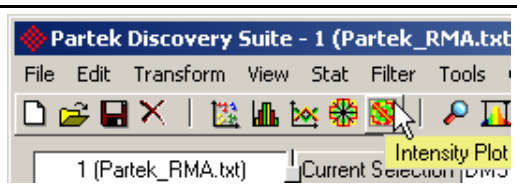


Figure 6. 97: Selecting the Intensity Plot accelerator button

To create an intensity plot, click on the accelerator button (Figure 6. 97) or select **View > Intensity Plot** on the Partek main menu. If any rows or columns are selected, then the intensity plot will be drawn on the selected rows or columns.

Intensity Plot Specific Menus

The File, Edit, View, and Help menus in the Intensity Plot viewer behave the same as the menus in the Scatter Plot viewer. Any differences will be notated below, otherwise see the **Viewing the Scatter Plot Results** section above.

The **Edit > Plot Properties > Style, Labels, Box & Whiskers, Titles, Axes, Color,** and **Labels** in the Intensity Plot behave the same as in the Scatter Plot - Plot Properties. Any differences will be notated below, otherwise see the **Scatter Plot Properties** section above.

The Mode buttons within the Intensity Plot Viewer behave the same as in the Mode buttons in the Scatter Plot. Any differences will be notated below, otherwise see the **Miscellaneous Viewer Options** section above.

Configuring the Intensity Plot

Rows/Columns to Show

The *Configure Plot* dialog for the intensity plot can be configured to show a subset of the data in the spreadsheet (Figure 6. 98).

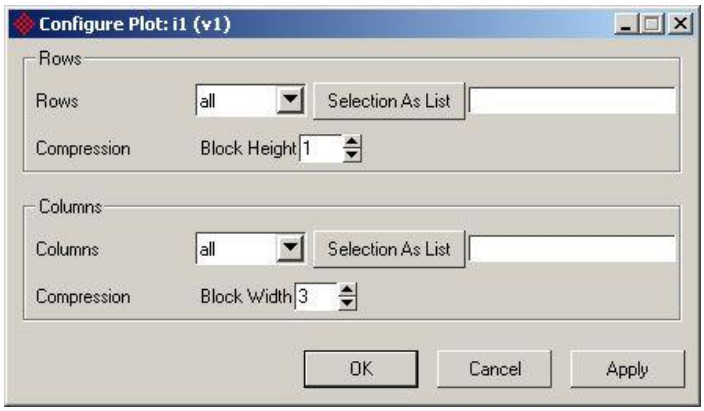


Figure 6. 98: Configuring the Configure Plot dialog

Compression

By default, the intensity plot is compressed so that each cell has a minimum size of 1 pixel. For example, if your spreadsheet has 1000 rows and your screen resolution is 1024x768, then the rows will have a compression of 2 (the top cells will contain the mean of the first 2 rows, the next cells down will contain the mean of the third and fourth rows, etc).

Increasing compression will increase the responsiveness of the viewer at the loss of some accuracy of the picture.

Intensity Plot Properties

Row Annotation

The columns in the spreadsheet can be edited under the *Row Annotation* page (Figure 6. 99).

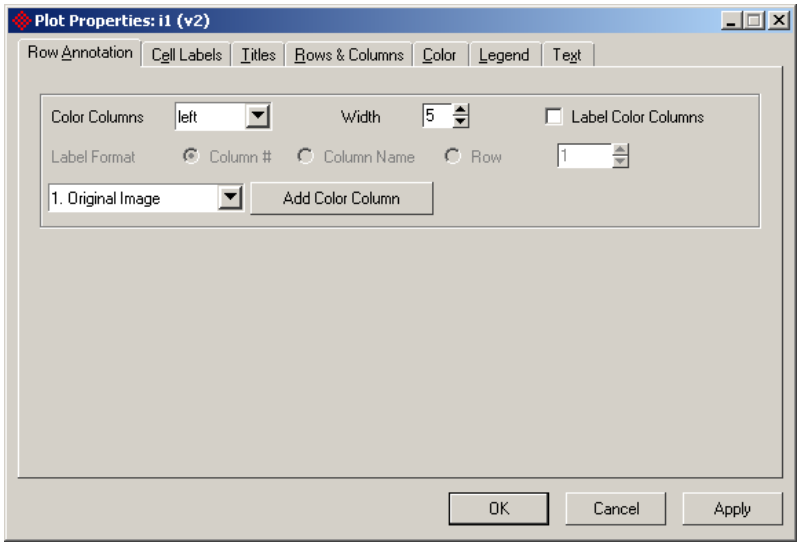


Figure 6. 99: Configuring the Row Annotation page

Adding Color Columns

For each column in the “Columns” list there will be a box indicating the value of the row. The first column in the list will be nearest the intensity plot. The color columns can be shown on either side of the intensity plot. The "Width" determines how many pixels wide each box will be. If the row is selected then the associated box will be one and a half times bigger.

Cell Labels

Label cells with their values on the *Cell Labels* page (Figure 6. 100).

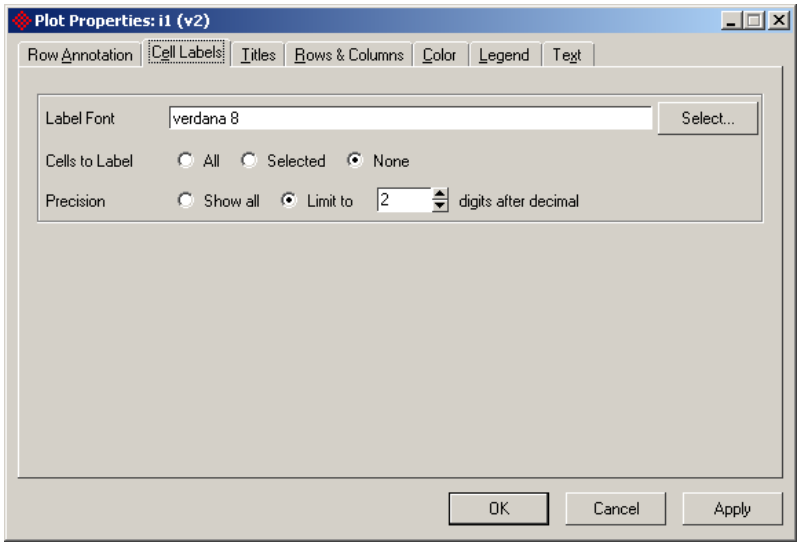


Figure 6. 100: Configuring the Cell Labels page

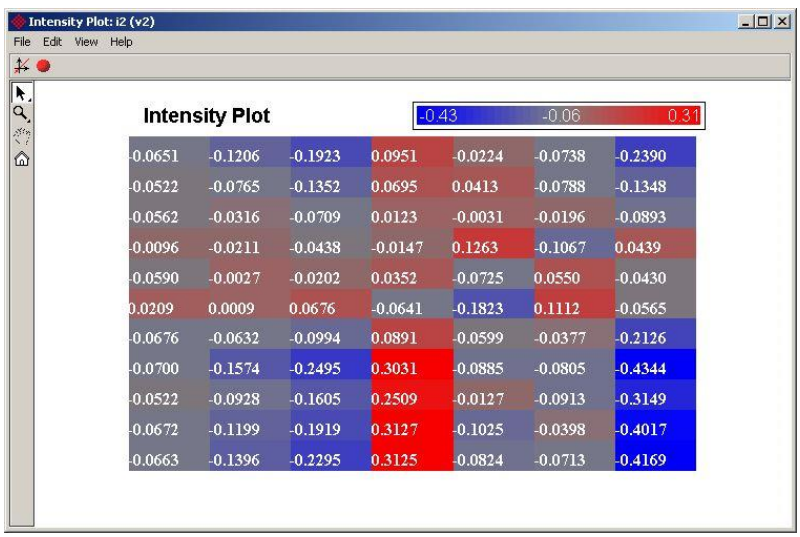


Figure 6. 101: Viewing an Intensity plot with cell labels

Rows & Columns

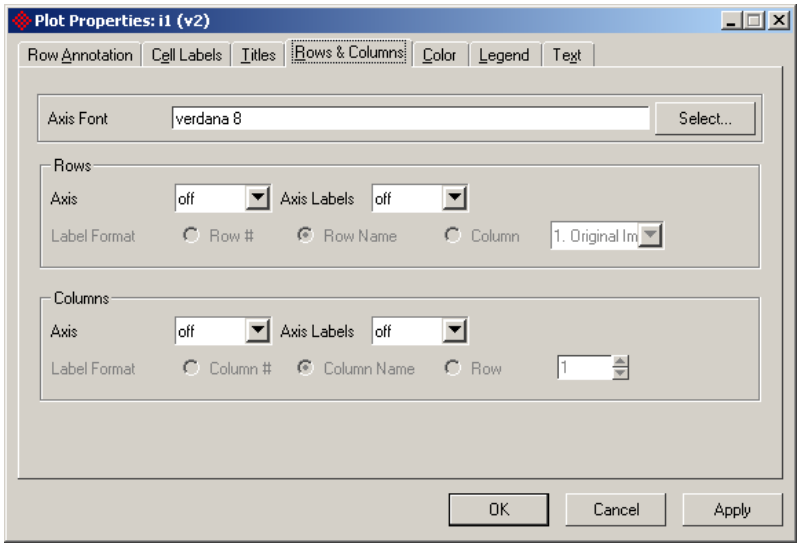


Figure 6. 102: Configuring the Rows & Columns page

Axis

The Axis contains five major ticks, which give the numeric values of the rows/columns.

Axis Labels

When axis labels are on, a label will be shown for each row/column of cells.

Color

The left and right tabs on the histogram represent the *Min* and *Max* values. Specify the range of the color represented by the continuous color map by typing in *Min* and *Max* values and pressing **Enter**. The positions of the tab on the histogram will be updated. You can also drag the tabs to specify the range; the *Min* and *Max* values will be updated while dragging.

The first color on the color map will be used if *First* is used as the *Min Outlier Color*. The last color on the color map will be used if *Last* is used as the *Max Outlier Color*. If outliers are colored as *Fixed*, they will be drawn using the specified color.

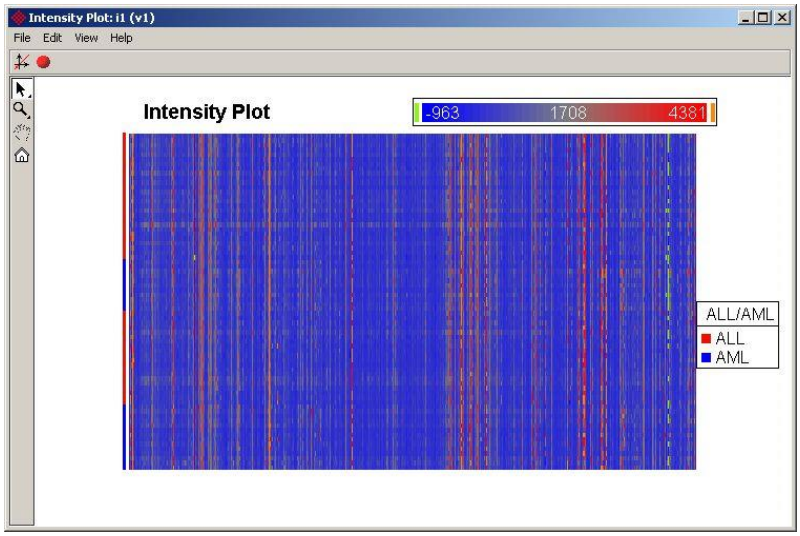


Figure 6. 103: Viewing an Intensity plot with Color Scaling in effect

The XY/Barchart Plot

An XY plot is a two dimensional graph that allows the examination of the effect of one or two categorical variables (factors) on a response variable (e.g. gene expression) (Figure 6. 104).

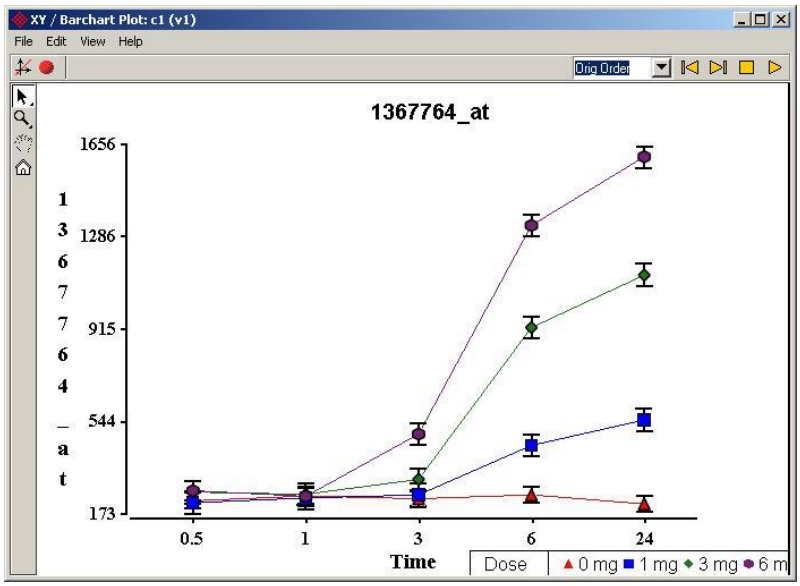


Figure 6. 104: Viewing an XY Plot

The Barchart plot is the same as the XY plot configured with line style set to bars (Figure 6. 105).

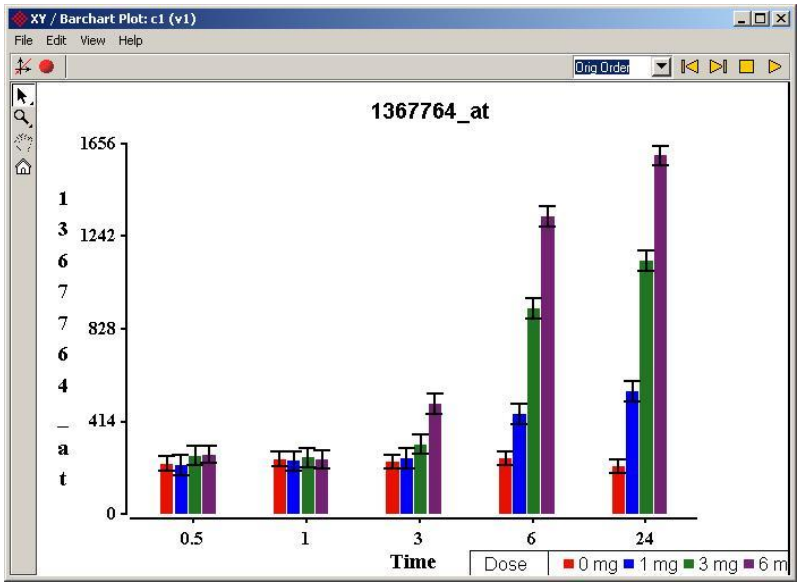


Figure 6. 105: Viewing a Barchart Plot

Opening a XY / Barchart Plot

To invoke a *XY / Barchart plot*, in the Partek main window, click menu **View > XY / Barchart Plot** (Figure 6. 106).

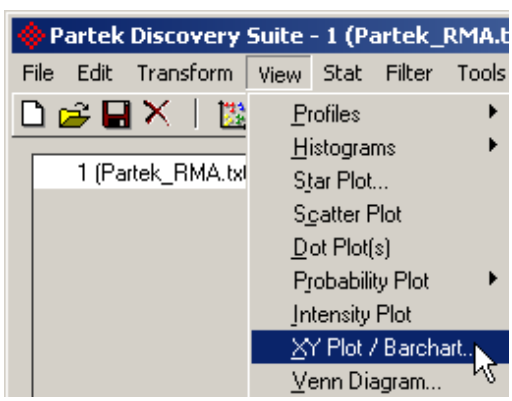


Figure 6. 106: Showing the XY / Barchart Plot menu item

Creating a XY / Barchart Plot

The *Create XY / Barchart Plot* dialog appears after selecting **XY/Barchart Plot** from the *View* menu. Here the XY Plot can be drawn as specified (Figure 6. 106). By default, if there is more than one categorical variable in the spreadsheet the XY plot is drawn based on two factors. If there is only one categorical variable then *Separate By* will be set to **None**. *Mean* and *LS-Mean* can be set here as well as on the **Plot Properties > Axes > Y-Axis** dialog.

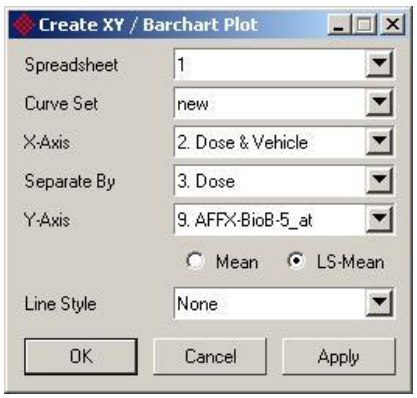


Figure 6. 107: Configuring the Create XY / Barchart Plot dialog

XY/Barchart Plot Specific Menus

The File, Edit, View, and Help menus in the XY/Barchart Plot viewer behave the same as the menus in the Scatter Plot viewer. Any differences will be notated below, otherwise see the **Viewing the Scatter Plot Results** section above.

The **Edit > Plot Properties > Style, Labels, Box & Whiskers, Titles, Axes, Color,** and **Labels** in the XY/Barchart Plot behave the same as in the Scatter Plot - Plot Properties. Any differences will be notated below, otherwise see the **Scatter Plot Properties** section above.

The Mode buttons within the XY/Barchart Plot Viewer behave the same as in the Mode buttons in the Scatter Plot. Any differences will be notated below, otherwise see the **Miscellaneous Viewer Options** section above.

Configuring the XY/Barchart Plot

The *Configure Plot* dialog allows the configuration of what values to plot, as well as the range of the Y-axis (Figure 6. 108).

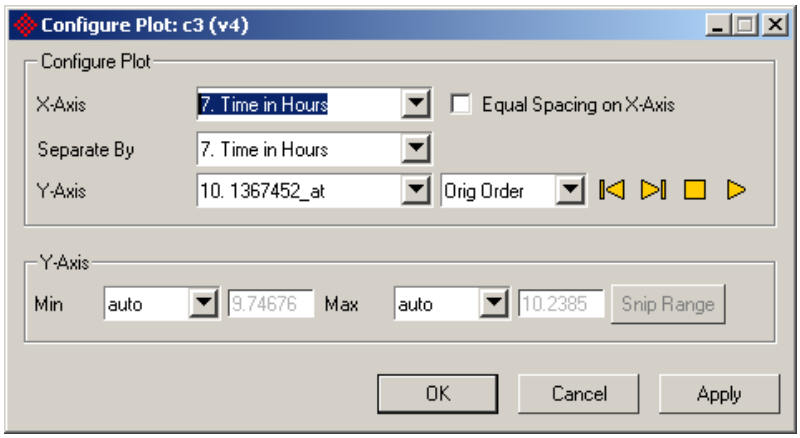


Figure 6. 108: Configuring the Configure Plot dialog

If **Equal Spacing on X-Axis** is unchecked then the spacing will be based on the numerical interpretation of the categories. This option is only available if the categories are numeric (Figure 6. 109).

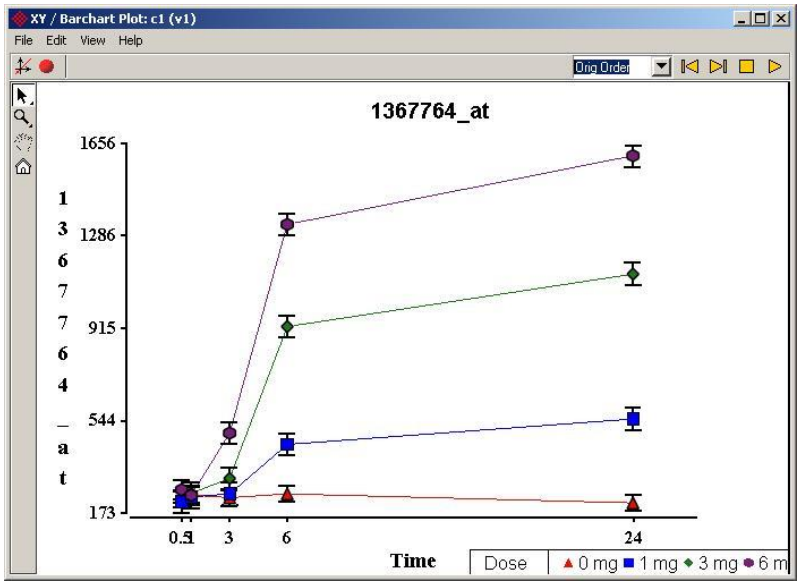


Figure 6. 109: Showing Unequal Spacing on the X-Axis

XY/Barchart Plot Properties

Error Bars

Error Bars can be drawn on the XY Plot. By default, the error bars are on standard error, but if the Y-Axis is set to *Mean* then they can be set to the standard deviation (Figure 6. 110).

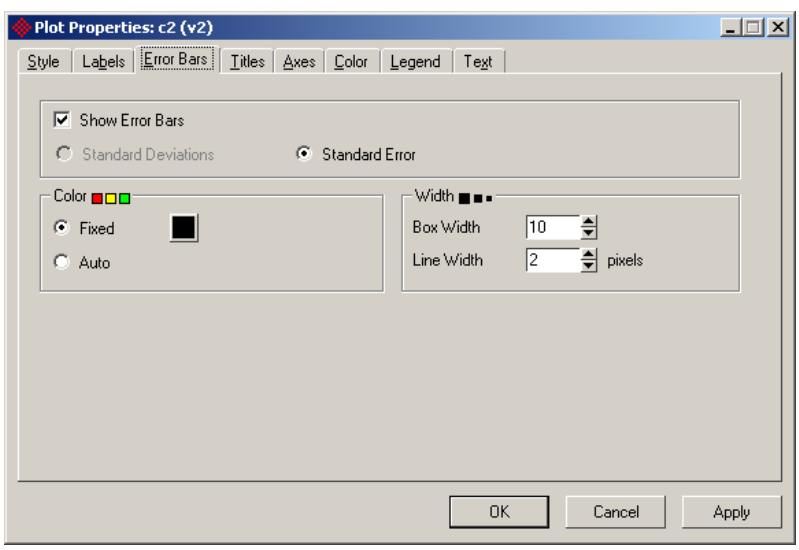


Figure 6. 110: Configuring the Error Bars page

The Sources of Variation Plot

The Sources of Variation Plot is a bar chart that shows the variation contributed by effects across all test variables (response variables) in the ANOVA model. The plot can only be invoked on the result spreadsheet of ANOVA or ANCOVA.

The X-axis of the plot represents the factors or interactions in the ANOVA model; the Y-axis represents the F ratio of the factors or interactions. "F ratio" is ANOVA's language for "signal to noise ratio". The "Average F Ratio" is the average signal to noise ratio of all the computed variables for each factor. It is quantitative. You may wish to examine the median, since the average can be influenced by very large effects on just a few variables.

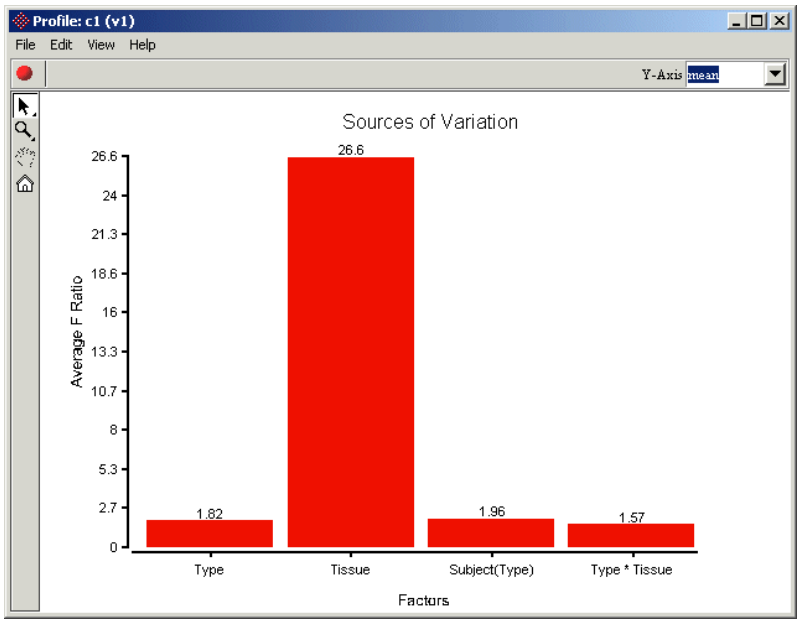


Figure 6. 111: Viewing a Barchart Plot that shows a Sources of Variation in ANOVA

Opening a Sources of Variation Plot

While the ANOVA result spreadsheet is active go to **View > Sources of Variation** in the Partek main window (Figure 6. 112). You can also make a SOV plot for a single variable by right-clicking on the row corresponding to that variable.

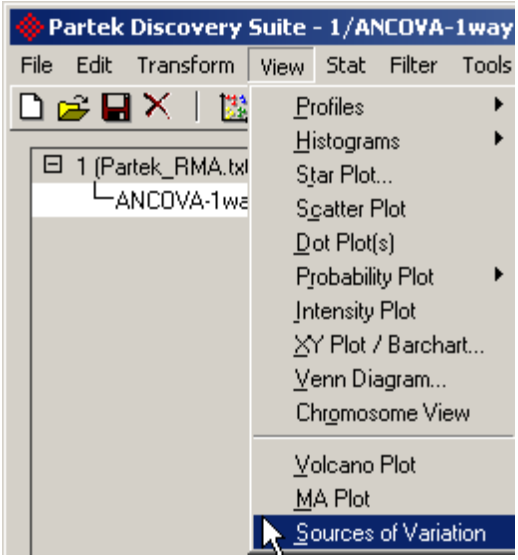


Figure 6. 112: Selecting the Source of Variation from the View menu

Sources of Variation Plot Menu Items

The File, Edit, View, and Help menus in the Sources of Variation Plot viewer behave the same as the menus in the Scatter Plot viewer. Any differences will be notated below, otherwise see the **Viewing the Scatter Plot Results** section above.

The **Edit > Plot Properties > Style, Labels, Box & Whiskers, Titles, Axes, Color,** and **Labels** in the Sources of Variation Plot behave the same as in the Scatter Plot - Plot Properties. Any differences will be notated below, otherwise see the **Scatter Plot Properties** section above.

The Mode buttons within the Sources of Variation Plot viewer behave the same as in the Mode buttons in the Scatter Plot. Any differences will be notated below, otherwise see the **Miscellaneous Viewer Options** section above.

Sources of Variation Plot Properties

Style

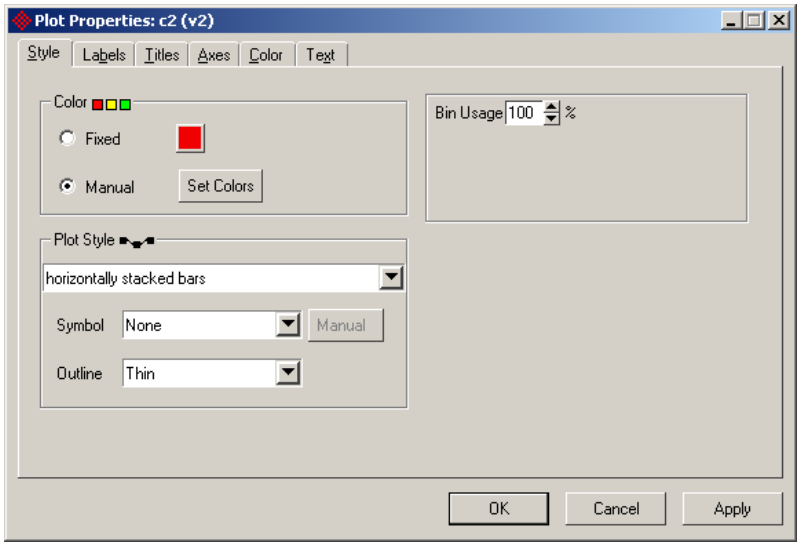


Figure 6. 113: Configuring the Style page

Bin Usage

The width of the bars is specified in the *Bin Usage* panel (Figure 6. 114).

Figure 6. 114: Configuring the Bin Usage panel

The Star Plot

The Star Plot is a way to examine the distribution of the variables. Each line (or set of points) represents a row (Figure 6. 115) or column (Figure 6. 116). The other dimension (columns or rows) is plotted around the theta axis, while the corresponding values are represented by the distance from the center.

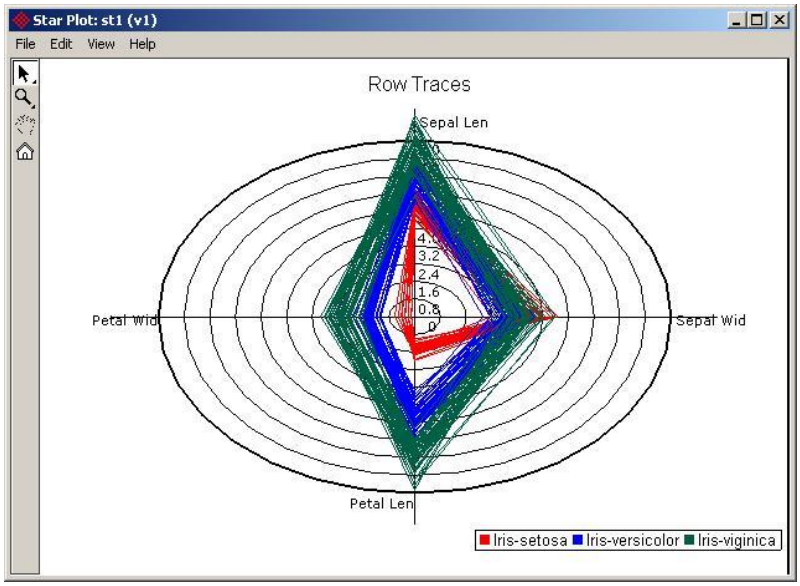


Figure 6. 115: Viewing a Star Plot on rows

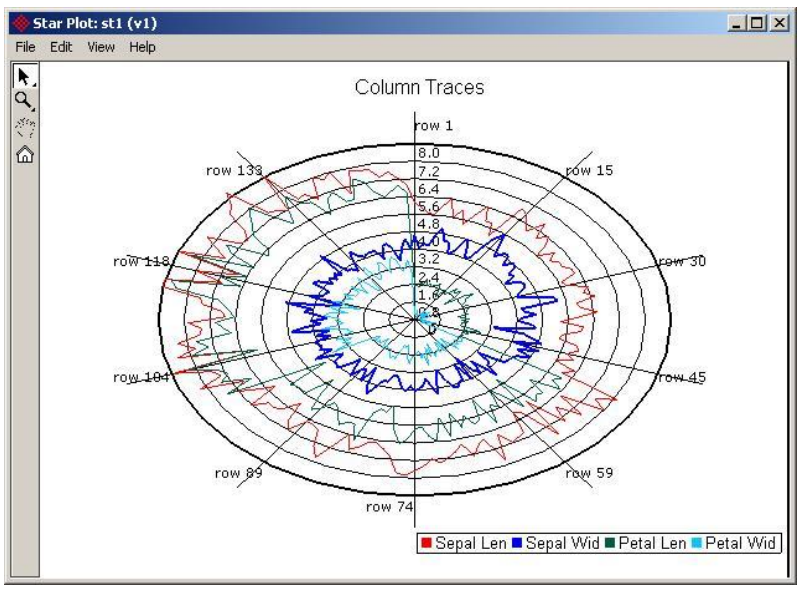


Figure 6. 116: Viewing a Star Plot on columns

Opening a Star Plot

To create a star plot, click **View > Star Plot…** in the Partek main window (Figure 6. 117), or click the *Star Plot* accelerator button on the tool bar (Figure 6. 118). Additionally, the star plot is available via the pop-up menu on rows or columns.

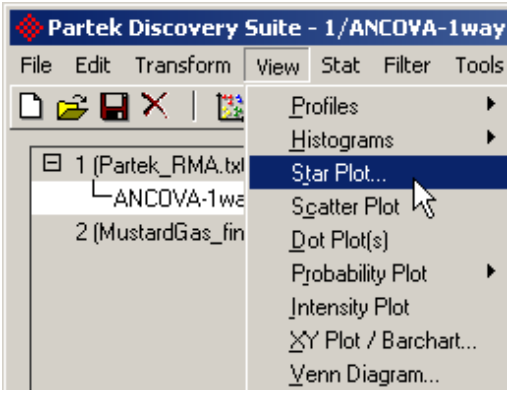


Figure 6. 117: Selecting the Star Plot menu option from the View menu

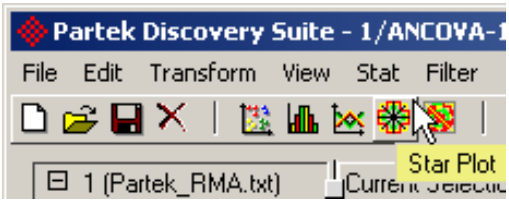


Figure 6. 118: Selecting the Star Plot accelerator button

Creating a Star Plot

Create and configure star plots in the *Create Star Plot* dialog (Figure 6. 119).

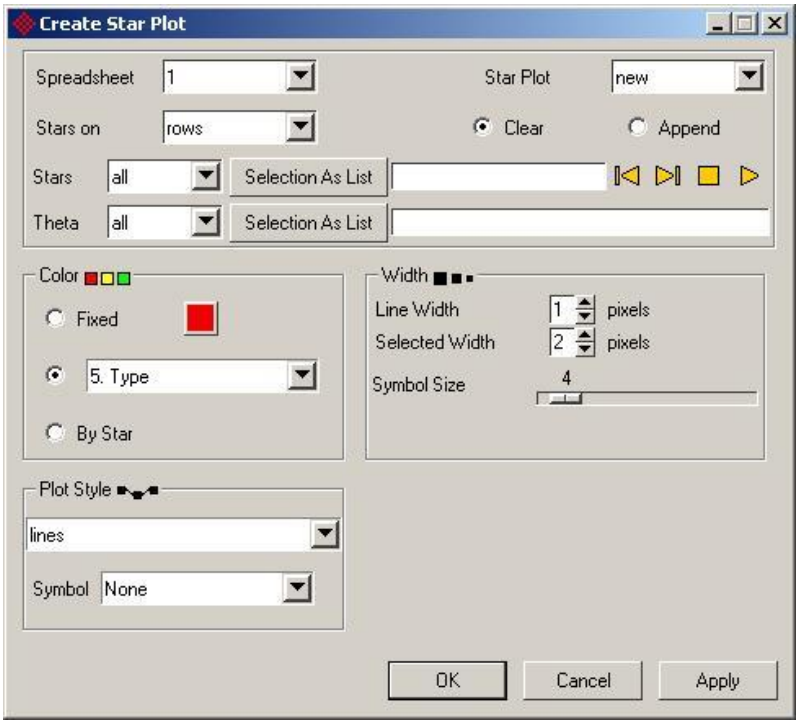


Figure 6. 119: Configuring the Create Star Plot dialog

What to Plot

Specify the existing *Star Plot* and *Append* to add stars to the existing star plot (Figure 6. 120).

Figure 6. 120: Configuring what to plot in the Star plot

With the *Stars* row/column entry there are four buttons (**Prev**, **Next**, **Stop**, and **Fly Through**). Pressing **Prev** or **Next** will put the appropriate number in the entry and plot it. **Fly Through** cycles through each row/column in the spreadsheet, starting at the number in the entry (the first row/ numeric column, if the entry is blank). Pressing the **Stop** button will end a fly through.

Nominal columns cannot be shown in a star plot.

Star Plot Specific Menu Items

The File, Edit, View, and Help menus in the Star Plot viewer behave the same as the menus in the Scatter Plot viewer. Any differences will be notated below, otherwise see the **Viewing the Scatter Plot Results** section above.

The **Edit > Plot Properties > Style, Labels, Box & Whiskers, Titles, Axes, Color,** and **Labels** in the Star Plot behave the same as in the Scatter Plot - Plot Properties. Any differences will be notated below, otherwise see the **Scatter Plot Properties** section above.

The Mode buttons within the Star Plot Viewer behave the same as in the Mode buttons in the Scatter Plot. Any differences will be notated below, otherwise see **Miscellaneous Viewer Options** section above.

Configuring the Star Plot

Configure Plot

The content of the plot is configured in this panel (Figure 6. 121).

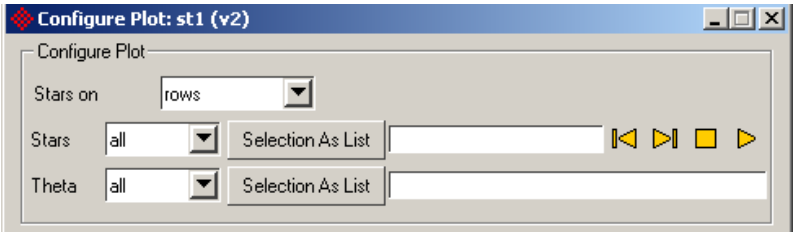


Figure 6. 121: Configuring the Star plot Configure Plot panel

Configure Axis

The scaling and range can be changed only on the radial axis (Figure 6. 122).

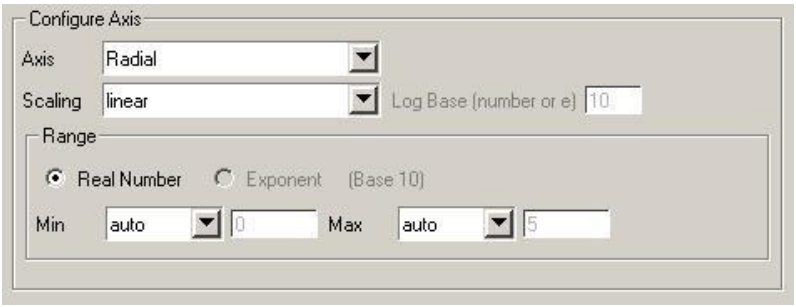


Figure 6. 122: Configuring the Star plot Axis panel

Fold scale is designed for columns that hold ratios. Axis labels that would be between 0 and 1 (exclusive) are shown as the negative inverse. The axis label that would be 1 is shown as "N/C" (no change). When the scaling is log, the log base can be set as either a *number* or *e*. If the scaling is non-linear and all values for the given axis are negative, then an error will be generated and the plot will remain linearly scaled, otherwise the values less than or equal to zero will simply not be shown. If scaling is set to *fold*, then the points will remain in the same place, but axis labels that would be negative are replaced with a dash. If the axis is log scaled then the points with negative values will not be shown.

If radial scaling is *independent (linear)* or *independent (log)* then the radial position of each point will be based on the minimum and maximum of the row/column associated with the theta position (Figure 6. 123). Radial min/max will be fixed at *auto*.

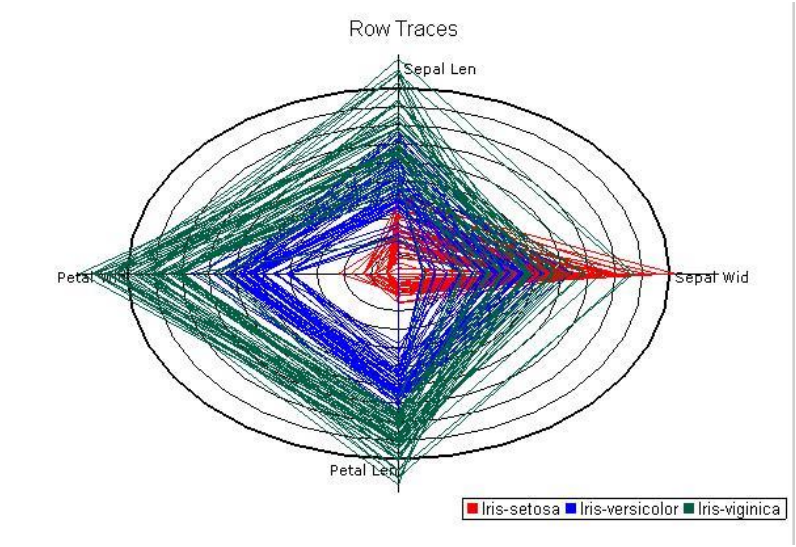


Figure 6. 123: Viewing Independent scaling in the Star plot

When the scaling is *log*, the range of the axis can be specified as either *Exponent* or *Real Number*.

The range of the radial axis can be specified by *min* and *max*. Set the parameters as *manual* first when editing.

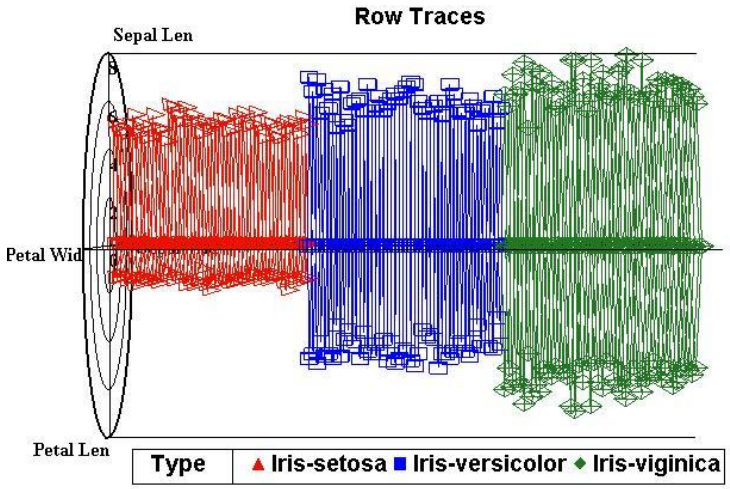


Figure 6. 124: Viewing a Star plot shaped by column

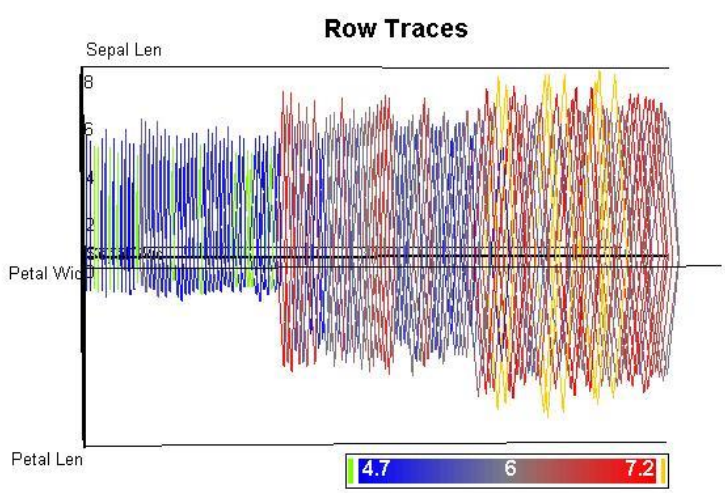


Figure 6. 125: Viewing a Star plot with Color Scaling in effect

The Probability Plot

The probability plot is a graph that indicates whether a data set is normally distributed. The data is plotted against a linear line of theoretical normality for the data set. Departures from the theoretical normality line are not considered normally distributed. Partek software offers two types of probability plots, normal and uniform (Figure 6. 126, Figure 6. 127).

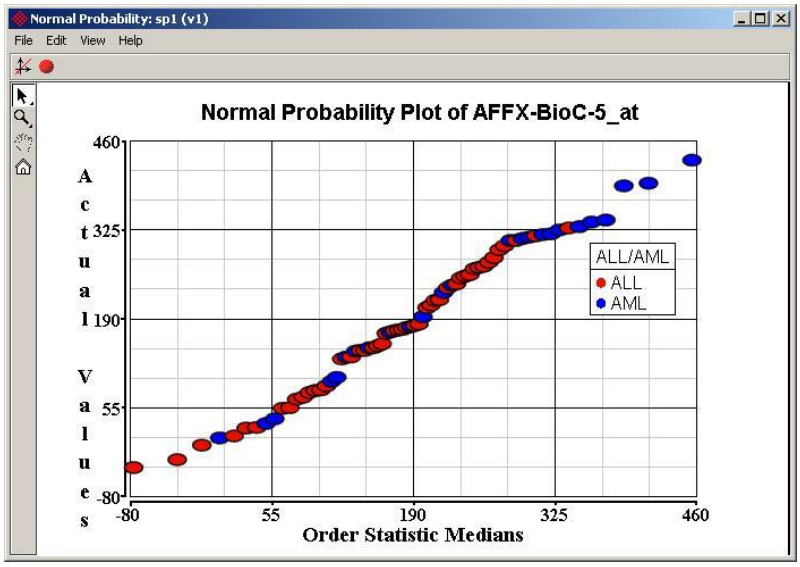


Figure 6. 126: Viewing a Normal Probability Plot

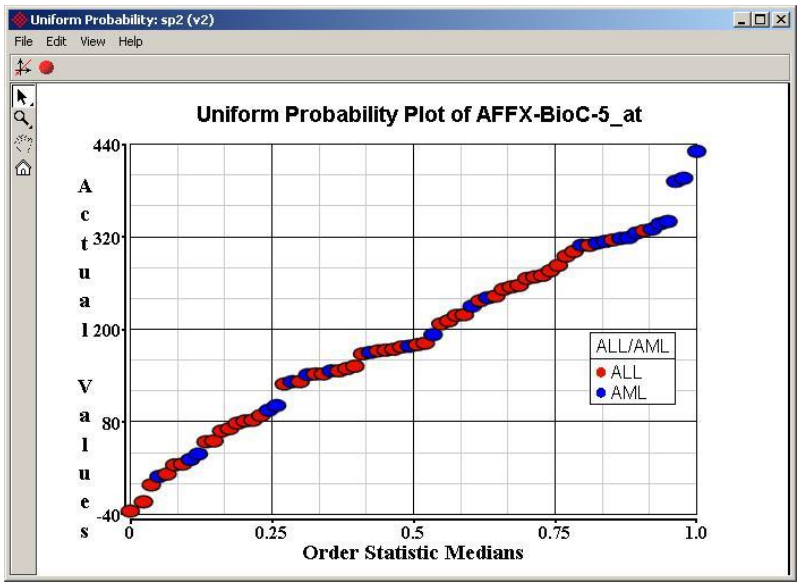


Figure 6. 127: Viewing a Uniform Probability Plot

Opening a Normal Probability Plot

To invoke a normal probability plot, in the Partek main window, click menu **View > Probability Plots > Normal Probability Plot** (Figure 6. 128). The uniform probability plot can be invoked by choosing **View > Probability Plots > Uniform Probability Plot**.

Note: At least one numerical column within the Partek spreadsheet must be selected prior to invoking the probability plot. A separate plot will open for each numeric column selected.

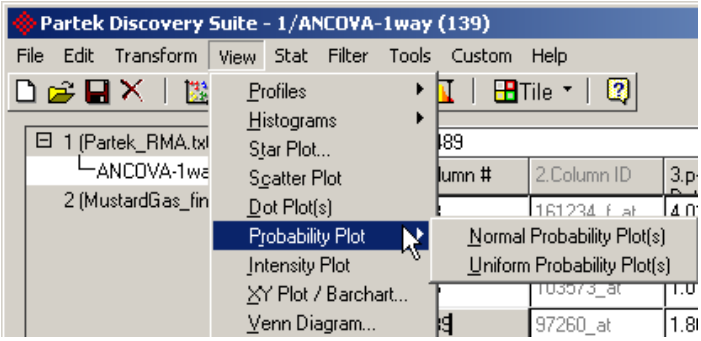


Figure 6. 128: Selecting the Normal Probability Plot menu option

Probability Plot Menu Items

The File, Edit, View, and Help menus in the Probability Plot viewer behave the same as the menus in the Scatter Plot viewer. Any differences will be notated below, otherwise see the **Viewing the Scatter Plot Results** section above.

The **Edit > Plot Properties > Style, Labels, Box & Whiskers, Titles, Axes, Color,** and **Labels** in the Probability Plot behave the same as in the Scatter Plot - Plot Properties. Any differences will be notated below, otherwise see the **Scatter Plot Properties** section above.

The Mode buttons within the Probability Plot viewer behave the same as in the Mode buttons in the Scatter Plot. Any differences will be notated below, otherwise see the **Miscellaneous Viewer Options** section above.

The Venn Diagram

The Venn diagram provides a way to visualize the relationship between two or three lists from any subsection of the diagram.

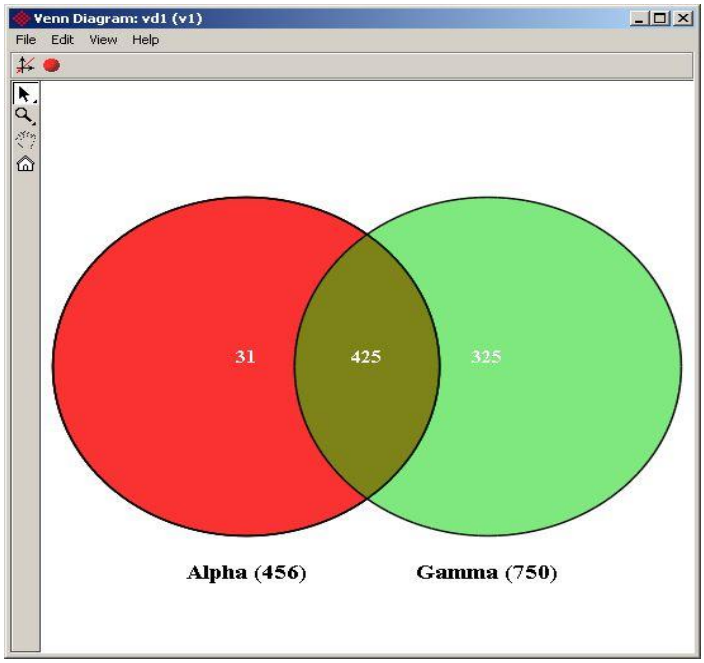


Figure 6. 129: Viewing a Venn diagram with two lists

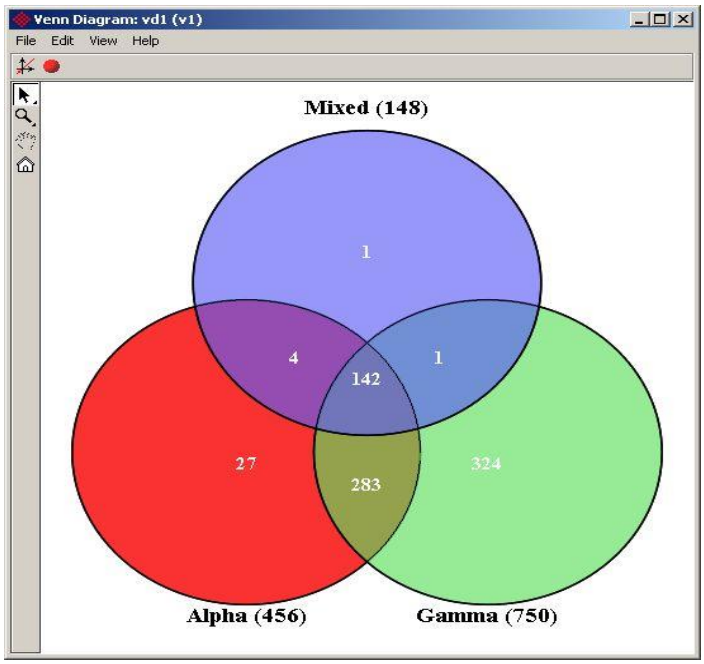


Figure 6. 130: Viewing Venn diagram with three lists

Opening the Venn Diagram

To create a Venn diagram, click **View > Venn Diagram** in the Partek main window**.** The Venn diagram is available from the *List Manager* (**Tools > List Manager**)

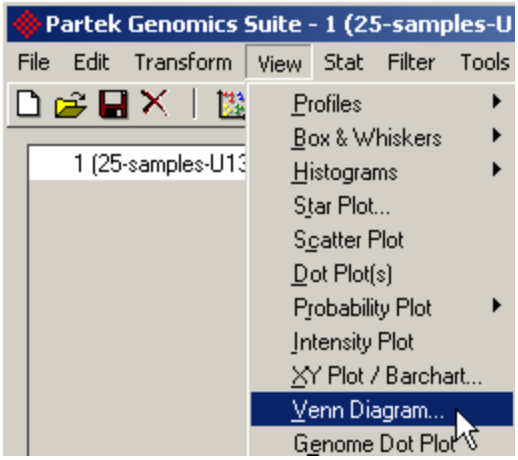


Figure 6. 131: Selecting the Venn diagram menu option from the View menu

Creating a Venn Diagram

From the list manager select two to three lists then click the **Venn Diagram** button (Figure 6. 132).

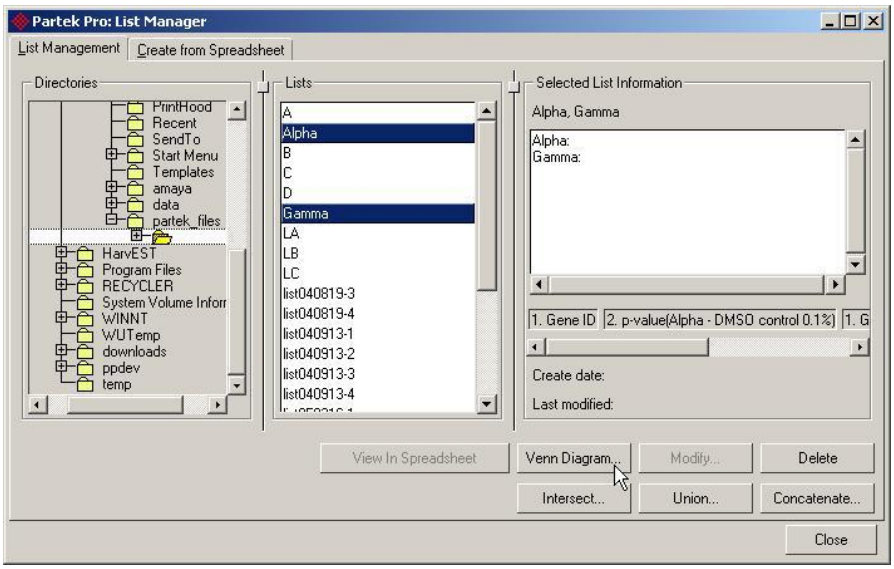


Figure 6. 132: Creating a Venn diagram from the list manager

Venn Diagram Menu Items

The File, Edit, View, and Help menus in the Venn diagram viewer behave the same as the menus in the Scatter Plot viewer. Any differences will be notated below, otherwise see the **Viewing the Scatter Plot Results** section above.

The **Edit > Plot Properties > Style, Labels, Box & Whiskers, Titles, Axes, Color,** and **Labels** in the Venn diagram behave the same as in the Scatter Plot - Plot Properties. Any differences will be notated below, otherwise see the **Scatter Plot Properties** section above.

The Mode buttons within the Venn diagram viewer behave the same as in the Mode buttons in the Scatter Plot. Any differences will be notated below, otherwise see the **Miscellaneous Viewer Options** section above.

Creating Lists from Selected Slices

Create List from Selected Slices will create a list from the currently selected slice. Using the pop-up menu, it is possible to create lists using *Union* or *Intersection* if a region with overlapping slices is selected (Figure 6. 133).

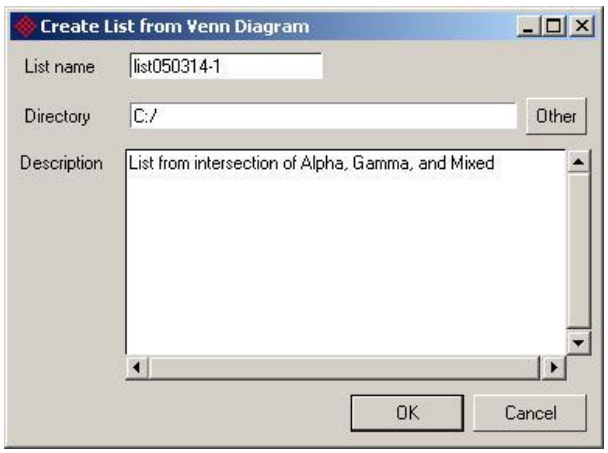


Figure 6. 133: Creating a list from a Venn diagram

Configuring the Venn Diagram

Configure Plot

The *Configure Plot* dialog allows the configuration of what values to plot (Figure 6. 134).

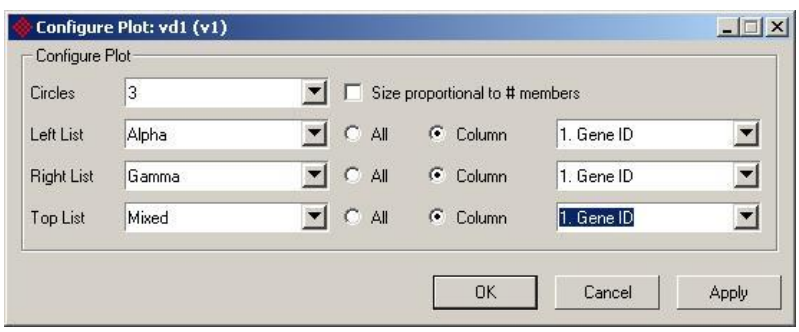


Figure 6. 134: Configuring the Venn diagram Configure Plot dialog

Partek® Genomics Suite Specific Visualizations

A Visual Overview

Plots on Result Spreadsheets –Volcano Plot and MA Plot

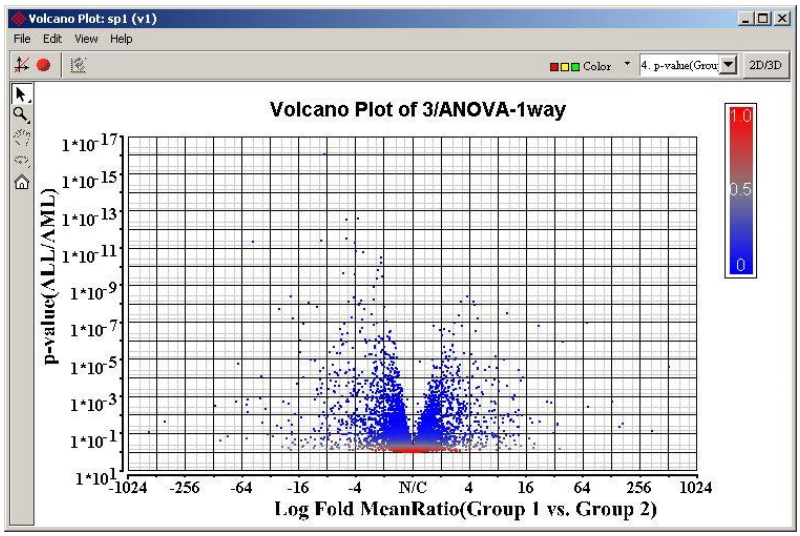


Figure 6. 135: View > Volcano Plot

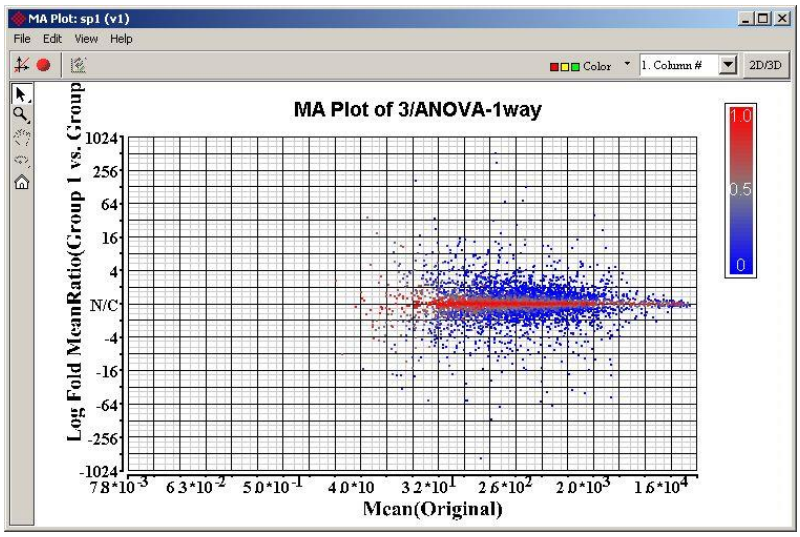


Figure 6. 136: View > MA Plot

Chromosome View 6.4

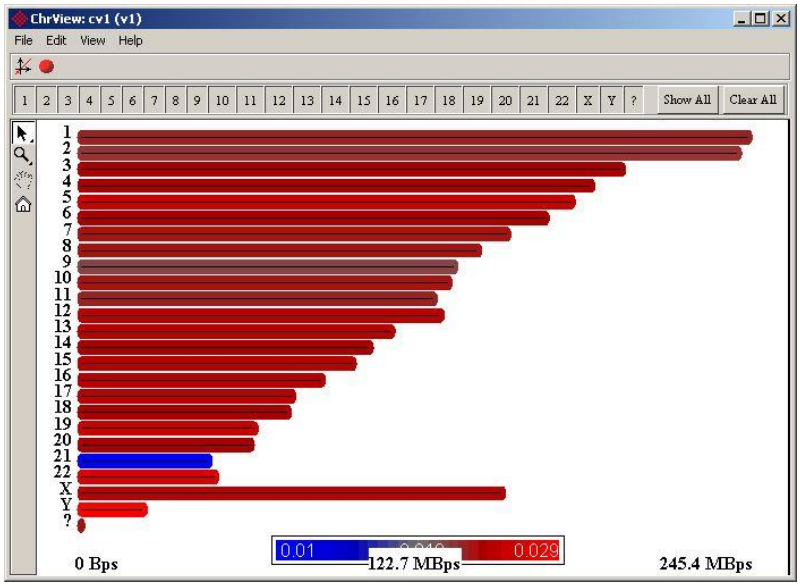


Figure 6. 137: View > Chromosome View 6.4

Chromosome View

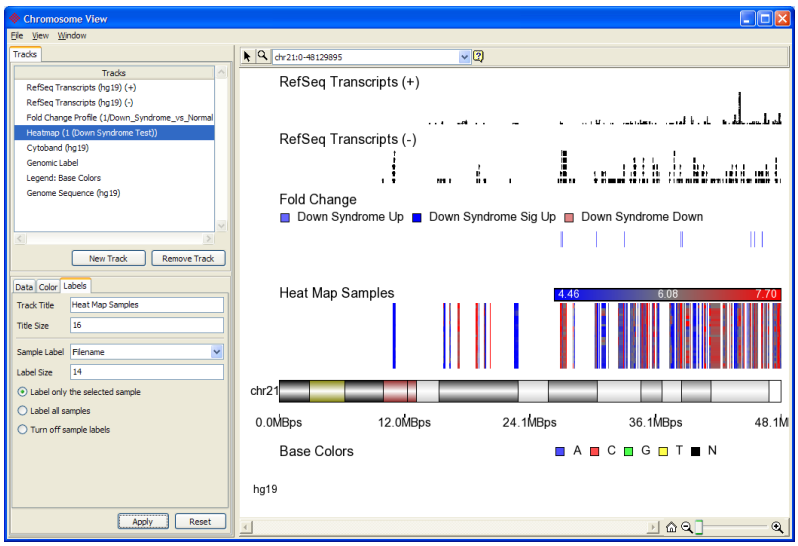


Figure 6. 138: View > Chromosome View

Save Images (Orig. Data)

Save Images is only available on statistical spreadsheets such as an ANOVA result.

It will invoke a view and save a .jpg image for each row in the result spreadsheet.

Specify the directory to contain the images. The filename of each image will be based on the probeset id (Figure 6.).

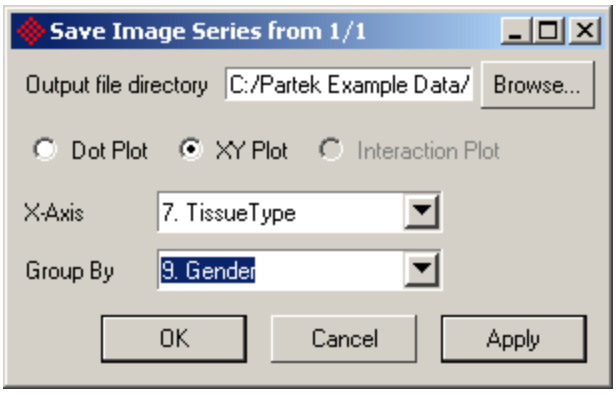


Figure 6. 139: View > Save Images (Orig. Data)

The Volcano Plot

The Volcano Plot is a special 2-D scatter plot to visualize the statistical results on the significance and amount of changes in all the tested variables between two different levels (groups) of a factor. The X-axis represents the fold change of the test variable under the two conditions; it is on log2 scale. The Y-axis shows the p-values from the t-Tests or ANOVA pair-wise comparison tests; it is on log10 scale.

Invoking a Volcano Plot

The volcano plot can only be invoked on the results spreadsheet of the ANOVA or the two-sample t-Test. By default, the result of the two-sample t-Test includes the p-value, mean of each group, and the mean ratio of the two groups. In ANOVA, however, the default results spreadsheet does not contain the mean and mean ratio of the compared groups, but they can be specified in the *Pairwise Comparisons* dialog invoked by clicking the **Result** button in the ANOVA dialog. When the results spreadsheet is active, click **View > Volcano Plot** from the Partek main window (Figure 6. 140), this will invoke the *Volcano Plot Configuration* dialog (Figure 6. 141).

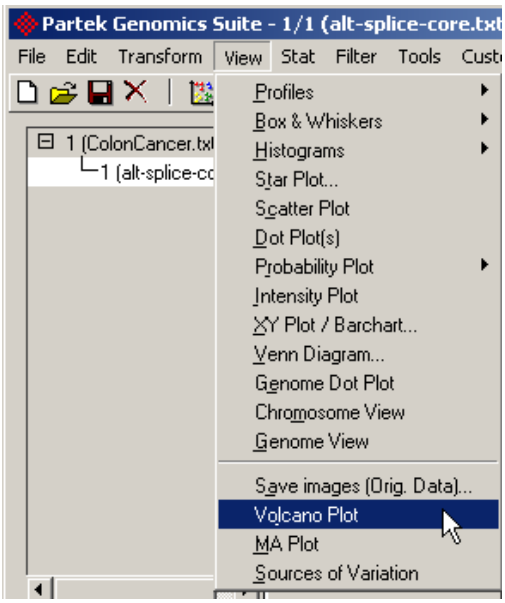


Figure 6. 140: Selecting the Volcano Plot from the View menu

By default, the X-axis is the ratio of means (MeanRatio) from the two subgroups, and the Y-axis is the p-value of the correspondent comparison. Each point on the plot represents a row of the results spreadsheet. The color of the point can be color coded by any of the columns in the results spreadsheet (Figure 6. 141).

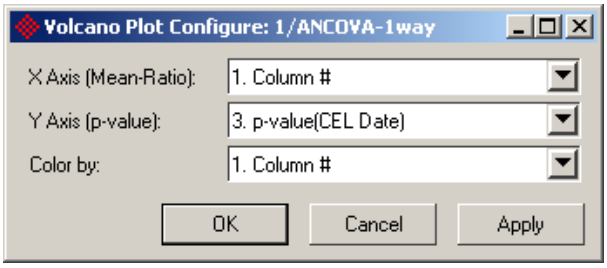


Figure 6. 141: Configuring the Volcano Plot dialog

Click **OK** to invoke the plot (Figure 6. 142).

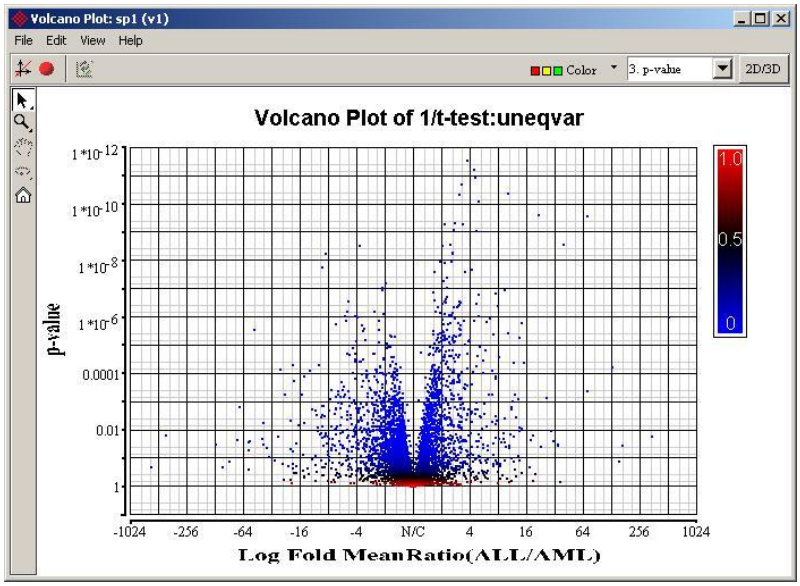


Figure 6. 142: Viewing the Volcano Plot of a Two Sample t-Test result

Volcano Plot Specific Menu Items

The File, Edit, View, and Help menus in the Volcano Plot viewer behave the same as the menus in the Scatter Plot viewer. Any differences will be notated below, otherwise see the **Viewing the Scatter Plot Results** section above.

The **Edit > Plot Properties > Style, Labels, Box & Whiskers, Titles, Axes, Color,** and **Labels** in the Volcano Plot behave the same as in the Scatter Plot - Plot Properties. Any differences will be notated below, otherwise see the **Scatter Plot Properties** section above.

The Mode buttons within the Volcano Plot viewer behave the same as in the Mode buttons in the Scatter Plot. Any differences will be notated below, otherwise see the **Miscellaneous Viewer Options** section above.

The Volcano Plot will start in mixed mode if the number of rows in the results spreadsheet exceeds the *Auto Mixed Mode* threshold set on the *Lines & Cursors* tab of the **Edit > Preferences** dialog.

The MA Plot

The MA Plot is a special 2-D scatter plot to visualize the statistical results on the significance and amount of changes in all the tested variables between two different levels (groups) of a factor. The X-axis represents the original mean of the variable; it is on log2 scale. The Y-axis represents the fold change of the test variable under the two conditions; it is on log2 scale.

Invoking a MA Plot

The MA plot can only be invoked on the results spreadsheet of the ANOVA or the two-sample t-Test. By default, the result of the two-sample t-Test includes the p-value, mean of each group, and the mean ratio of the two groups. In ANOVA, however, the default results spreadsheet does not contain the mean and mean ratio of the compared groups, but they can be specified in the *Pairwise Comparisons* dialog invoked by clicking the **Result...** button in the ANOVA dialog. When the results spreadsheet is active, click **View > MA Plot** from the Partek main window (Figure 6. 143), this will invoke the *MA Plot Configuration* dialog (Figure 6. 144).

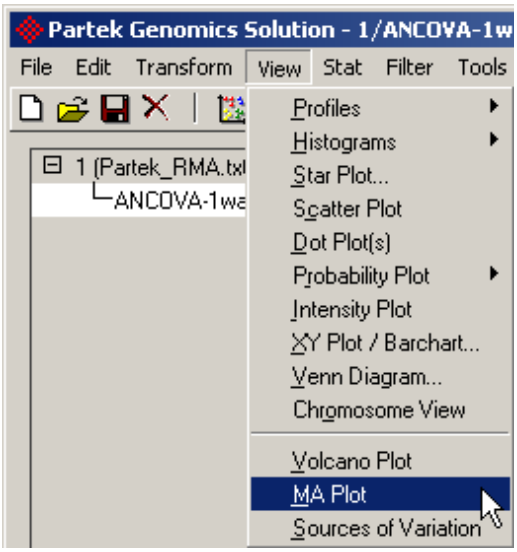


Figure 6. 143: Selecting the MA Plot from the View menu

The X-axis of the MA plot contains the original mean of the data. If the parent spreadsheet is available then this column is added if it is not already in the spreadsheet. If the spreadsheet does not contain the original mean and the parent spreadsheet is not available, then the MA Plot cannot be drawn. By default, the Y-axis is the ratio of means (MeanRatio) from the two subgroups. Each point on the plot represents a row of the results spreadsheet. The color of the points can be color coded by any of the columns in the results spreadsheet.

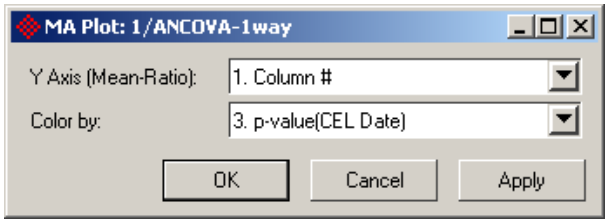


Figure 6. 144: Configuring the MA Plot

Select **OK** within the *MA Plot / ANCOVA-1 way* dialog to invoke the plot (Figure 6. 145).

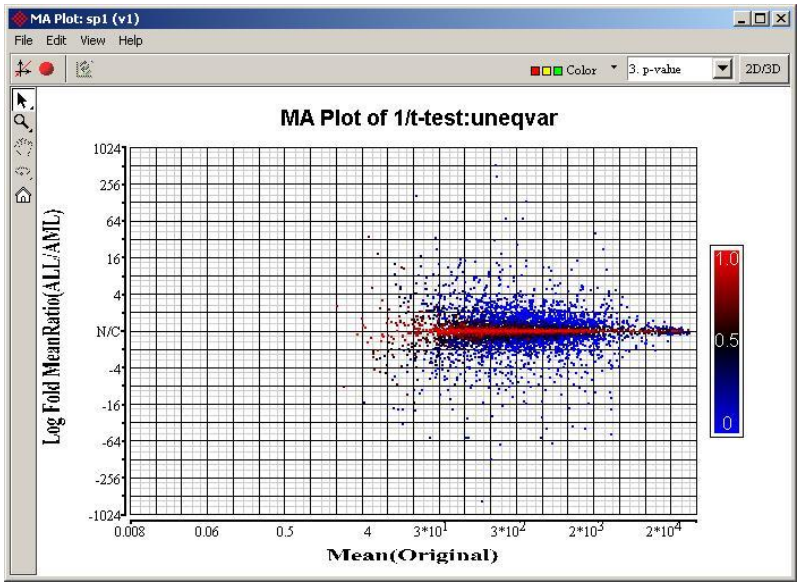


Figure 6. 145: Viewing a MA Plot of a Two Sample t-Test result

MA Plot Specific Menu Items

The File, Edit, View, and Help menus in the MA Plot viewer behave the same as the menus in the Scatter Plot viewer. Any differences will be notated below, otherwise see the **Viewing the Scatter Plot Results** section above.

The **Edit > Plot Properties > Style, Labels, Box & Whiskers, Titles, Axes, Color,** and **Labels** in the MA Plot behave the same as in the Scatter Plot - Plot Properties. Any differences will be notated below, otherwise see the **Scatter Plot Properties** section above.

The Mode buttons within the MA Plot viewer behave the same as in the Mode buttons in the Scatter Plot. Any differences will be notated below, otherwise see the **Miscellaneous Viewer Options** section above.

The MA plot will start in *Mixed Mode* if the number of rows in the results spreadsheet exceeds the *Auto Mixed Mode* threshold set on the *Lines & Cursors* tab of the **Edit > Preferences** dialog.

The Chromosome View 6.4

The Chromosome View 6.4 provides a way to examine the genomic location of probe sets. In addition, the genome view can summarize information by chromosome or show the location (by base pairs) of each probe set.

Creating the Chromosome View 6.4

The views are invoked from the *View* menu. The profiles can be configured on the *Profiles* tab of the *Plot Properties* dialog.

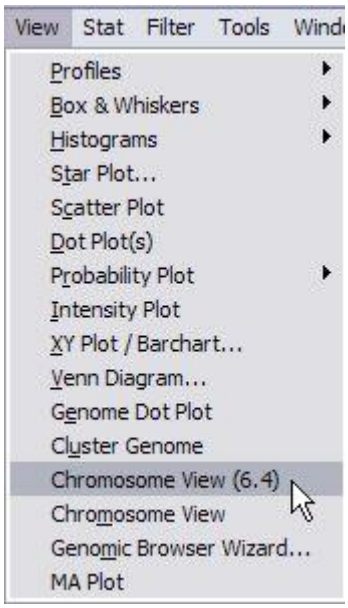


Figure 6. 146: Selecting the Chromosome View 6.4 from the View menu

The *Chromosome View 6.4* plots the mean of each probe set across the entire genome (Figure 6. 147).

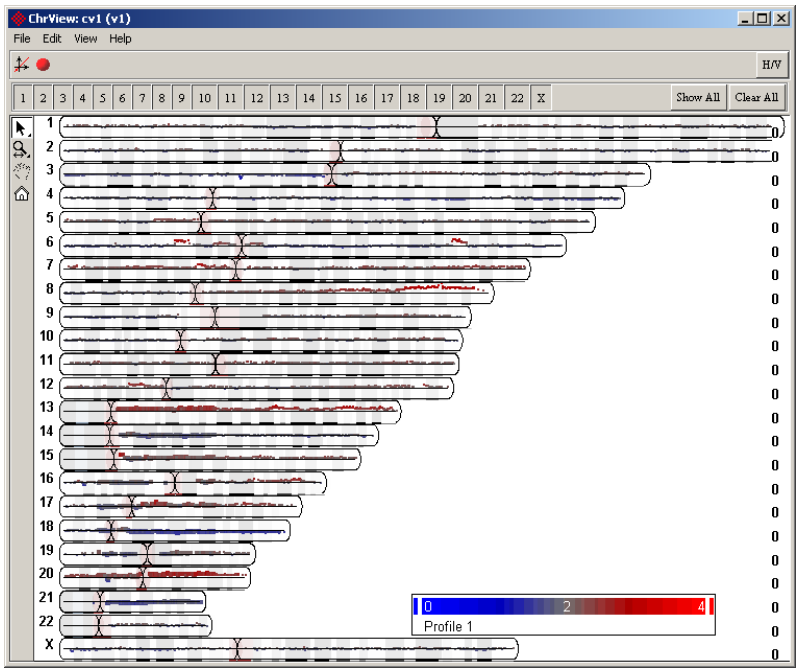


Figure 6. 147: Genome View (horizontal)

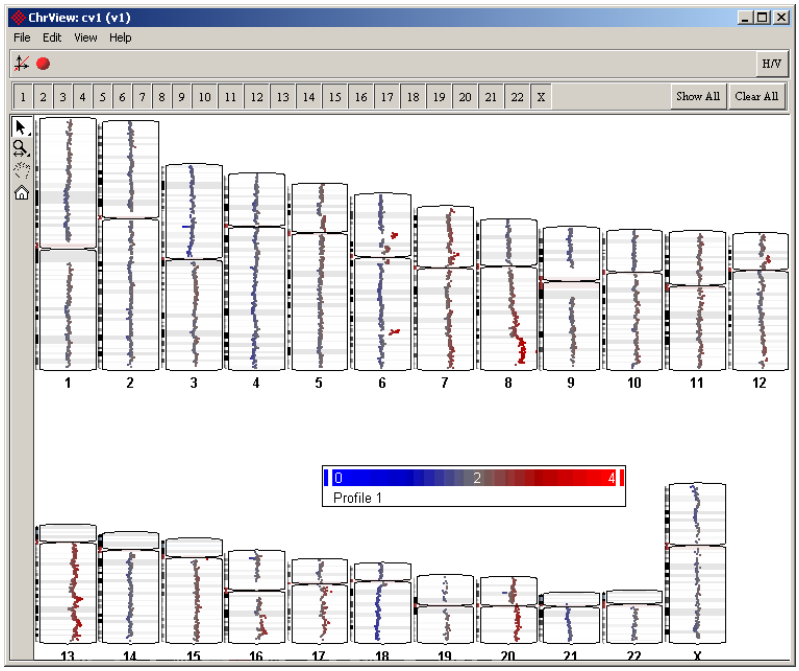


Figure 6. 148: Genome View (vertical)

The *Chromosome View* is invoked on the first chromosome with a profile for the selected sample(s) on top and a heat map with all samples on bottom (Figure 6. 149).

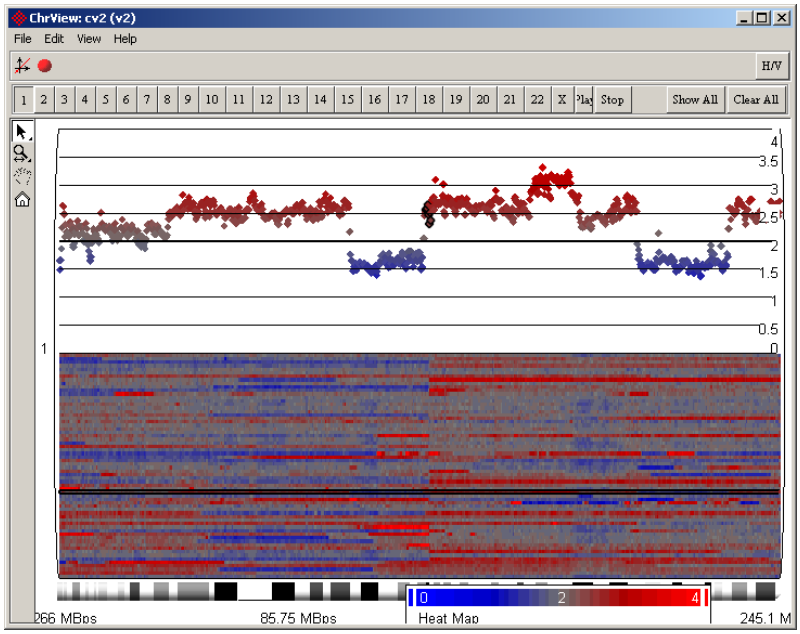


Figure 6. 149: Chromosome View

The *Genome Dot Plot* is most appropriate for copy number. It shows the genome in-line with a marker for every 100,000 base pairs. The length of each marker is determined by the number of samples with an average in that area exceeding the

threshold. The dot plot is configured on the *Dot Plot* tab of the *Plot Properties* dialog (Figure 6. 150).

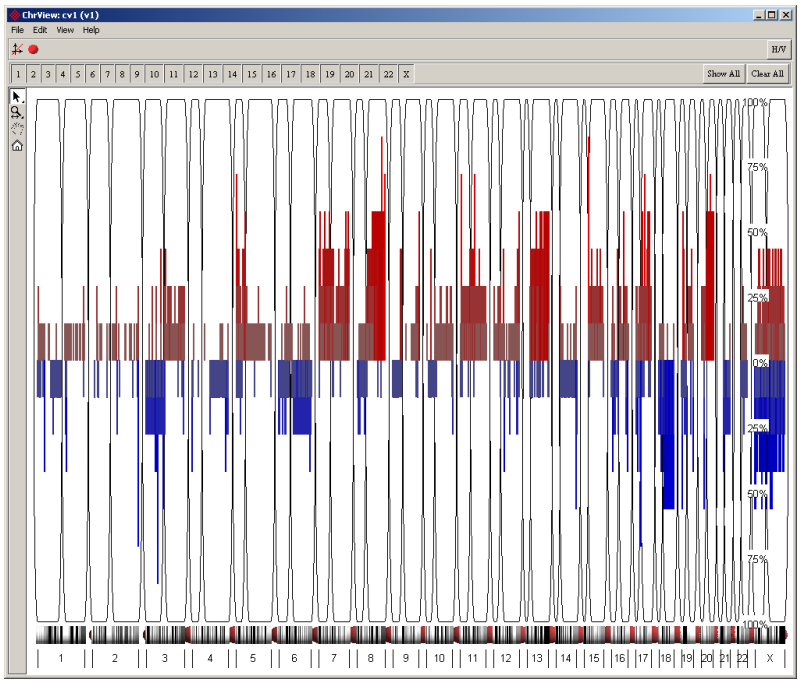


Figure 6. 150: Genome Dot Plot

The *Region View* is only available for spreadsheets, such as .bed files, that have the region property (**File > Properties**) (Figure 6. 151 & Figure 6. 152).

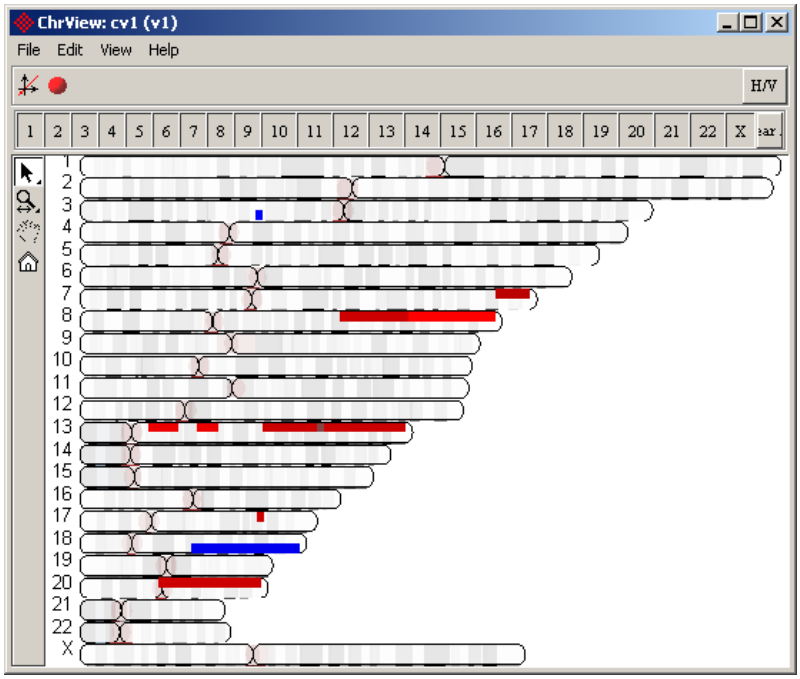


Figure 6. 151: Region Genome View

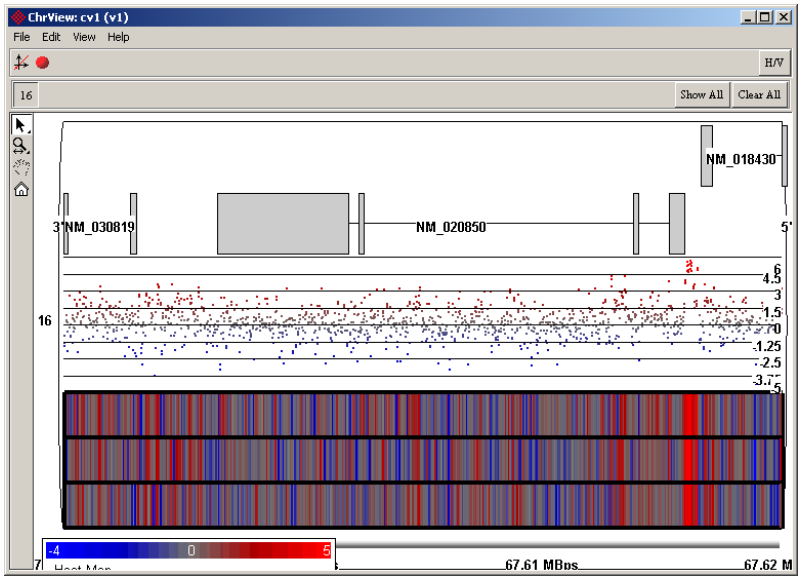


Figure 6. 152: Region View

In order to invoke genomic views, the spreadsheet must be properly associated with the correct annotation file.

- Select **File** > **Properties** to add or edit the association

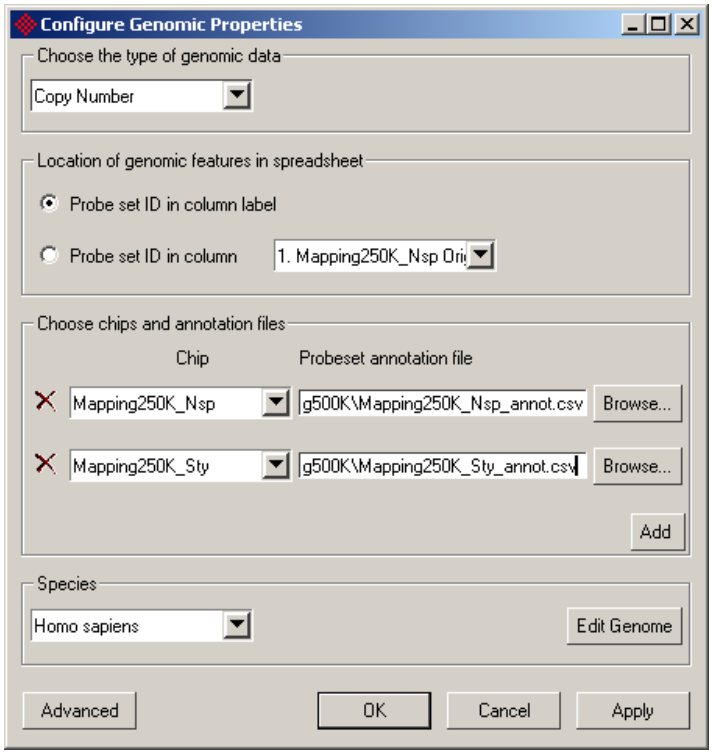


Figure 6. 153: Editing Genomic Properties of spreadsheet 1

Partek comes with configurations for several species. You can click **Edit Genome** within the *Configure Genomic Properties* dialog to configure a new species or to update an existing entry. The *Edit Genome* dialog is shown in Figure 6. 154 below.

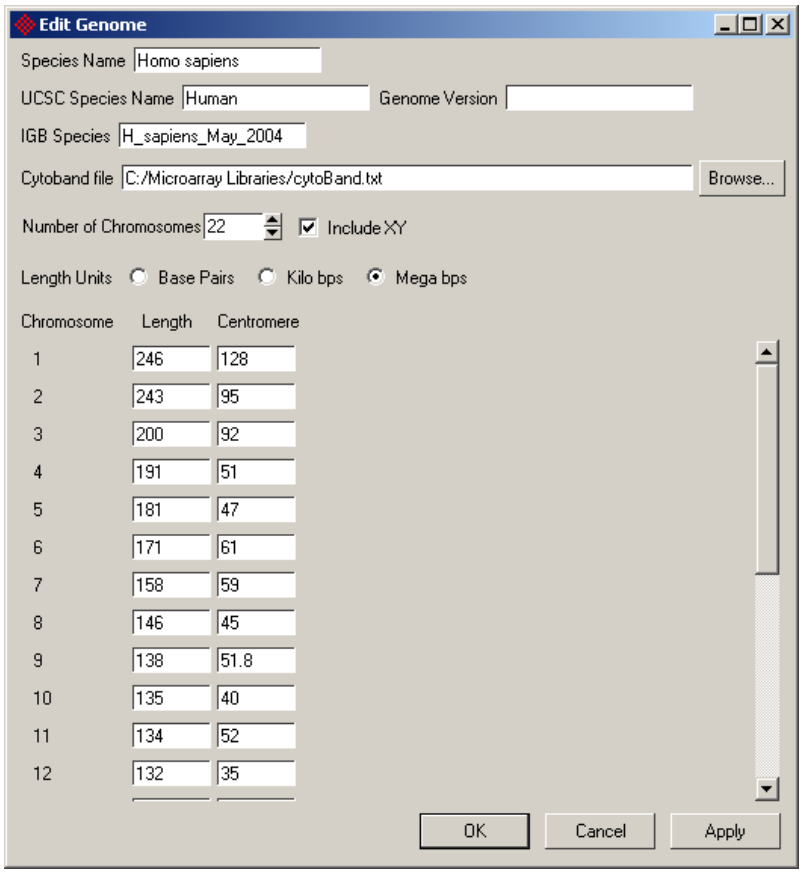


Figure 6. 154: Associating a species name using the Edit Genome dialog

Cytoband files can be obtained from UCSC: http://hgdownload.cse.ucsc.edu/downloads.html. On the UCSC page, choose the species, then the genome version, then click on the Annotation database link and download the cytoband file and unzip it.

If the annotation file for this spreadsheet is not in the *Annotation File* combo box, then click **Add New Annotation File…** You will only have to do this once per annotation file. The latest Affymetrix® annotation files can be obtained from http://www.affymetrix.com/support/technical/byproduct.affx.

After applying the *Configure Annotation* dialog, saving the spreadsheet will save the annotation file association. The annotation can be edited by selecting **File > Properties** from the Partek main window.

Looking at Chromosomes

When chromosomes are stained, they reveal light and dark bands. These bands are used in specifying gene locations on the chromosome. The centromere is the region that connects sister chromatids. The p arm is the shorter arm extending from the centromere.

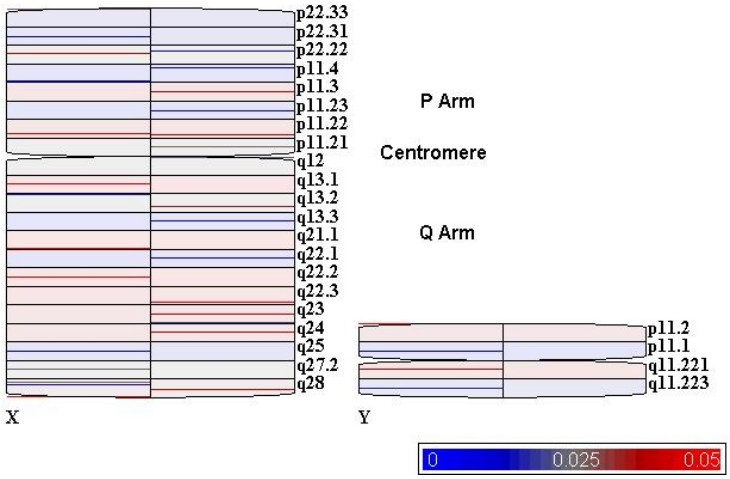


Figure 6. 155: Viewing Chromosome features

Chromosome View 6.4 Specific Menu Items

The File, Edit, View, and Help menus in the Genome View behave the same as the menus in the Scatter Plot viewer. Any differences will be notated below, otherwise see the **Viewing the Scatter Plot Results** section above.

The **Edit > Plot Properties > Style, Labels, Box & Whiskers, Titles, Axes, Color,** and **Labels** in the Genome Viewer behave the same as in the Scatter Plot - Plot Properties. Any differences will be notated below, otherwise see the **Scatter Plot Properties** section above.

The Mode buttons within the Genome View behave the same as in the Mode buttons in the Scatter Plot. Any differences will be notated below, otherwise see the **Miscellaneous Viewer Options** section above.

Configuring the Chromosome View 6.4

In the *Configure Plot* dialog for the chromosome view, you can choose which chromosomes to show. The number of probe sets is displayed next to each chromosome label (Figure 6. 156).

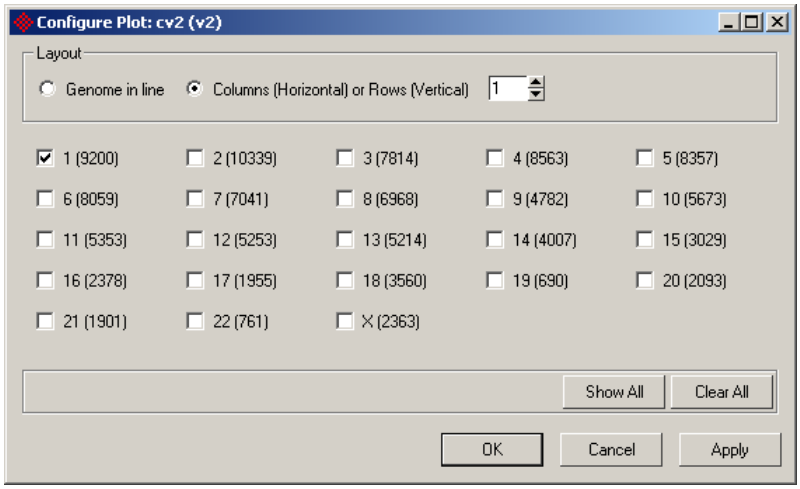


Figure 6. 156: Configuring the Genome View Plot dialog

If probe sets are on rows, you can add and remove *Criteria* to determine which probe sets are shown. By default, a criteria based on p-value will be added. This can be configured on the *Lines & Cursors* tab of the *Preferences* dialog (**Edit > Preferences** from the Partek main window)

If the *Layout* is set to **Genome in line** then all chromosomes will be drawn in succession horizontally (Figure 6. 157).

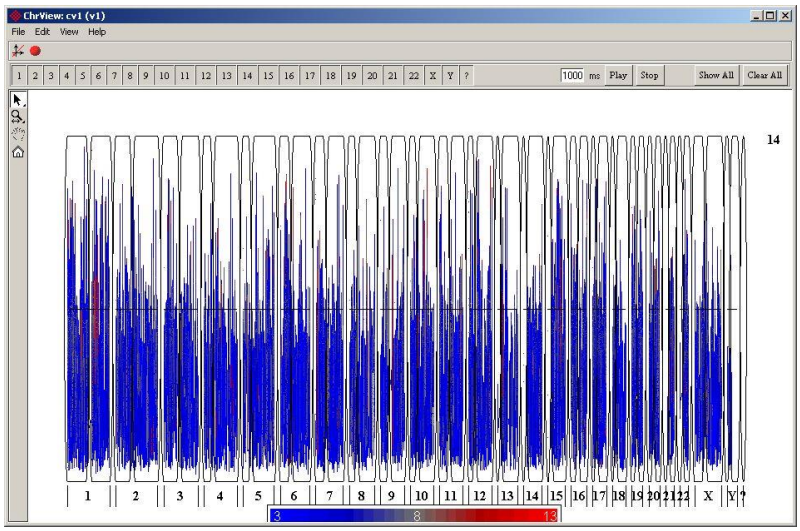


Figure 6. 157: Viewing the Genome in a line

Style

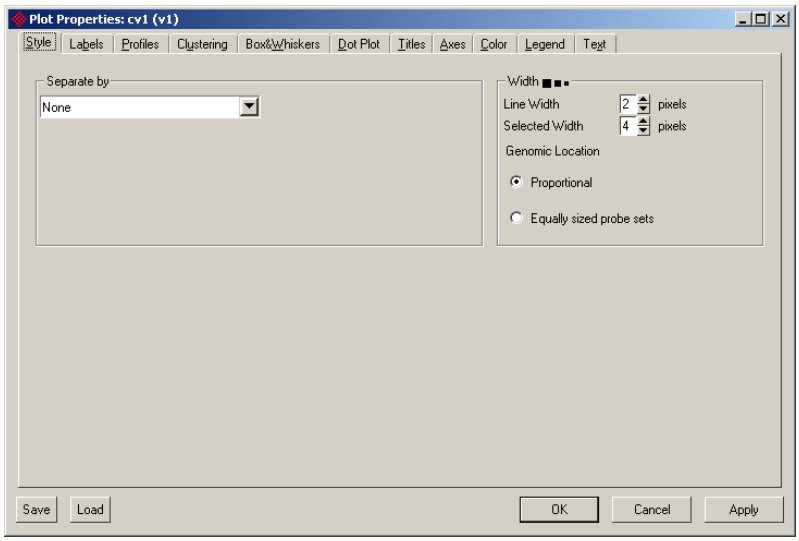


Figure 6. 158: Configuring the Style page

Width

These spin boxes determine the size of points and width of lines for profile and probe set styles. The width is specified in pixels (Figure 6. 159).

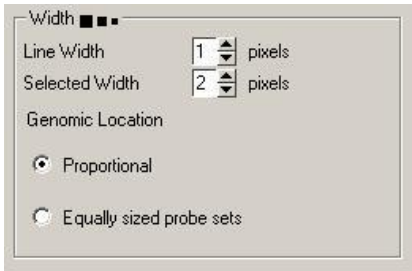


Figure 6. 159: Configuring the Width panel

If *Genomic Location* is **Proportional** then the location of each probe set is based on the base pair location (Figure 6. 160a). If *Genomic Location* is **Equally sized probe sets** then each probe set will be drawn in order flush with the next (Figure 6. 160b). The region of each probe set on a given chromosome will be the same. The size of chromosomes will remain the same (based on number of base pairs on the chromosome).

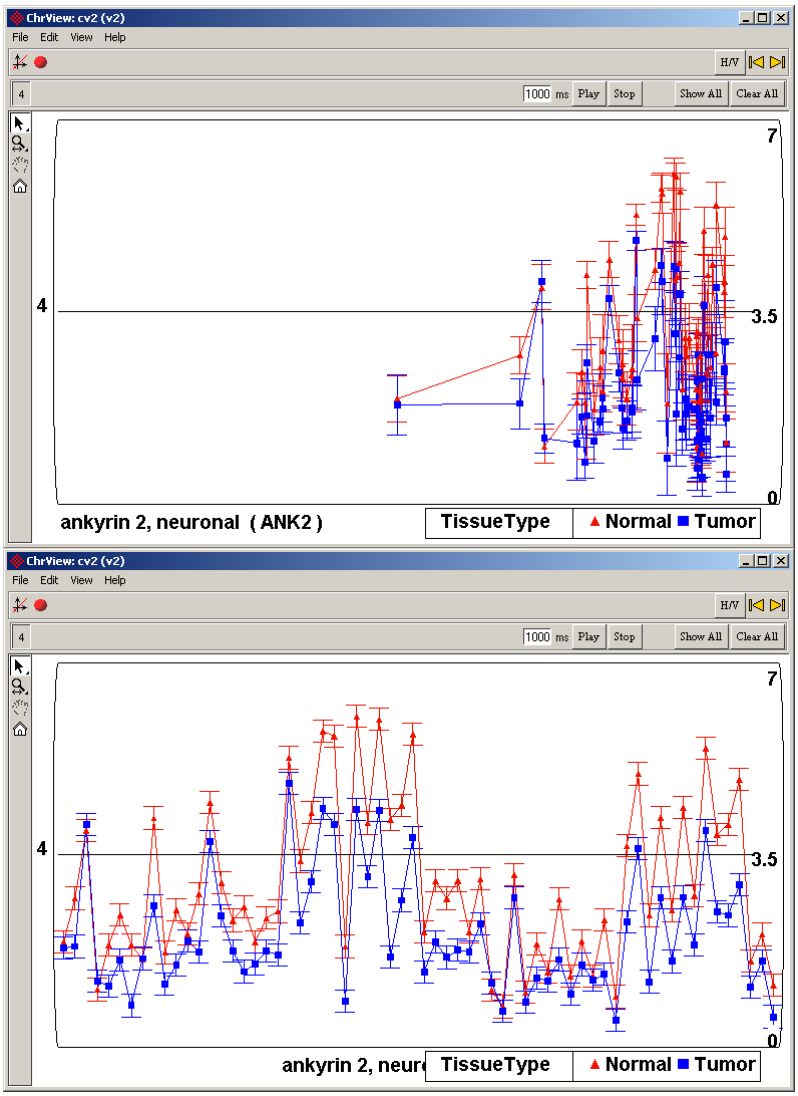


Figure 6. 160: Viewing the Genomic Location Proportional (top) and Equally sized probe sets (bottom)

Separate by

The *Separate by* combo box allows you to display a separate chromosome for each category in a column (Figure 6. 161).

Figure 6. 161: Configuring the Separate by panel

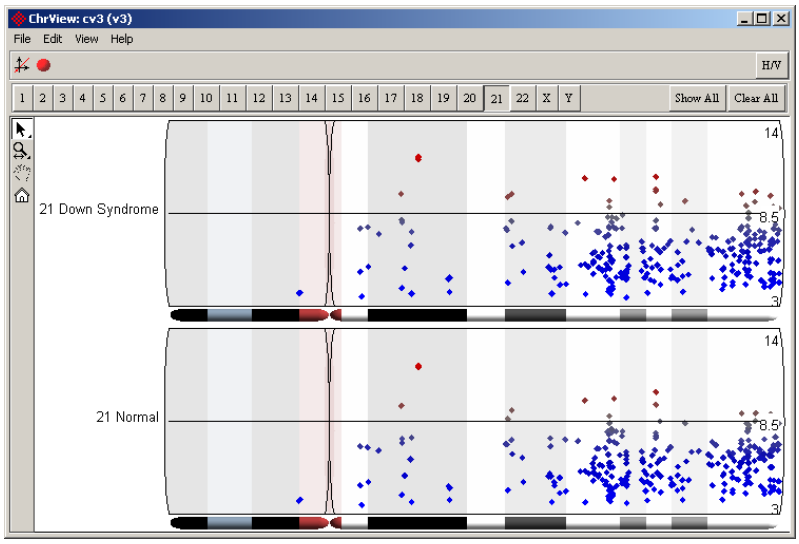


Figure 6. 162: Viewing Chromosome 21 separated by disease type in the Genome View

Labels

On the *Labels* tab you can configure how much of the label to show if you have the plot separated by a nominal variable (on the *Style* tab) (Figure 6. 163).

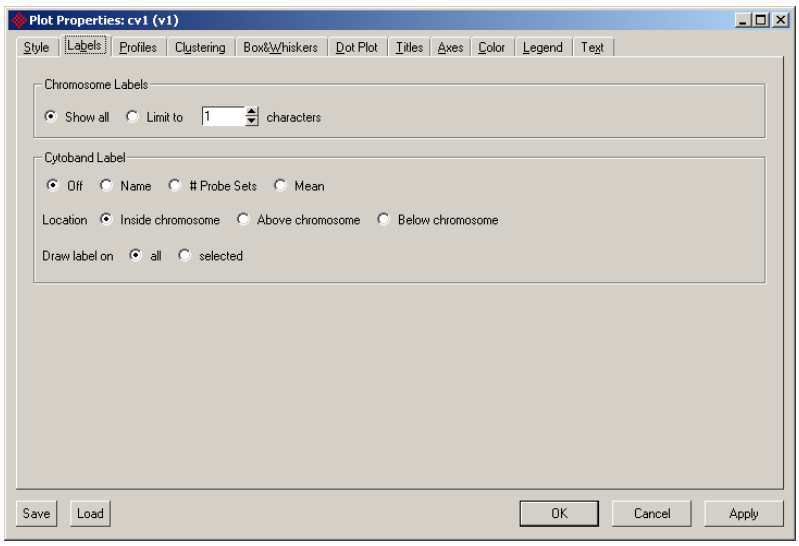


Figure 6. 163: Configuring the Labels page

Profiles

Profiles and heat maps can be added and removed in the *Profiles* tab. Select **Configure** to change the properties of an existing profile (Figure 6. 164). Check **HMM Smoothing** to overlay HMM states on the by-sample profile. See Chapter 10 for more information on HMM Smoothing. A profile with HMM overlay is shown in Figure 6. 165.

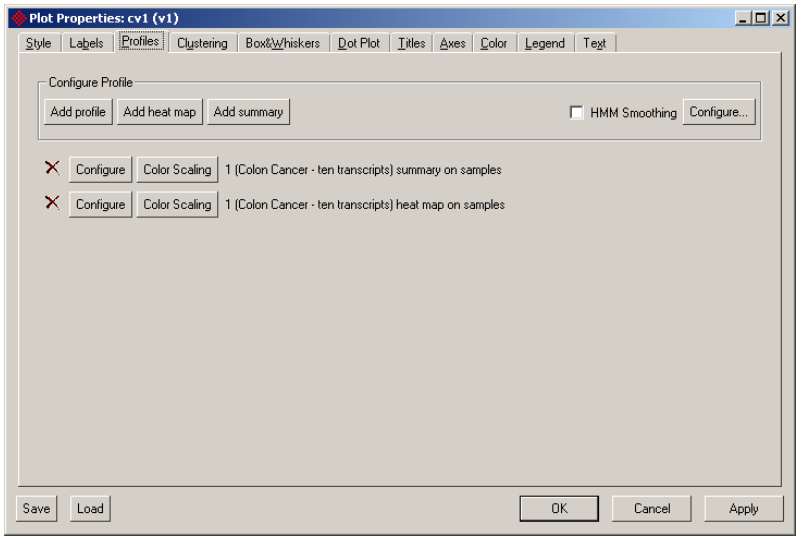


Figure 6. 164: Configuring the Profiles page

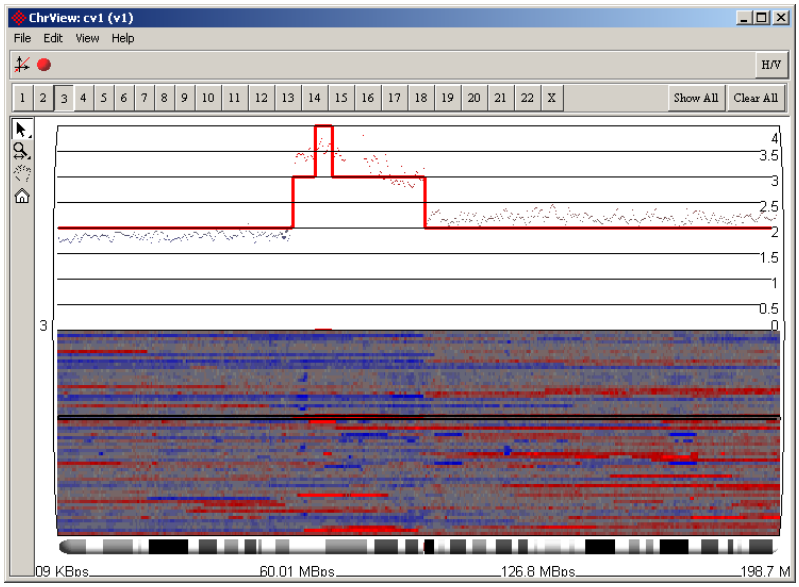


Figure 6. 165: Profile with HMM overlay

On the top line, you can select any spreadsheet, which is linked to genomic locations. If you draw the profile on the mean then there will be one line with the height determined by the mean of the column (if genes are on columns) or the mean of the row (if genes are on rows).

If genes are on columns, then you can separate the profile by a categorical variable, which will result in one line for each category of that column. If genes are on rows, you can specify a column from which the profile will derive its height.

If genes are on columns, then you also have the option to *Separate by sample*, which draws a line for each row in the spreadsheet.

If genes are on columns, then you can color the profile by any spreadsheet, which uses the same annotation file. If genes are on rows then the color spreadsheet must be the same as the profile spreadsheet.

You can color by any numeric column if genes are on rows and by any categorical column if the profile is separated by sample.

The legends for the profiles are edited on the *Text* tab. You can double click a legend to go to the appropriate configuration page.

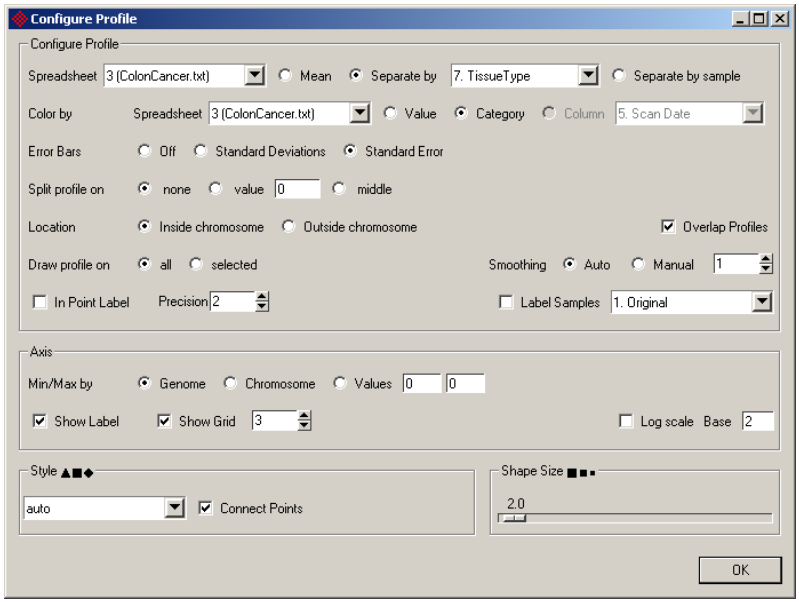


Figure 6. 166: Configuring the Chromosome View 6.4 Profile dialog

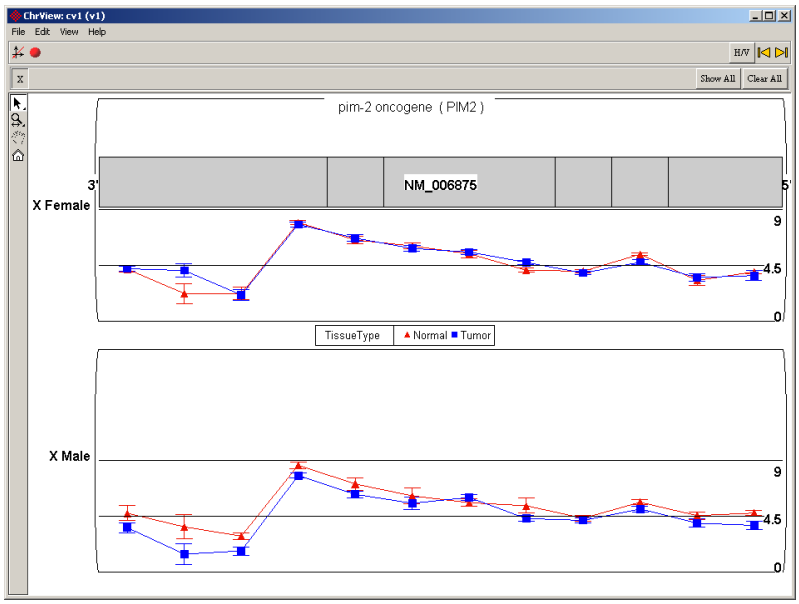


Figure 6. 167: Viewing a profile split by tissue type inside a chromosome separated by gender

Clustering

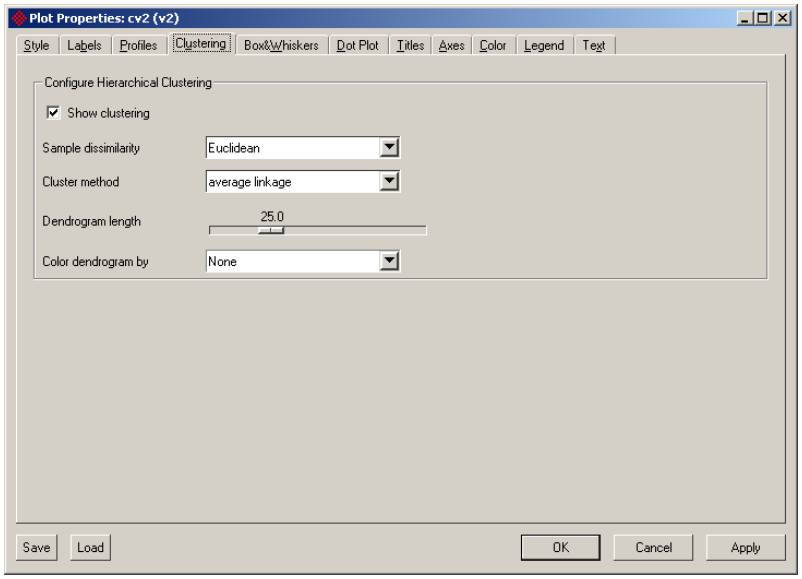


Figure 6. 168: Clustering page

On the *Clustering* tab, you can add a hierarchical clustering dendrogram to the genomic view. You can configure how to cluster and how much screen space to give to the dendrogram. See Chapter 8 for more information on hierarchical clustering.

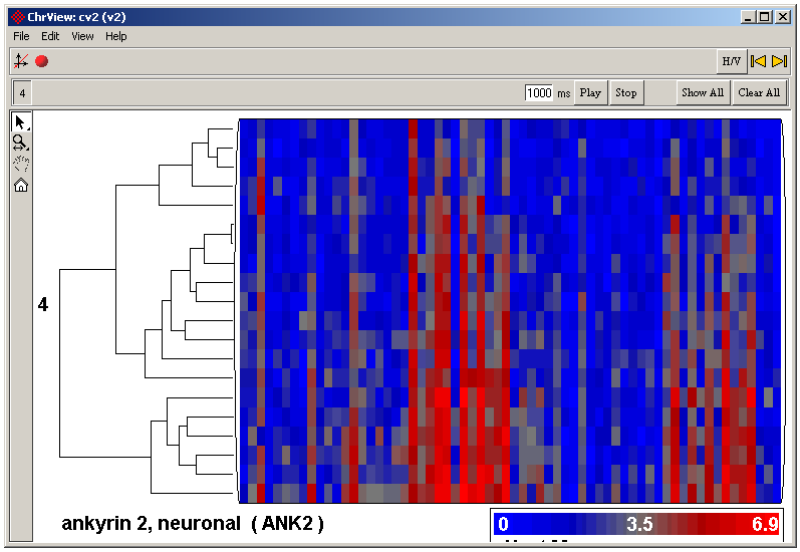


Figure 6. 169: Dendrogram on gene view

Box & Whiskers

On the *Box and Whisker* tab, you can add Box & Whiskers to visualize the distribution of the data (Figure 6. 170).

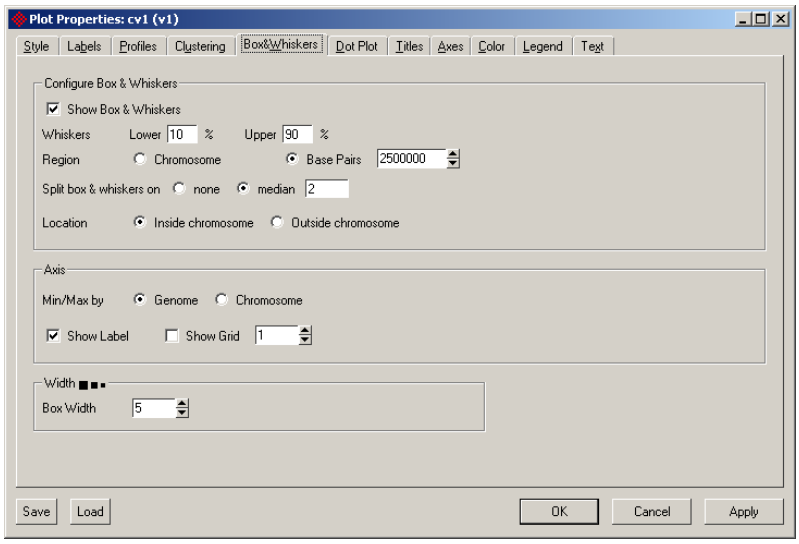


Figure 6. 170: Configuring the Box & Whiskers page

If the box and whiskers are split on a value, then the box and whiskers for a given region will be drawn above the center if the median is greater than the given value and below the center if the median is less than the given value. The center value therefore will be two different values when the biggest value in the upper percentile below the center is greater than the smallest value in the lower percentile.

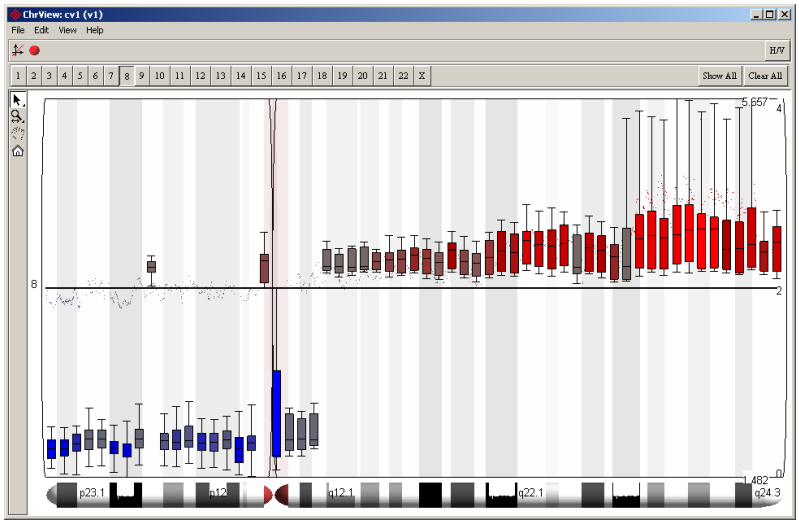


Figure 6. 171: Viewing Box & Whiskers per 0.25 MBp, separated on 2

Dot Plot

Dot Plots are another method to visualize the distribution of the data.

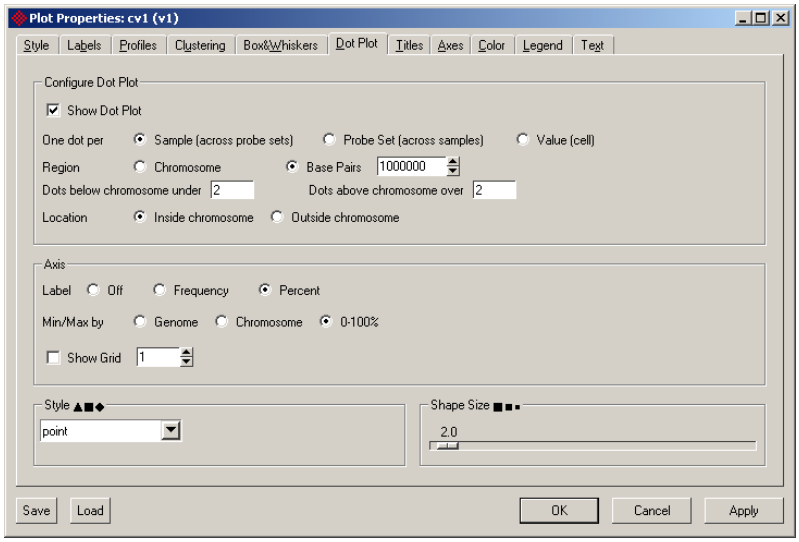


Figure 6. 172: Configuring the Dot Plot page

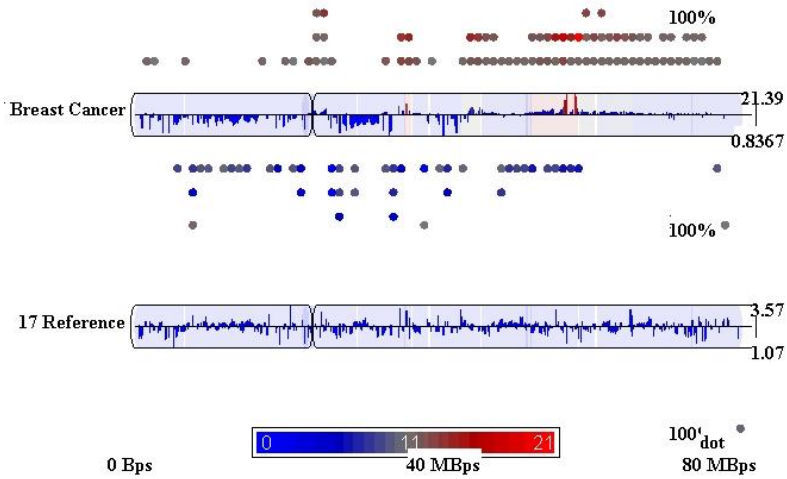


Figure 6. 173: Viewing a Dot Plot on copy number data

If one dot is drawn per sample, then the height of the stack will never exceed the number of samples. If one dot is drawn per probe set then the maximum height of the stack is equal to the number of probe sets in the region. If one dot is drawn per value then the maximum height of the stack is equal to the number of probe sets in the region times the number of samples.

A region of one base pair will guarantee that every probe set is in a unique region. If the region is greater than one base pair then a probe set will be assigned to a region based on its middle.

Orientation – Vertical

You can toggle the orientation by clicking the **H/V** (horizontal/vertical) button on the chromosome toolbar (Figure 6. 174).

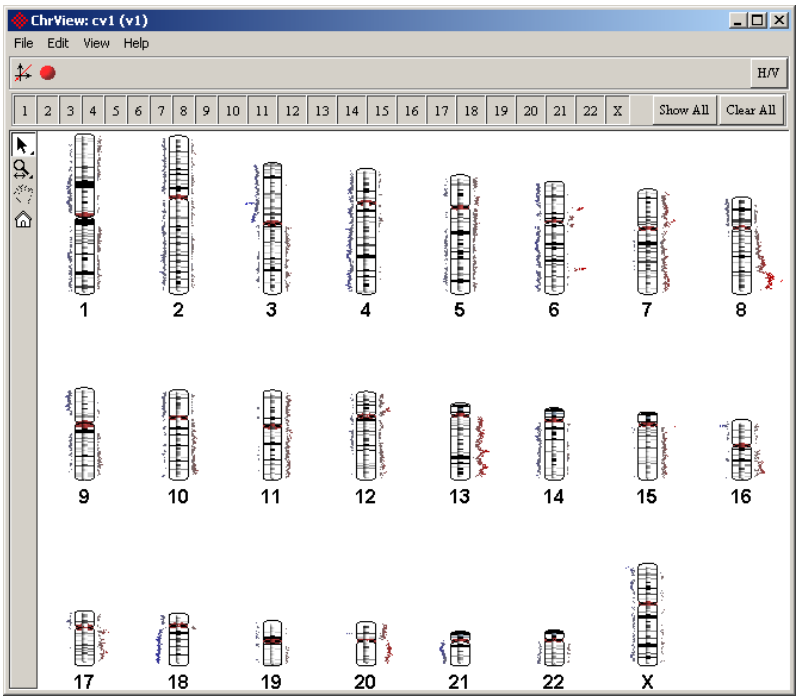


Figure 6. 174: Viewing Vertical chromosomes

The default orientation can be configured from the *Preferences* dialog (**Edit > Preferences** of the spreadsheet window).

File > Save Images from Regions

From the *Save Images from Regions* dialog, you can save .jpeg images for each region in a database file (Figure 6. 175).

Figure 6. 175: Saving Images from Regions

File > Dump to Spreadsheet

Once you have zoomed into a region or selected interesting probe sets, you can create new spreadsheets that summarize the region.

Acceptable choices for the *Gene Database* include annotation files, .bed files, and .pgx files.

The .pgx format is (tab-delimited):
name/key chromosome start end (optional additional columns)

Choosing **Wiggle file for UCSC** will invoke a browser with the UCSC data upload page and put the output file name in your clipboard.

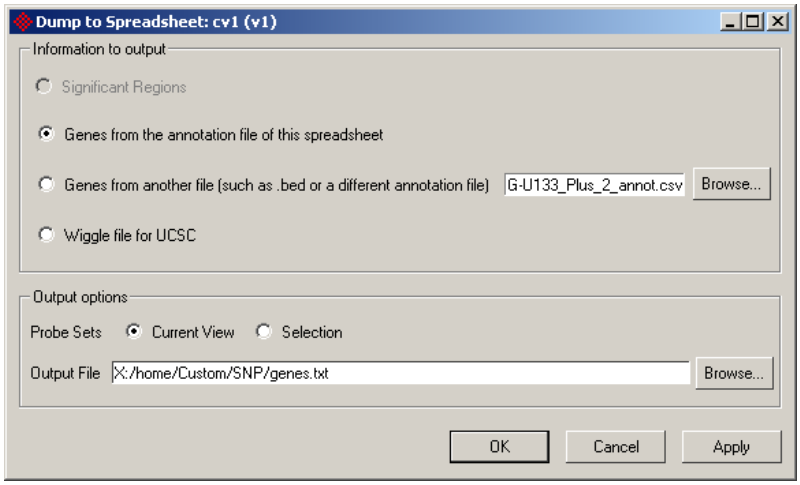


Figure 6. 176: Dumping to a new spreadsheet

File > Add Genomic Features

From the *Add Genomic Features* dialog, you can add regions from a database or a region spreadsheet (Figure 6. 177).

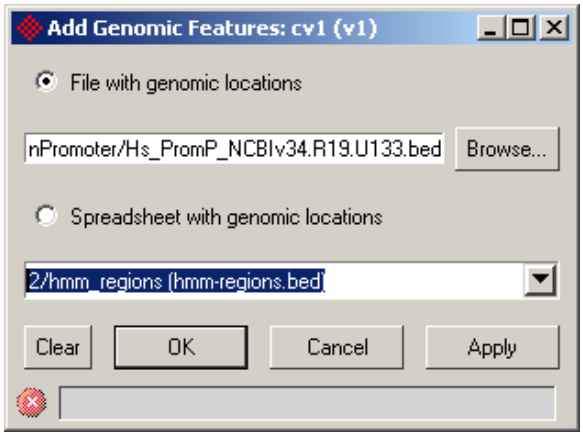


Figure 6. 177: Adding genomic features

The appearance and labeling of these regions can be configured from the *Plot Properties* dialog in the genes panel on the *Axes* tab.

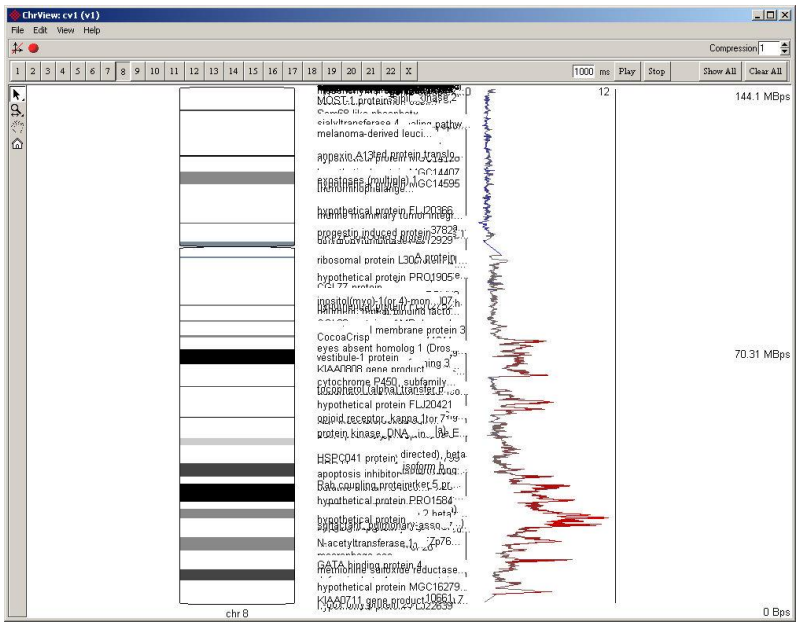


Figure 6. 178: Viewing Chromosome 8 labeled with reference genes

View > Zoom

You can manually specify start and end zoom coordinates (Figure 6. 179) or you can select a chromosome and a feature to automatically fill in the entries.

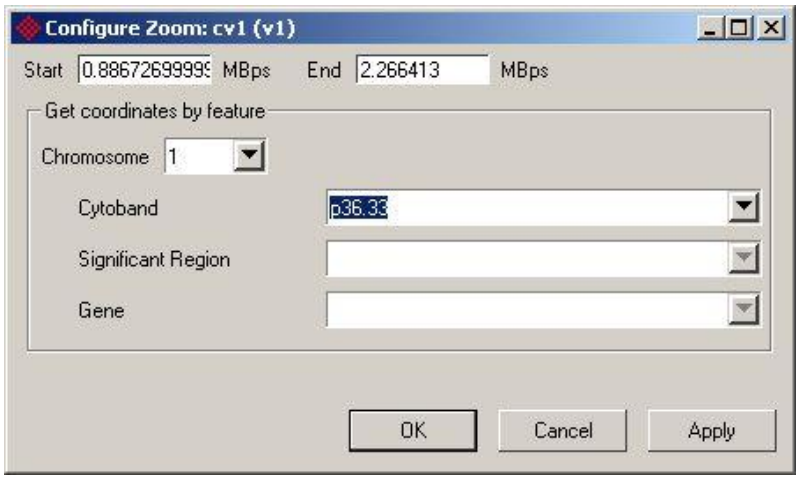


Figure 6. 179: Configuring the Zoom

View > HTML Report

Once you have zoomed into a region or selected interesting probe sets then you can create HTML reports to discover more information about the region.

Click **Choose annotations** to show a list of columns in the annotation file (Figure 6. 180).

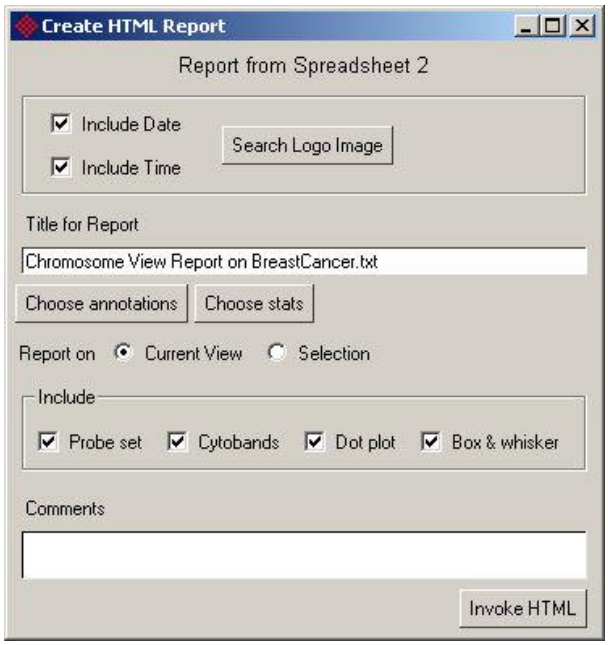


Figure 6. 180: Creating an HTML Report

The HTML report contains links to a number of resources including the NCBI, the UCSC Genome browser, and Affymetrix IGB.

Chromosome View 6.4 Specific Toolbars

The genome view has 3 additional tool bars: chromosome, profile and dot plot. The chromosome tool bar is on by default, the other two are turned on using the **View > Toolbar** entries.

The Chromosome toolbar hides and shows the chromosomes. Left click to toggle and right click to show only that chromosome (Figure 6. 181).

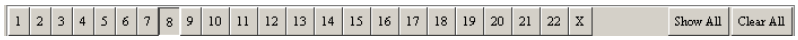


Figure 6. 181: Viewing the Chromosome navigation tool bar

On the Dot Plot toolbar, aspects of the Dot Plot can be configured (Figure 6. 182).

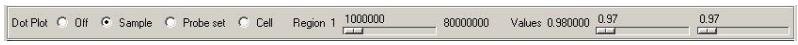


Figure 6. 182: Viewing the Dot Plot tool bar

Visualization of Multivariate Data

A common tool for visualizing multivariate data is the scatter plot.

- From the spreadsheet menu bar, select **View > Scatter Plot**. A new viewer will appear with a scatter plot in it (Figure 6. 183).

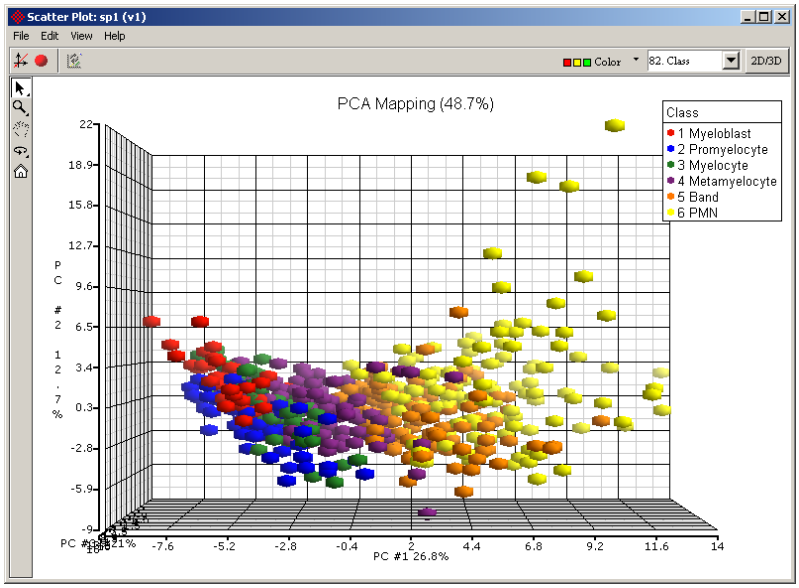


Figure 6. 183: Viewing a PCA mapped scatter plot of Blood Cell data

This scatter plot shows one point for each white blood cell (row in the spreadsheet). The points are color coded by the stage of the white blood cell. Note that the X, Y, and Z axes are labeled with PC #1, PC #2, and PC #3, respectively. This is because the data is high-dimensional and is being mapped to 3-D for display (in Partek, the default mapping uses Principal Components Analysis).

Data Mapping

Click the *Select Coordinates* icon on the viewer toolbar () to access the toolbar of Figure 6. 184. The mapping of the data onto the scatter plot is configured here. The data can be plotted using 2 or 3 columns or by using selected Principal Components (PC'S) of the original data.

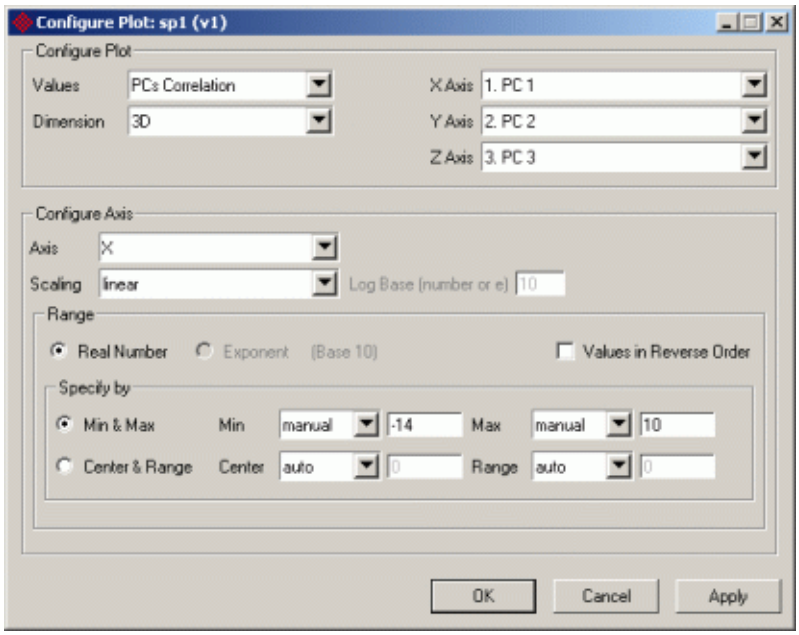


Figure 6. 184: Configuring the Scatter plot

Refreshing the Scatter Plot

Applying a row filter to the spreadsheet will not cause the PCs to be recomputed, but it will cause the indicator in the *Standard Toolbar* to become active. Selecting the refresh accelerator button or applying the *Configure Plot* dialog will recompute the PCs (Figure 6. 185).

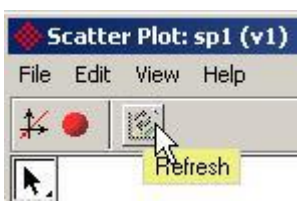


Figure 6. 185: Selecting the Refresh accelerator button

Filtering or deleting columns or deleting rows will cause the PCs to be recomputed. If variables rather than PCs are plotted, then clicking the **Refresh** button will update the axis minimums and maximums.

Picking

Points can be selected two ways: from the spreadsheet or from the graphics.

- Choose the *Standard Selection Mode* ( ) from the viewer tool bar.
- Select a row of the spreadsheet by clicking on the row heading. The row will be highlighted and the point will be selected in all displayed graphics. Hold down the **<Ctrl>** key and click on it again, it will be deselected both on the spreadsheet and in any graphics displayed from that spreadsheet.

- After making the first selection, hold down **<Ctrl>** to select multiple rows that are not next to each other; hold down **<Shift>** to select the rows that are next to each other. Clicking the upper left empty cell in the spreadsheet will deselect all.
- Left click on any point in the graphics and the spreadsheet will scroll to that row and select it. As a result, that row will be highlighted in all graphics. Press **<Ctrl>** and left click on the point again; it will be deselected in the graphics and on the spreadsheet. Holding down **<Ctrl>** or **<Shift>** while left clicking on the points in graphics will allow multiple selections. Clicking on an empty space in the graph will deselect the points and the corresponding rows in the spreadsheet.
- Hold down the left mouse key and drag the mouse to draw a bounding box on the graph. This will also allow multiple selections.
- Select the *User-defined Selection Mode* ( ). Here you can invoke user-defined operations to occur when selecting a point.

Chromosome View

The *Chromosome View* in Partek® Genomic Suite™ (Partek GS) (Figure 6. 186) is a visualization tool for genomic data. It is initialized by **View > Chromosome View.** It can be accessed from certain workflows under the *Visualization* tab by selecting **Plot Chromosome View**, or by <right-clicking> on a gene row header and selecting **Browse to Location**. The view can display multiple levels of genomic data simultaneously, including *Heat Maps*, *Reference Sequence Transcripts*, *Reference Genomes*, *Amplification* and *Deletion* sites and more. The chromosome view initially displays chromosome one (1) by default. The panel on the left allows you to customize the way the data is presented in the view.

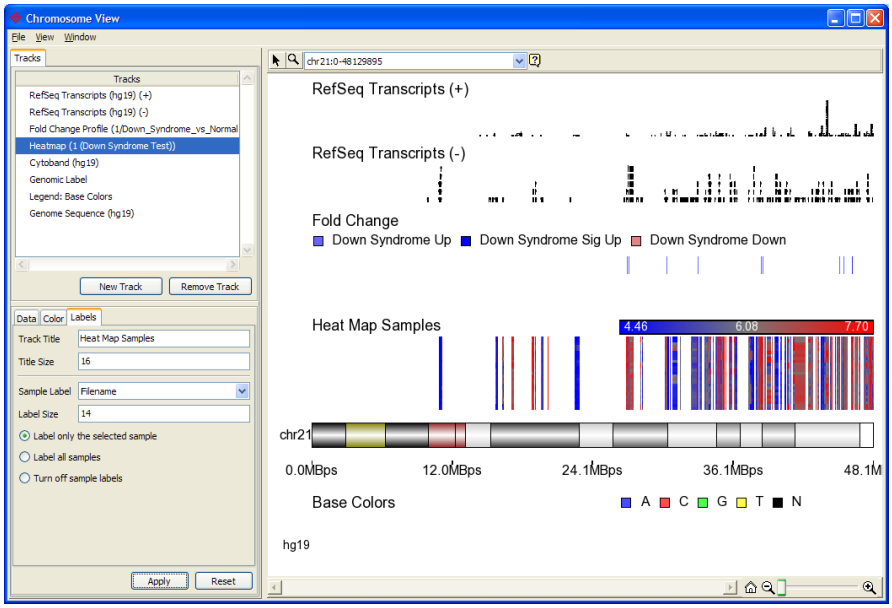


Figure 6. 186: Viewing the Chromosome View

When the *Chromosome View* is first initiated, you will be asked to choose a default annotation if one has not already been chosen though the analytical process (Figure 6. 187). There are three (3) options available for the human genome build 19 in Figure 6. 187, *RefSeq Transcripts*, *Ensembl Transcripts*, and *Do not download any file at this time.* Partek GS will attempt to automatically download the chosen annotation data. If *Do not download any file at this time* is chosen, only a Cytoband file from UCSC will attempt to be downloaded.

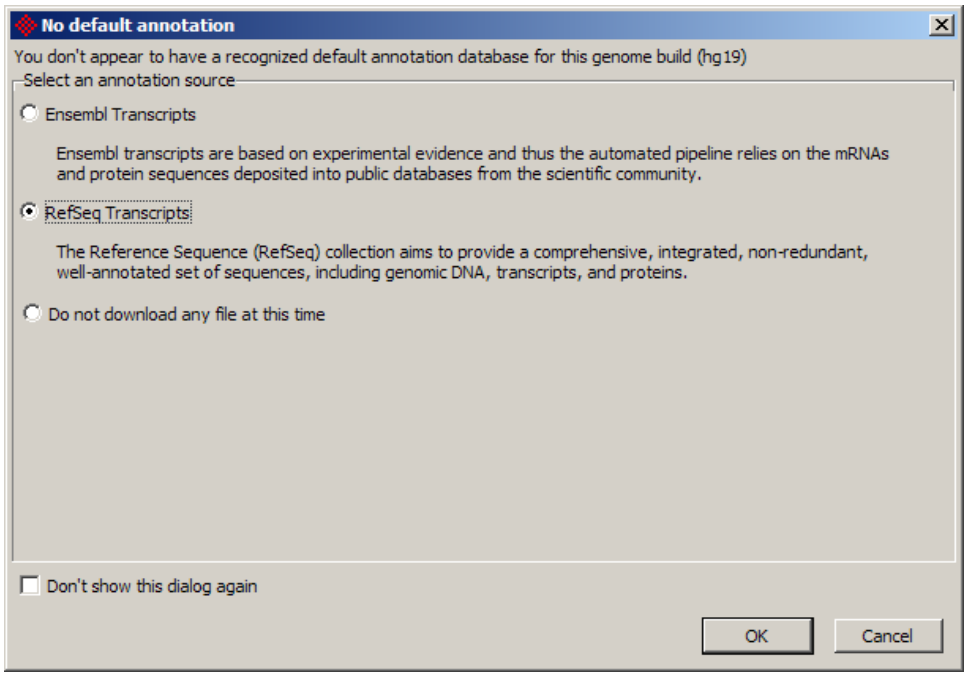


Figure 6. 187: Viewing Annotation options when opening the Chromosome View for the first time

Search Bar

The search bar can be found above the view (Figure 6. 188).

Figure 6. 188: Viewing the Search bar of Chromosome View

Use the search bar to zoom to genomic features that are available in annotation tracks. Type or paste in genomic positions such as chr6:40544957-49169085 or VEGFA (Figure 6. 189).

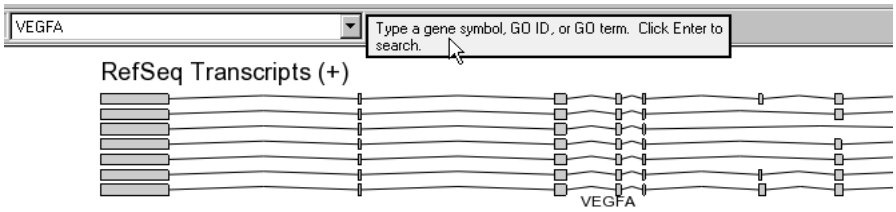


Figure 6. 189: Finding genomic features with the search box

The *Search* bar will display a dropdown list of the last ten searches (Figure 6. 190).

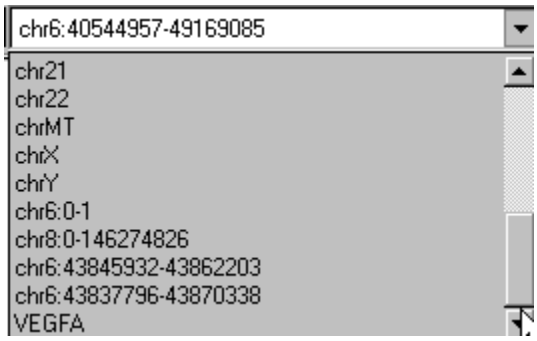


Figure 6. 190: Viewing the Search bar history

Modes

Selection Mode

<Left-click> on a track in selection mode (Figure 6. 191) to select the track in the track panel and to edit the properties of the selected track. Individual tracks have unique editable settings. <Right-click> on a track for an option to remove it. <Left-click> on heat maps to select samples. <Left-click> and drag on the Cytoband track to zoom to that region.

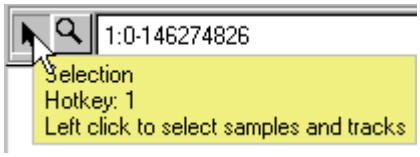


Figure 6. 191: Entering Selection Mode

Navigation Mode

In *Navigation* mode (Figure 6. 192), <left-click> to zoom in on a selected region. <Right-clicking> will re-center the plot.

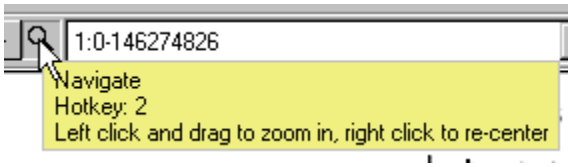


Figure 6. 192: Entering Navigation Mode

The mouse wheel will zoom in and out in both *Selection* and *Navigation* modes.

Mouse Over

Information about features under the mouse cursor (i.e. genes, sample ID, chromosome position) is displayed in the top-right corner (Figure 6. 193).

Figure 6. 193: Viewing MouseOver information

Use the slider (Figure 6. 194) or the magnifying glass button in the bottom right corner of the view to zoom in and out. Select the *Home* button to reset the view to the full chromosome.

Figure 6. 194: Viewing the Zoom Slider

Tracks

From the *Tracks* tab, you can add or remove tracks and configure the properties of tracks in the view (Figure 6. 195). With multiple tracks selected, changing common properties such as title or font size will be applied to all selected tracks, but unique properties such as track title will only be applied to the bottom selected track. Each track has editable parameters controlled by the tabs below track list. If changes are made to the track, select **Apply** to update the track with the changes. Select **Reset** to change the values back to the default settings.

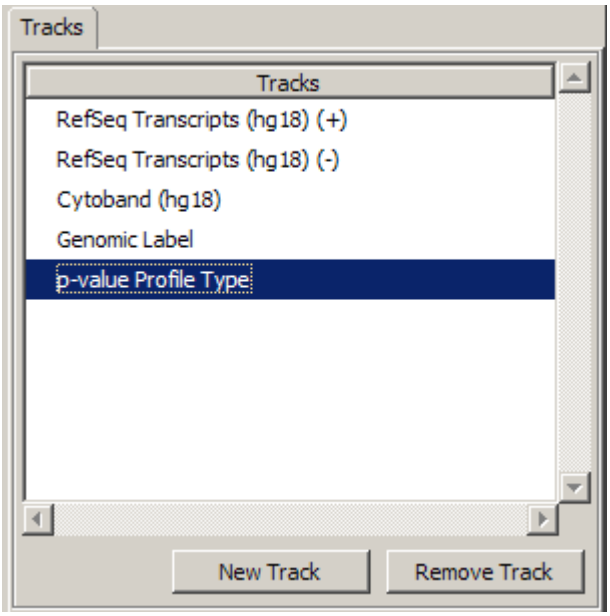


Figure 6. 195: Viewing the Tracks panel, which shows a list of tracks displayed in the view

New Track Options

Select **New Track** to prompt the *TrackWizard* dialog (Figure 6. 196), which shows the options to add new tracks.

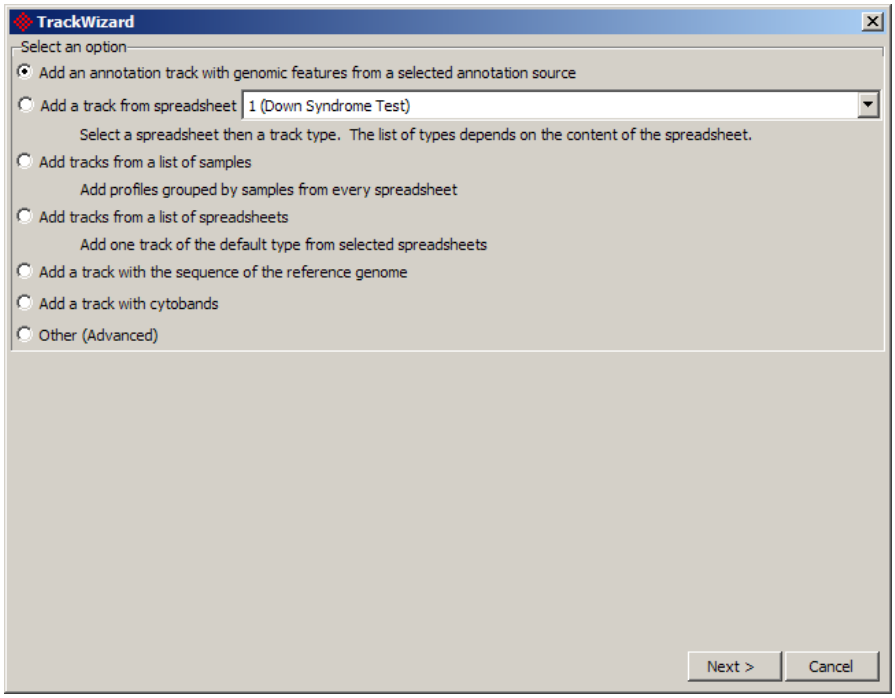


Figure 6. 196: Adding a new track with the TrackWizard

1) *Add an annotation track with genomic features from a selected annotation source* for a list of available annotations. Choose an available annotation and select *Create*, <u>or</u> select *Manage available annotations* to add a new annotation. Partek will attempt to automatically download the annotation file chosen if it is not already available on the local system, as indicated by the *Download Required* message highlighted in red (Figure 6. 197)

Annotation Track

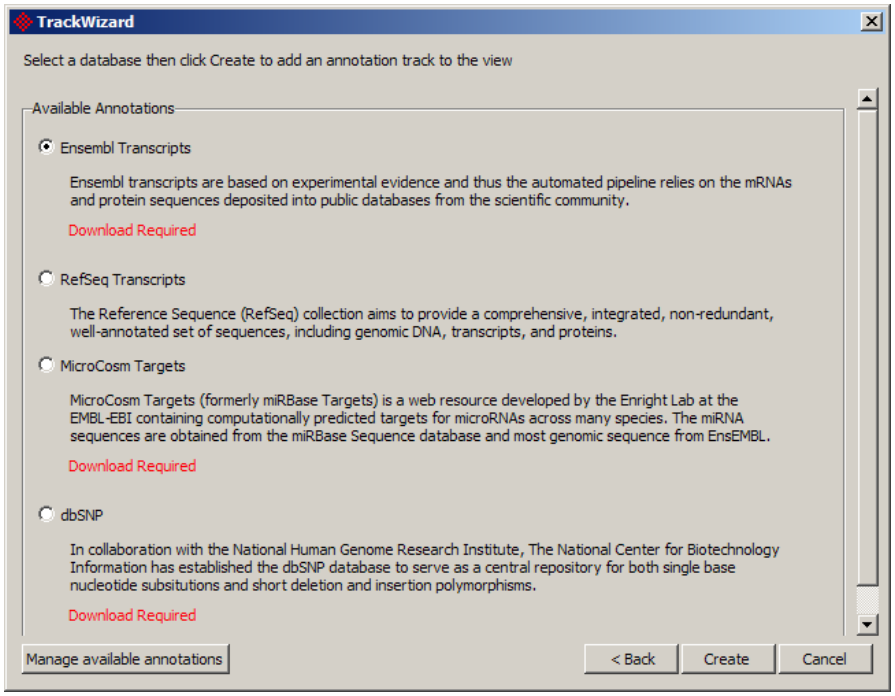


Figure 6. 197: Adding an annotation track to the Chromosome View

Figure 6. 198 is an example of one of the available annotation tracks Partek GS will attempt to download, *RefSeq*. For RefSeq, two gene annotation tracks are added, one filtered to the positive (+) strand (5' on the left and 3' on the right) and the other track filtered to the negative (-) strand (3' on the left and 5' on the right). An option to display both positive (+) and negative (-) strands on one track is available from the *Strand* drop-down in the *Tracks* panel (Figure 15).

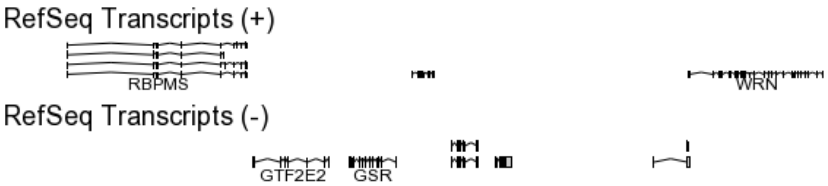


Figure 6. 198: Viewing Gene annotation tracks of positive and negative strands with every gene labeled

By default, each stack of genes is labeled with the *Gene ID* at the bottom of the track (Figure 6. 199). This is a result of the *Label every gene* option selected as the default setting. *Label every isoform* will draw the transcript id on top of each transcript.

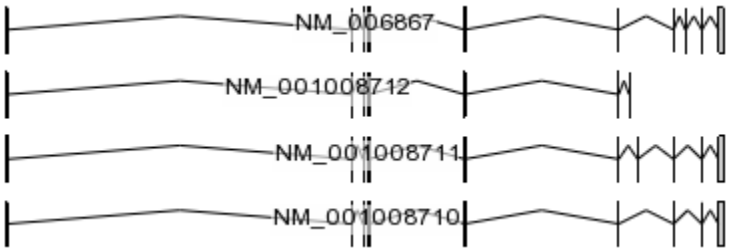


Figure 6. 199: Viewing known isoforms of gene location

The track height slider adjusts the relative height of the track (Figure 6. 200). Select the track from the drop down box of the track to change the height of that track. To increase the track height, move the slider bar to the left and select **Apply**. To decrease the track height, move the slider to the right and select **Apply**.

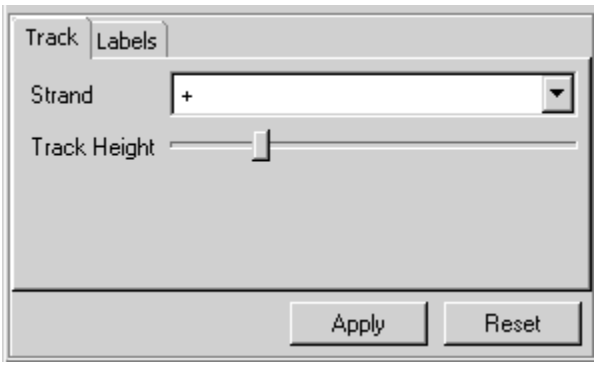


Figure 6. 200: Configuring the track height

The *Labels* tab configures the track title attributes and how to display gene or isoform labels (Figure 6. 201).

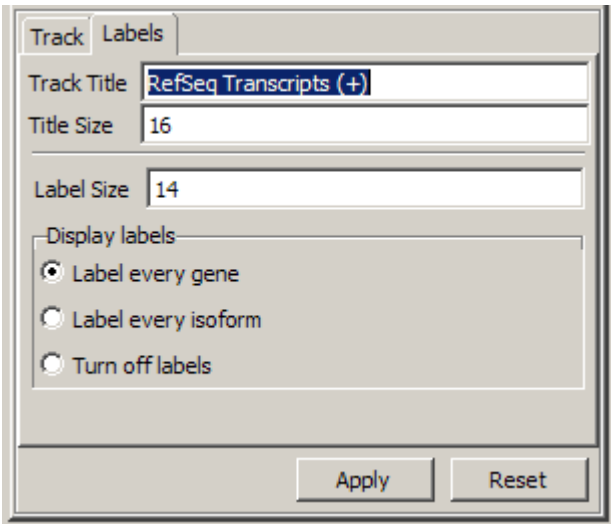


Figure 6. 201: Editing the Annotation Track label properties

The default title size and label sizes can be configured from **Edit > Preferences** from the main window.

2) *Add a track from spreadsheet* to create a track based on samples of the spreadsheet. When adding a track from a spreadsheet, the list of options is determined by the type of spreadsheet. If there is only type of track appropriate for the spreadsheet, then selecting *Next* will add the track. Figure 6. 202, Figure 6. 203, and Figure 6. 204 give examples of track options available from the different spreadsheet types.

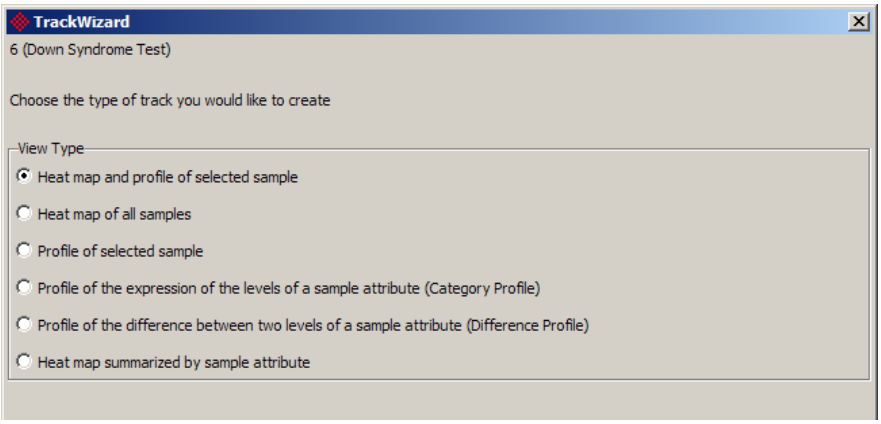


Figure 6. 202: Adding a New Track using the option "Add a track from spreadsheet" with a Sample spreadsheet selected from dropdown menu. Descriptions of these tracks are mentioned below

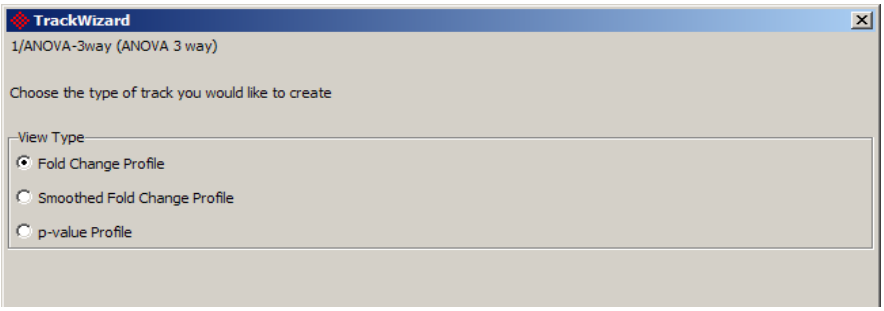


Figure 6. 203: Adding a New Track using the option "Add a track from spreadsheet" with an ANOVA results spreadsheet selected from dropdown menu

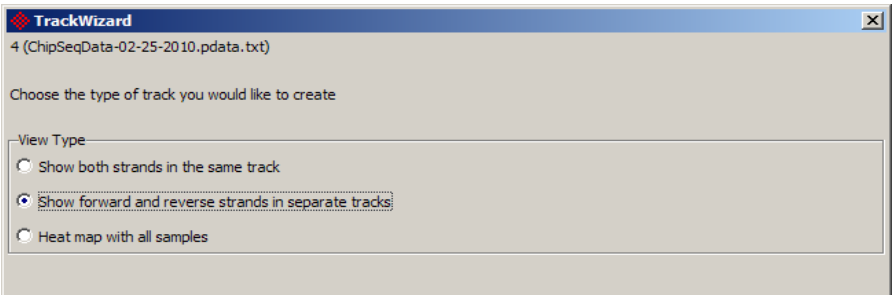


Figure 6. 204: Adding a New Track using the option "Add a track from spreadsheet" with a ChipSeq reads spreadsheet selected from dropdown menu

3) Add tracks from a list of samples to create a profile track by selecting samples (Figure 6. 205). Profiles are grouped by samples from every spreadsheet. Choose the samples by individually checking them and selecting the *Create* button, or choose them by a sample attribute and level attribute and select the *Check* button to select all the samples with the specified attribute. Profile tracks are generally most appropriate for Copy Number visualization, but have a broader purpose. Please see the Copy Number tracks section for more information on Profile tracks.

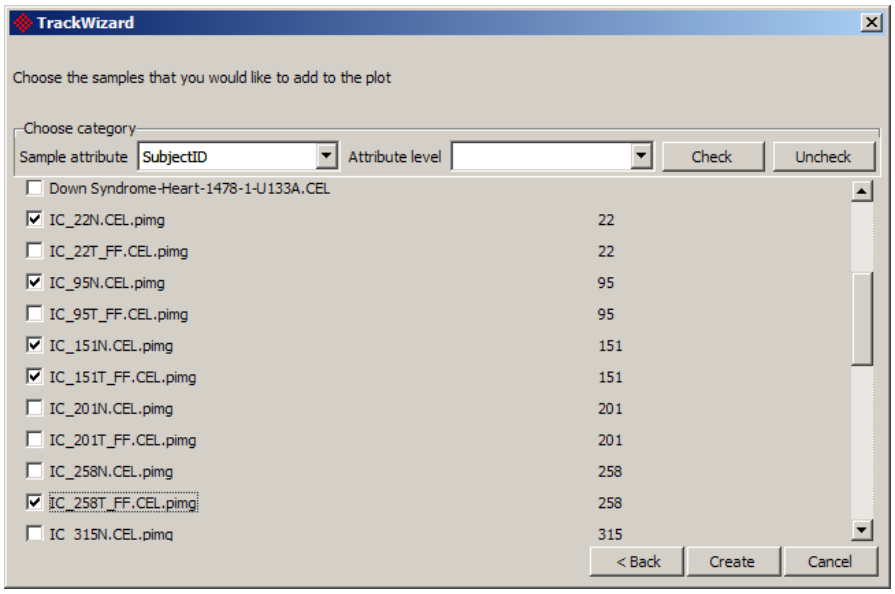


Figure 6. 205: Adding tracks from a list of samples

4) Add tracks from a list of spreadsheets to create a track from the spreadsheet list (Figure 6. 206). The option adds one track of the default type from the selected spreadsheets. The spreadsheet list contains all the spreadsheets currently open

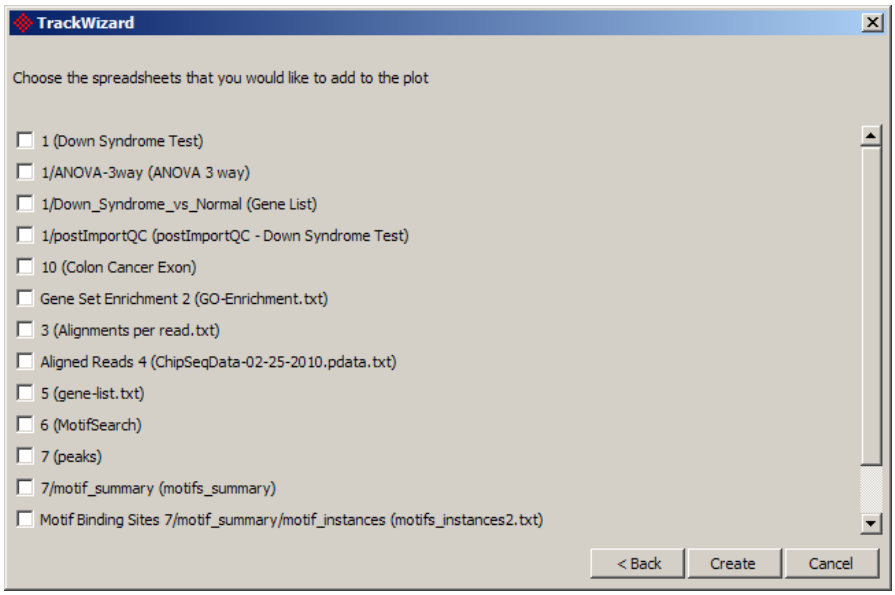


Figure 6. 206: Adding a default track from the list of spreadsheets

5) *Add a track with the sequence of the reference genome* to create a track with a known genome such as the human genome 18 (hg18). If one is not available, you will be asked to download or create a .2bit file from fasta files (Figure 6. 207).

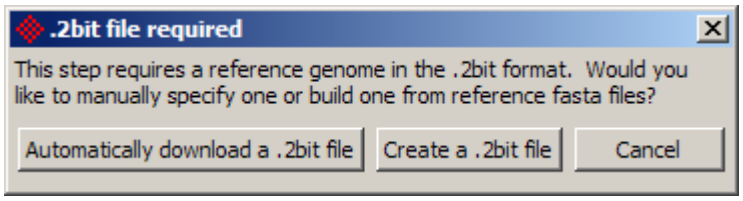


Figure 6. 207: Viewing the .2bit file dialog, Partek Genomics Suite attempts to download a .2bit file for the reference genome track

Reference Genome Track

The *Reference Genome* track displays the individual base pairs of the imported reference genome (Figure 6. 208). To import this track into the view, select *New Track* and then *Add a track with the sequence of the reference genome*. The base pairs will not be visible until the view is zoomed in.

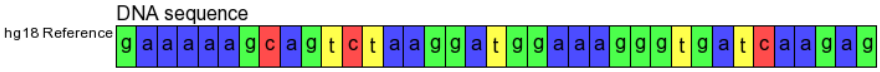


Figure 6. 208: Viewing the base pairs displayed from the hg18 Reference Genome track

The *Track Height* slider can be used to adjust the height of the *Reference Genome* track (Figure 6. 209). Moving the slider to the right will increase the track height, moving the slider to the left will decrease the track height.

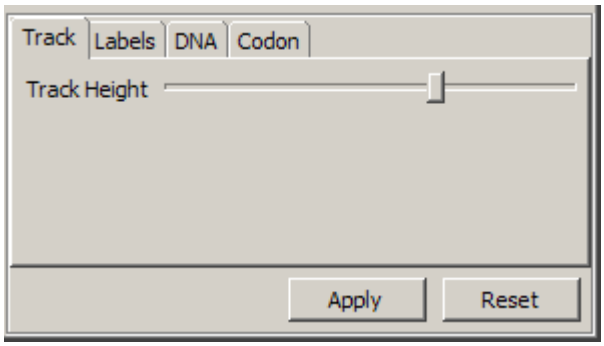


Figure 6. 209: Adjusting the Reference Genome track height

The Reference Genome *Track Label, Label Font Size and Base Font Size* can be adjusted from the *Labels* tab (Figure 6. 210).

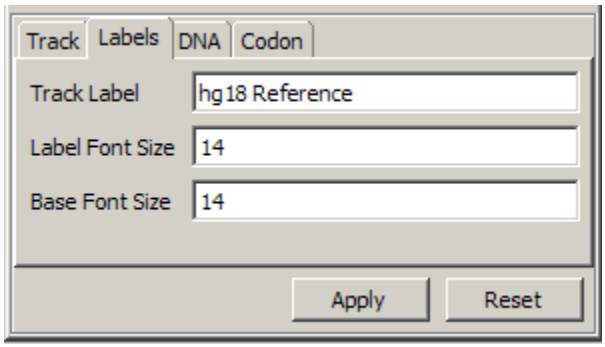


Figure 6. 210: Editing the label properties of the Reference Genome track

Uncheck *Show Bases* from the *DNA* tab to hide the reference sequence. Use this option if you want to only display the codons of the reference sequence (Figure 6. 211). Changing the color here will change base colors on the *Color* tab for the Base Colors Legend. Codons can be displayed to determine if a given mutation results in a change in protein. Select *Configure base colors* to change the colors of the bases from the default colors.

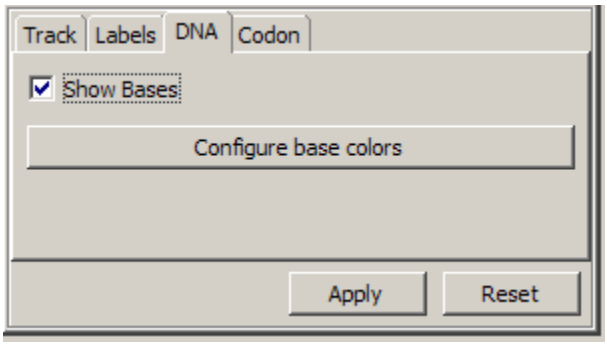


Figure 6. 211: Showing bases & configuring base colors of the Reference Genome track

The *Codon* tab is used to configure how the codons are displayed (Figure 6. 212).

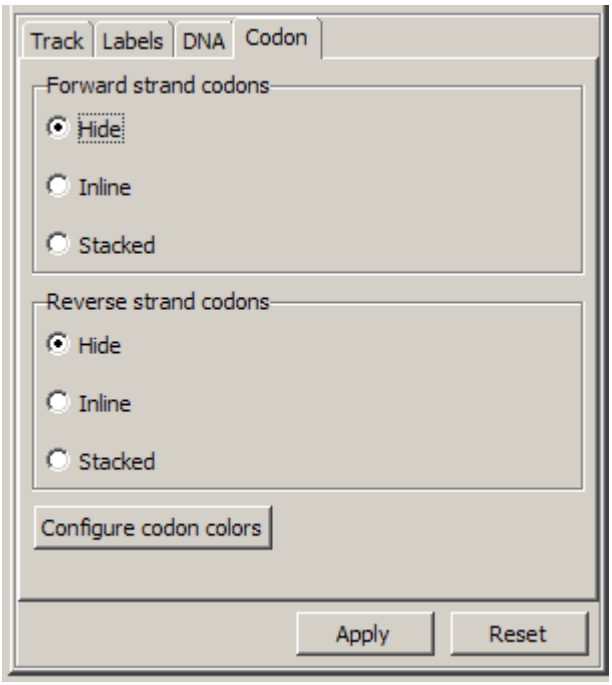


Figure 6. 212: Changing the codon display options of the Reference Genome track

The *Stacked* codon view displays three rows, each row revealing the potential starting base of the codon (Figure 6. 213). Each codon will be drawn over the three (3) bases it covers.

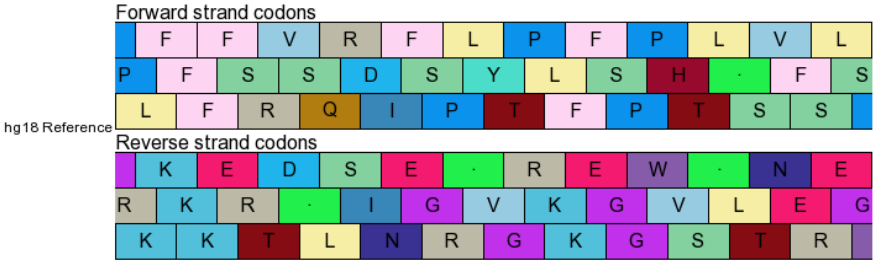


Figure 6. 213: Viewing a Stacked codon view, which is showing forward and reverse strand codons

The *Inline* codon view displays the codon to be transcribed if that base is the first base of three in the codon (Figure 6. 214).

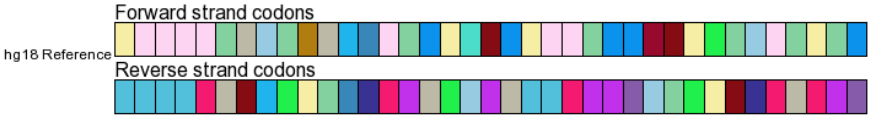


Figure 6. 214: Viewing the Inline codon view, which is showing forward and reverse strand codons

Selecting the *Configure codon colors* (Figure 6. 214) will open a color palette used to choose a color for each codon (Figure 6. 215).

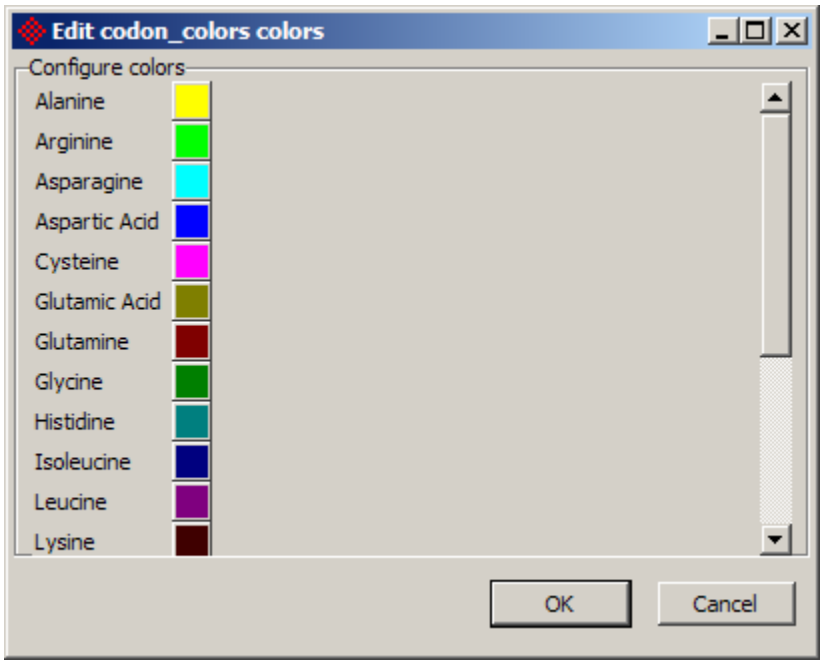


Figure 6. 215: Viewing the Color palette to change the codon colors

The *Base Color Legend* track displays the color palette being used for the *Reference Genome* track (Figure 6. 216). It can be changed by selecting the *Configure colors* button on the *Color* tab (Figure 6. 217).

Figure 6. 216: Viewing the Legend for base colors of the Reference Genome tracks

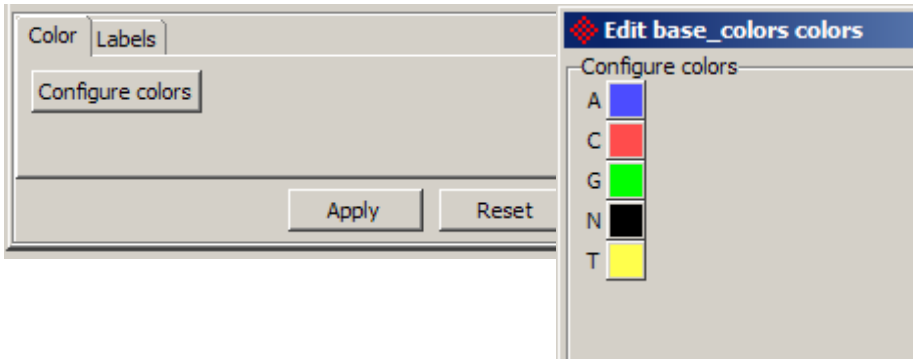


Figure 6. 217: Configuring base colors for the Reference Genome and Legend

The *Labels* tab can be used to configure the legend *Track Title* for the base pair colors (Figure 6. 218).

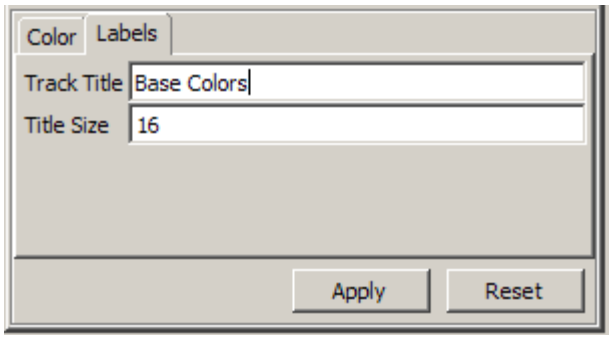


Figure 6. 218: Configuring the label for base colors

6) *Add a track with cytobands* to create a cytoband track. If one is not currently available, Partek GS will attempt to automatically download one.

Cytoband Track

The cytoband displays the chromosomal bands for the current view (Figure 6. 219). The chromosome number is displayed on the left side of the cytoband.

Figure 6. 219: Viewing a Cytoband with a chromosome label

The *Style* tab controls the brightness and labeling of the Cytoband track (Figure 6. 219). Check *Label Cytobands* to display the chromosomal band description below the cytoband (Figure 6. 220).

Figure 6. 220: Viewing a Cytoband with chromosomal band descriptions

The *Center Brightness* slider will adjust the brightness of the Cytoband (Figure 6. 221). Moving the slider to the left will decrease the brightness. Moving the slider to the right will increase the brightness.

Figure 6. 221: Viewing the Cytoband track properties with Center Brightness to the right

Genomic Label Track

The *Genomic Label* track displays the relative base scale of the current view (Figure 6. 222). It is typically loaded along with the Cytoband track.

Figure 1: Viewing the Genomic Label Track

The *Text Size* and *Number of tick marks* can be adjusted from the *Axis* tab of the *Genomic Label* track (Figure 38)

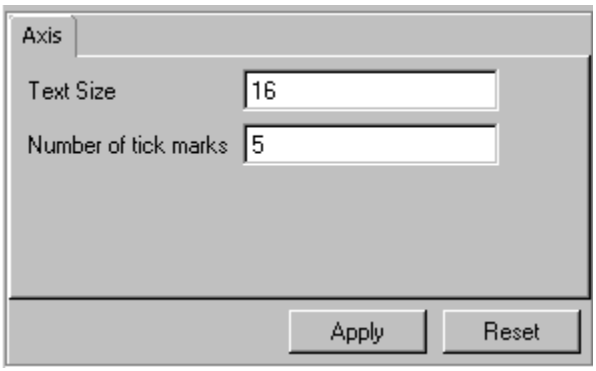


Figure 6. 222: Changing the text size and tick mark frequency of the Genomic Label track

7) *Other(Advanced)* to create a custom track type with specific data. Types of tracks include *Plot Title, Heat Map, Sequence Heat Map, Heat Map summarized by sample attribute, Profile, Annotation, Spreadsheet with genomic regions, Region bar profile, Regions separated by a categorical, Profile split by sample attribute (Category Profile), Color Map,* and *Profile of the difference between two levels of a sample attribute (Difference Profile)* (Figure 6. 223).

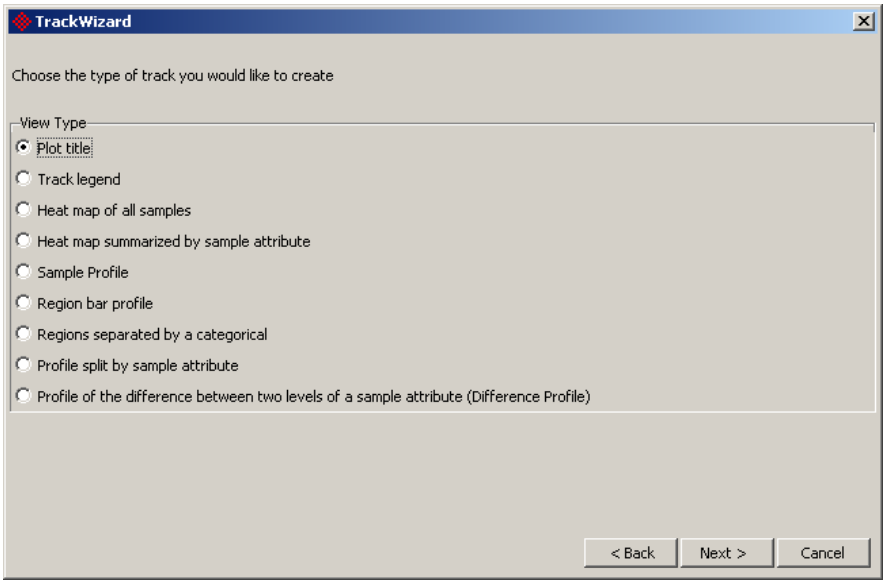


Figure 6. 223: Viewing available track types from the Other(Advanced) option of the New Track Wizard

Plot Title Track

Plot Title option will add a *Title* and *Title Size* to the *Chromosome View* (Figure 6. 224). Select the *Create* button to add the track.

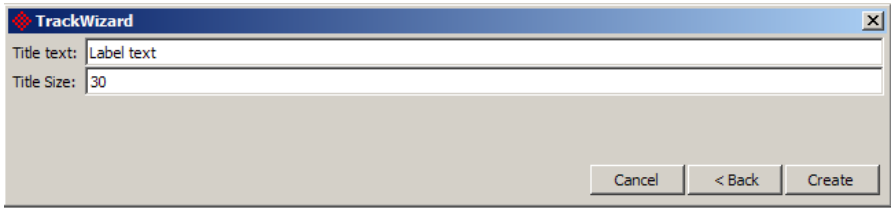


Figure 6. 224: Editing the Title Text and Title Size of the Plot Title track

Even after the *Plot Title* track has been added to the *Chromosome View*, it can still be edited. Figure 6. 225 shows how the labels of the *Title* and *Title Size* can be edited using the *Labels* tab.

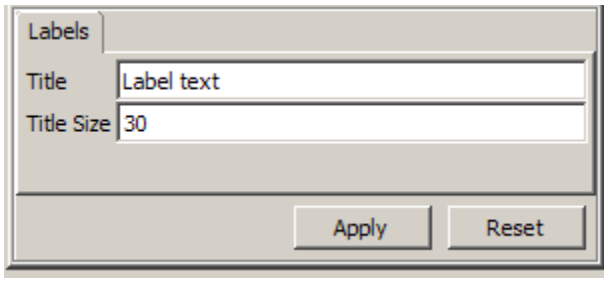


Figure 6. 225: Editing the Title and Title Size properties of the Plot Title track

Heat Map Track

Heat Map option will add a heat map based on the list of available spreadsheet samples. Selecting *Plot Chromosome View* from the workflow or choosing

View>Chromosome View with the sample spreadsheet selected will produce a *Heat Map* (Figure 6. 226).

A *Heat Map* track can be displayed on any spreadsheet that has genomic features (i.e. probesets) on columns. Individual heat maps can be drawn for each sample or for all samples within a selected spreadsheet. The heat map will attempt to display one marker per pixel. If the data is too dense to display and there is more than one marker mapped to a pixel, the color displayed will be the mean of the markers in that pixel.

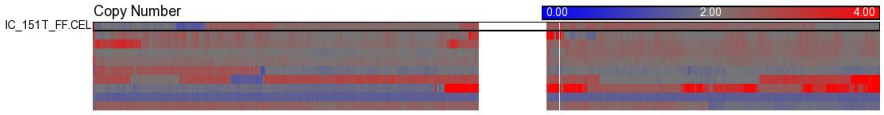


Figure 6. 226: Viewing the Heat Map track; each row corresponds to one sample

Use the *Data* tab to configure the heat map to display all of the samples in the spreadsheet or specify samples to display based on sample attribute (Figure 6. 227). Check the *Smooth Data* checkbox to turn on/off smoothing. Copy Number (and Log2ratio) data is smoothed by default, whereas other data is not. See Chapter 10 of the Partek Documentation for more information on Gaussian smoothing.

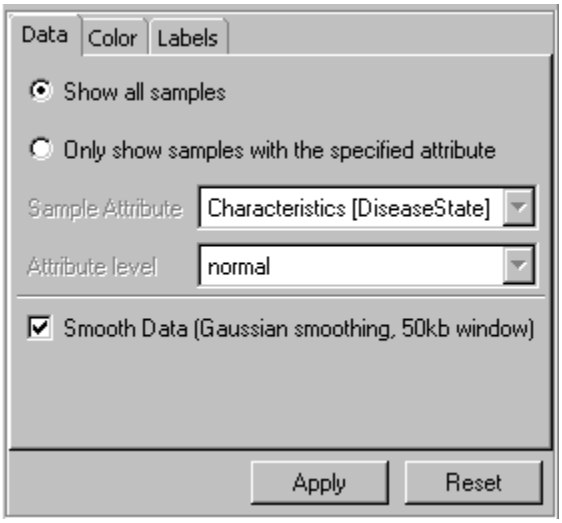


Figure 6. 227: Displaying sample options of the Heat Map track

The *Color* tab can be used to configure the color display of the heat map (Figure 6. 228). The *Min/Max* input options control lets you add the scale for how expression values are displayed corresponding to a given color. The *Min/Max* values are automatically determined by the range of the data.

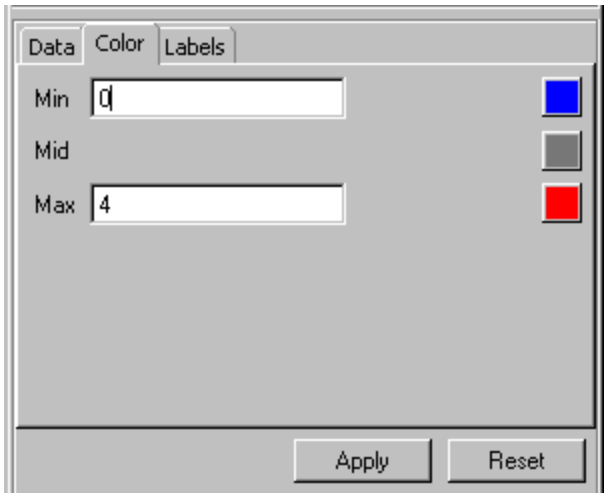


Figure 6. 228: Configuring the color options of Heat Map track

Use the *Labels* tab to adjust the *Track Title*, *Track Size*, or *Label Size*, and to turn on/off sample labels (Figure 6. 229).

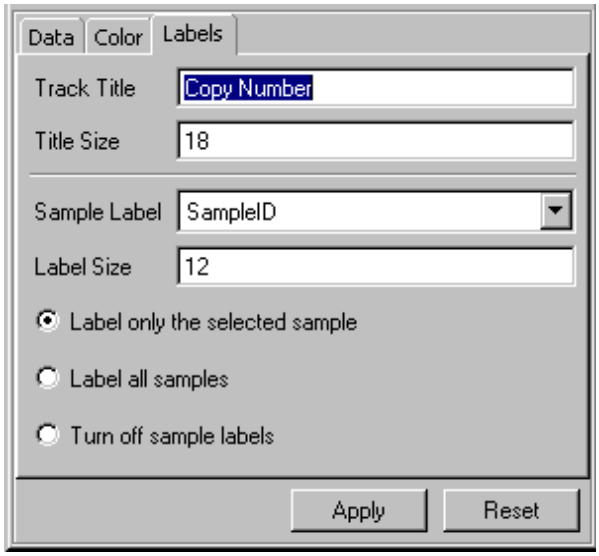


Figure 6. 229: Editing the label properties of heat map

Add Sequence Heat Map* is discussed under **Chip-seq Tracks under **Workflow Specific Track Description**.

Heat Map Summarized by Sample Attribute Track

The *Heat Map summarized by sample attribute* option displays a heat map track with the expression values summarized across a specified sample attribute (Figure 6. 230).

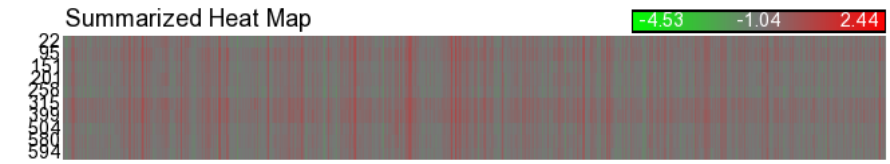


Figure 6. 230: Viewing the Heat Map summarized by sample attribute (Subject ID)

The *Data* tab (Figure 6. 231) lets you control how the *Heat Map* summarized by sample attribute track is to be displayed using the available options of the *Data* tab. The heat map can be changed to display a sample attribute and whether or not smoothing is turn on or off. For more on Gaussian smoothing, please see Chapter 10 of the Partek Documentation.

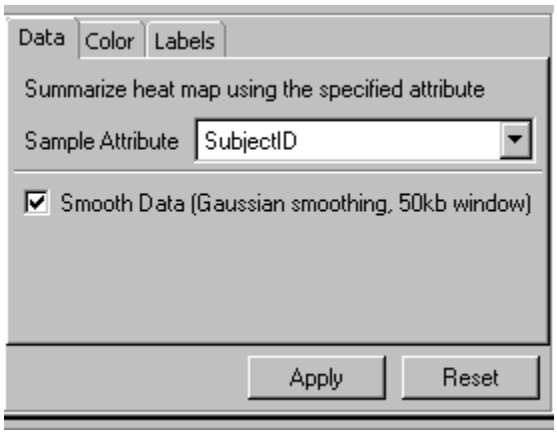


Figure 6. 231: Viewing the Data tab properties of the Heat Map summarized by sample attribute track

The color of the *Heat Map* summarized by the sample attribute can be adjusted using options on the *Color* tab (Figure 6. 232). The *Min/Max* expression intensity scale can be changed as well as the color range.

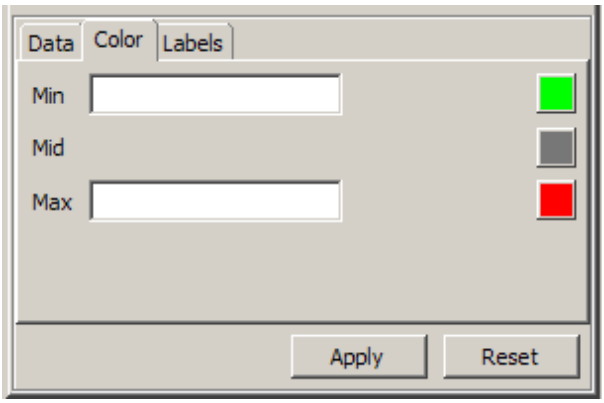


Figure 6. 232: Viewing the Color tab of the Heat Map summarized by sample attribute

The *Labels* tab is used to edit the values of the *Track Title*, *Title Size*, and *Label Size* of the *Heat Map* summarized by sample attribute (Figure 6. 233).

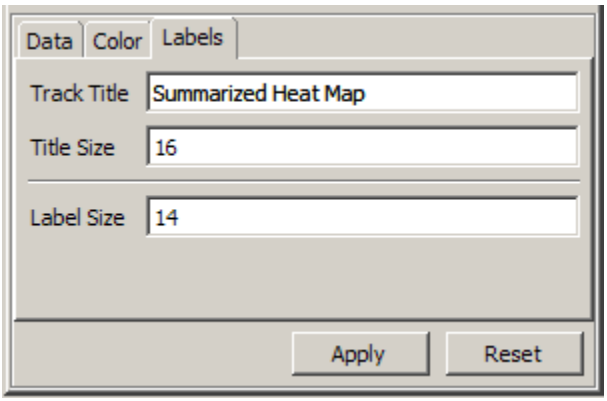


Figure 6. 233: Editing the Track Title, Title Size, and Label Size of the Heat Map summarized by sample attribute

Profile* track is discussed under **Copy Number tracks under **Workflow Specific Track Description.**

Profile Split by sample attribute* description can be found under the **Category Profile > **Gene/Exon Expression tracks** under **Workflow Specific Tracks**.

Legend Track

The *Legend* option allows you to add a descriptive color legend for most of the common tracks (Figure 6. 234). Choose the type of legend you would like to display, add a title to the legend using the *Display name of Legend* input box, and select *Create*.

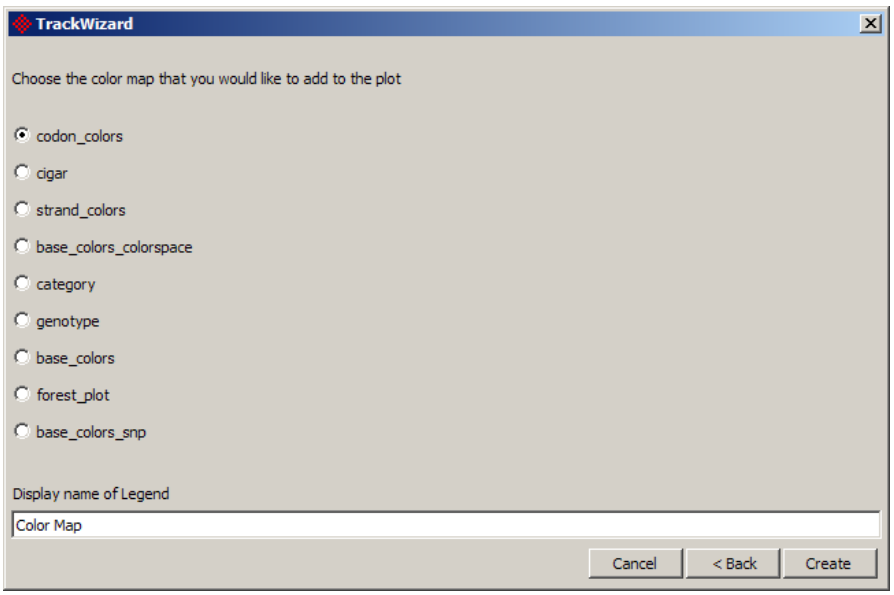


Figure 6. 234: Adding a Color Map track to the Chromosome View

The colors of the *Color Map* track can be changed using the *Color* tab (Figure 6. 235). Select the *Configure colors* button to adjust the way the properties of the *Color Map* are displayed.

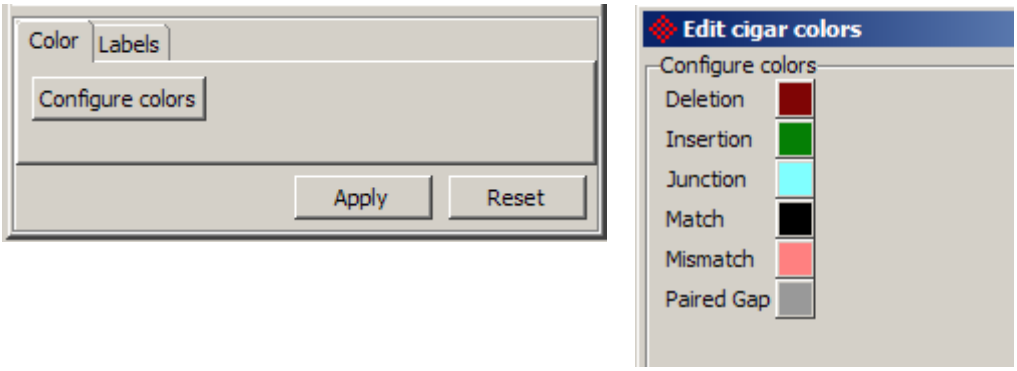


Figure 6. 235: Configuring the colors of the cigar Color Map

The *Labels* tab of the *Color Map* is used to edit the *Track Title* and *Title Size* of the *Color Map* (Figure 6. 236).

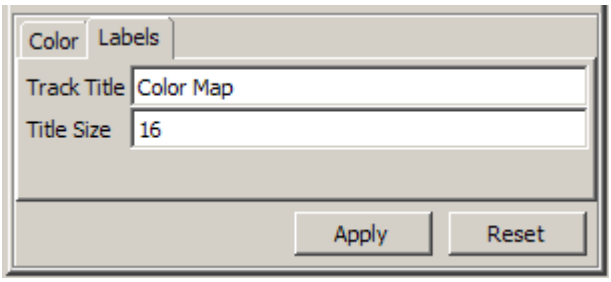


Figure 6. 236: Editing the Track Title and Title Size of the Color Map track

Profile of the difference between two levels of a sample attribute* description can be found under **Difference Profile > **Gene/Exon Expression Tracks** under **Workflow Specific Track Description.**

Workflow Specific Tracks

The Workflow Specific Tracks section explores tracks that are generally most appropriate for Copy Number, Gene/Exon Expression and Next Generation Sequencing workflows, although not exclusive to those workflows.

Copy Number Tracks

The *Chromosome View* can display copy number information in different ways including amplifications, deletions and individual intensity values for region or whole chromosome views, allele specific copy number, and Loss of Heterozygosity.

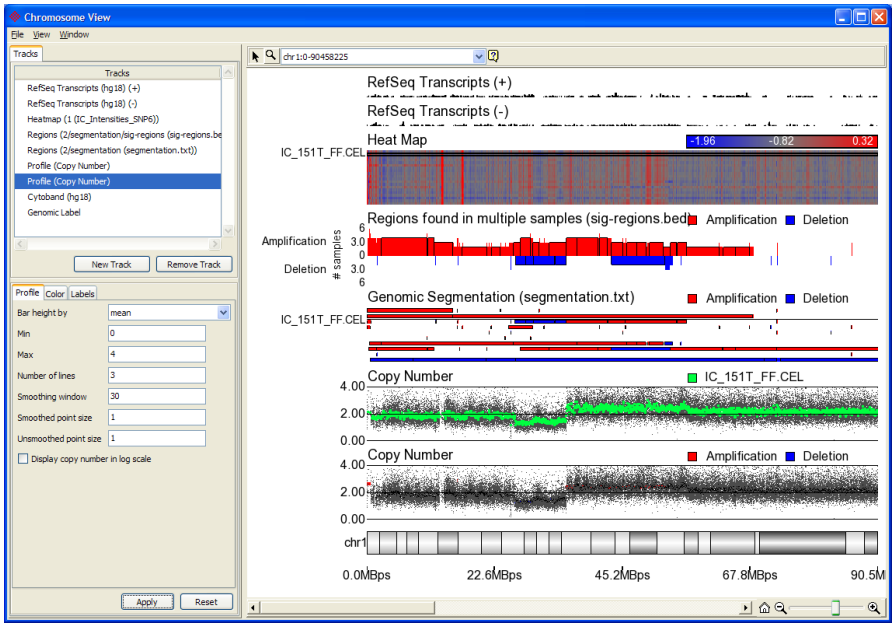


Figure 6. 237: Viewing the Chromosome View featuring Copy Number data

Sig-regions Histogram Track

The result of *Find Regions in Multiple Samples* from the Copy Number workflow will be displayed as a histogram track in the *Chromosome View*. Amplifications extend above the center; deletions extend below the center (Figure 6. 238).

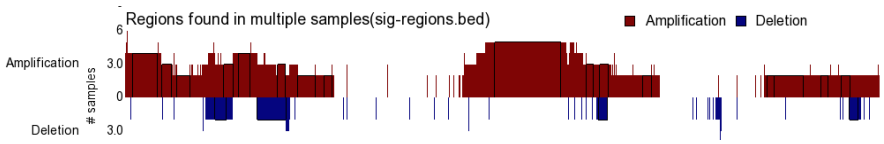


Figure 6. 238: Viewing the Histogram bar height by # Samples (Y-axis)

From the *Profile* tab, check *Separate bars by* to separate the bars by the available attributes in the sig-regions spreadsheet (Figure 6. 239). The histogram bar height is determined by the selected attribute in the *bar height by* drop down menu. The default selection is the *Copy Number* column attribute with the bar height as *# samples*. The *Min* and *Max* values set the Y-axis scale of the track. The *Bars come from* feature sets the baseline from which to extend the histogram height.

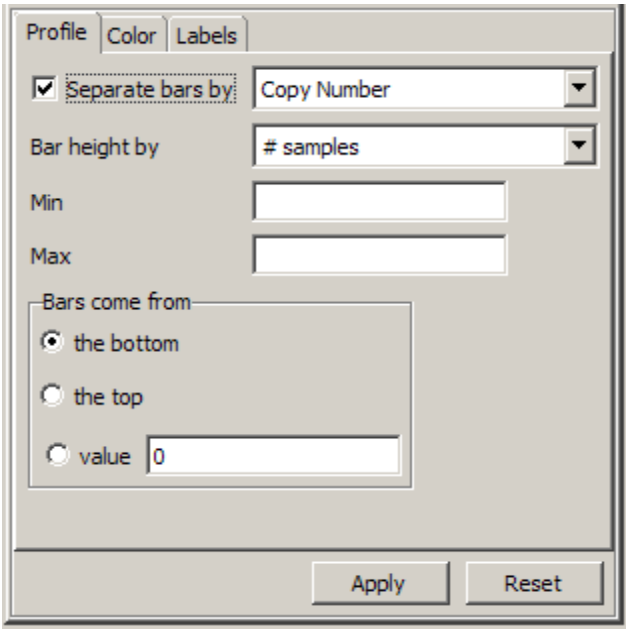


Figure 6. 239: Configuring the sig-regions histogram

Color bars by will determine which attribute is used to color the histogram bars. Select *Configure category colors* to change the colors of the histogram bars by attribute (Figure 6. 240). The default value is Copy Number with amplifications drawn in Red and deletions drawn in Blue. The *Min* and *Max* values set the color range intensity values but will only be noticeable if specific attributes are selected for the *Color bars by* drop down menu.

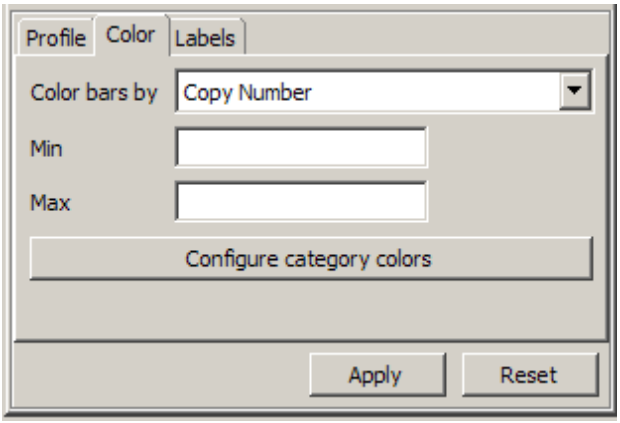


Figure 6. 240: Configuring the Color sig-regions histogram by column attribute

The *Track Title*, *Title Size*, and *Label Size* can be changed from the *Labels* tab (Figure 6. 241).

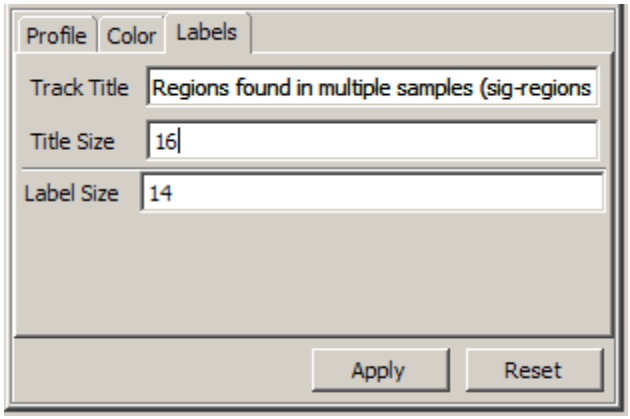


Figure 6. 241: Editing label properties of the sig-regions histogram

Segmentation Track

The *Segmentation* track displays regions of copy number variation. The results are bars plotted with Amplifications and Deletions lengths by sample (Figure 6. 242). By default, one row will be drawn for each sample.

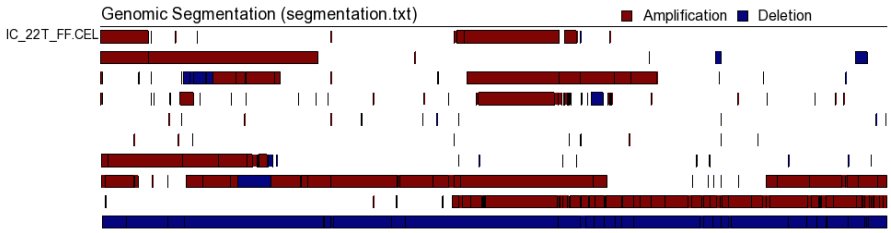


Figure 6. 242: Viewing the Segmentation results; Amplification=red, Deletion=blue

The *Segmentation* track can be displayed by attributes of the spreadsheet. Use the dropdown menu of the *Profile* tab to select bar separation (Figure 6. 243). The *Track Height* can be adjusted using the track height slider. Moving the slider to the right will increase the track height, moving the slider to the left will decrease the track height.

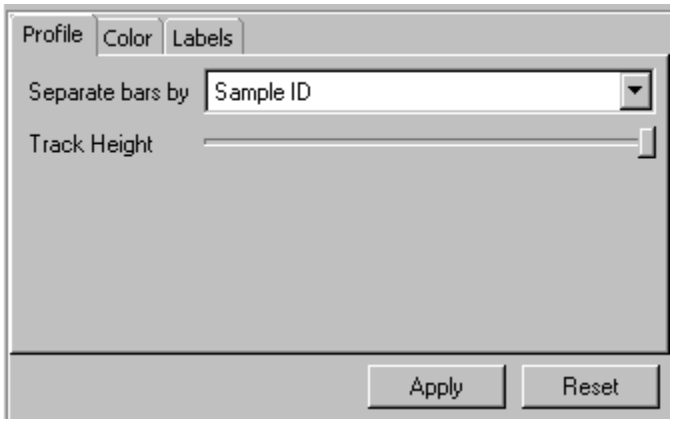


Figure 6. 243: Configuring the Segmentation track display

The *Color* tab gives selections to color the bars. Select **Configure category colors** to change the colors of the bars by attribute (Figure 6. 244). The default value is Copy Number with amplifications drawn in Red and deletions drawn in Blue. The *Min* and *Max* values set the color range intensity values but will only be noticeable if specific attributes are selected for the *Color bars by* drop down menu.

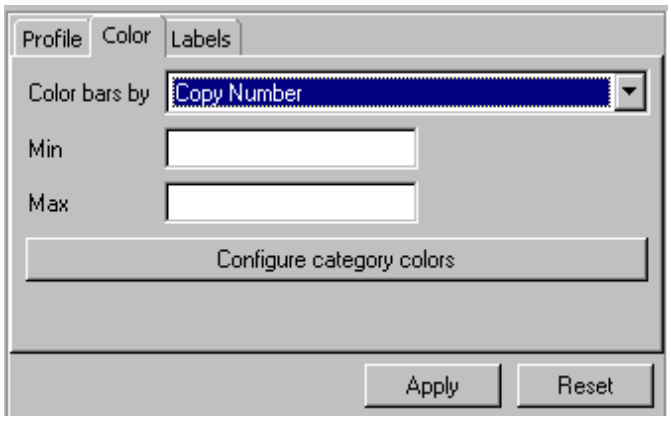


Figure 6. 244: Coloring the histogram by attribute of the Segmentation track

The *Track Title*, *Title Size* and *Font Size* can be changed from the *Labels* tab. By default, the first sample in the spreadsheet is selected (Figure 6. 245). Deselect *Label on the selected sample* to label all the samples.

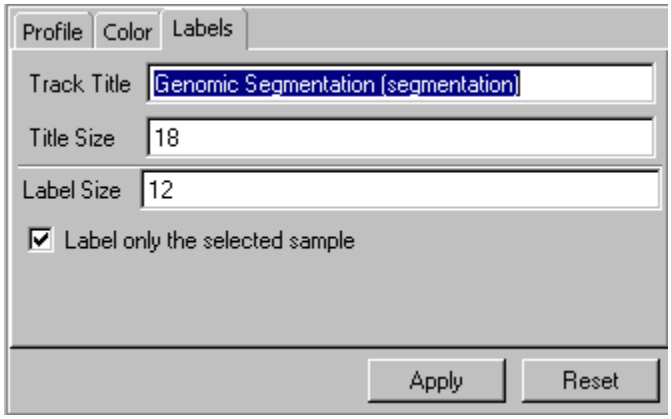


Figure 6. 245: Editing the Track Title and Label Size of the Segmentation track

Profile Track

The *Profile* track displays the expression of individual markers in smoothed and unsmoothed form. The position of the smoothed points is based on the median of the points within a *Smoothing window*. A profile track can be created for each sample, or samples can be displayed overlapping each other. Figure 6. 246 shows a copy number profile for one (1) sample.

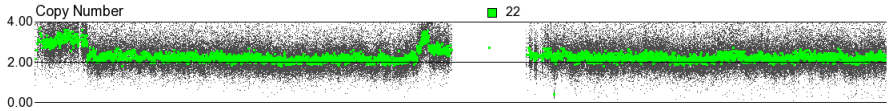


Figure 6. 246: Viewing the Profile track of smoothed copy number sample

Change the *Y-axis* tab configuration options to set the *Min/Max* values of the Y-axis scale, *Number of grid lines*, *Smoothing window* size, *Smoothed point size*, *Unsmoothed point size*, or *Display the copy number profile in log scale* (Figure 6. 247).

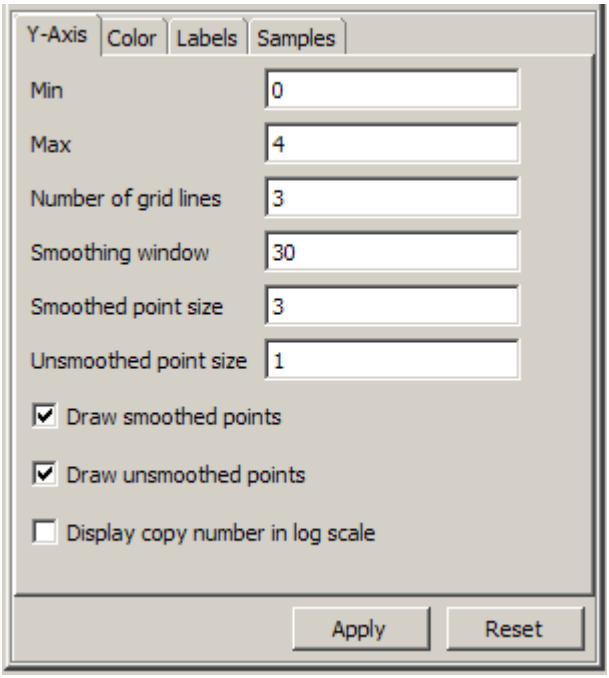


Figure 6. 247: Configuring the Profile track plot options

The *Color* tab can be configured to change the how the colors of the smoothed & unsmoothed points are displayed (Figure 6. 248). Select *Configure category colors* to change how the points are colored.

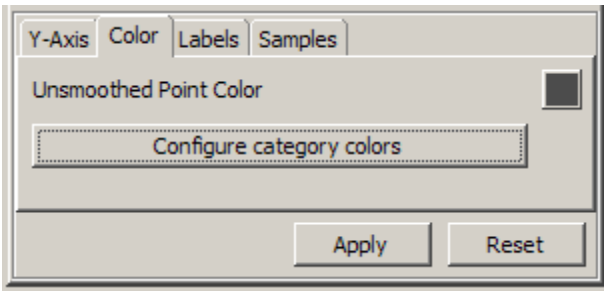


Figure 6. 248: Configuring the color properties of Profile track

The *Labels* tab allows you to edit the *Track Title* and *Title Size* (Figure 6. 249).

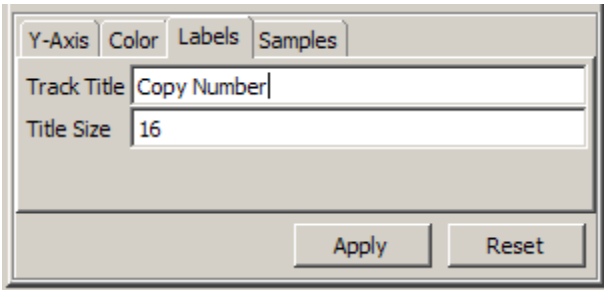


Figure 6. 249: Editing Track Title, track Title Size of the Profile track

The *Samples* tab lets you choose which samples are to be displayed on the selected *Profile* track (Figure 6. 250). Samples can be displayed individually or can overlap with each other. Selecting the *Set Samples* button will prompt the same dialog as Figure 6. 205.

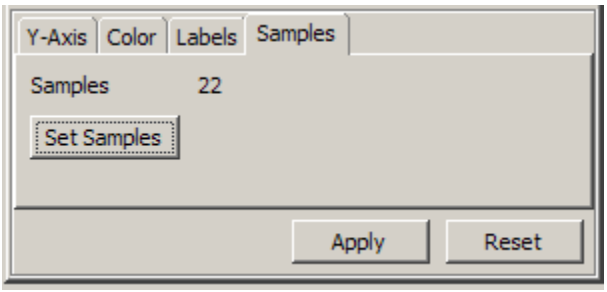


Figure 6. 250: Viewing the Samples tab of the Profile track

The *Profile* track displays the heat map sample selected by the viewer. Figure 6. 251 below shows a highlighted sample on the heat map with a corresponding copy number *Profile* track updated with the sample profile. Other sample profiles can be added (below) but will not be updated with the selection of samples on the heat map.

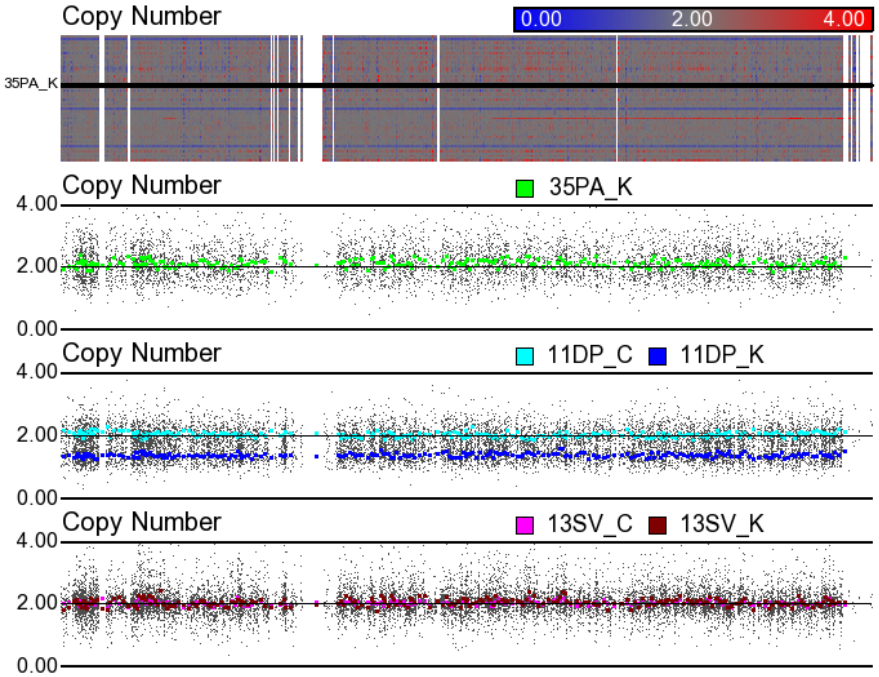


Figure 6. 251: Viewing the heat map that has sample 35PA_K selected. The middle track has manually selected the normal and tumor samples for subject 11DP (the bottom track for subject 13SV). These two tracks will not change as samples are selected in the heat map

Gene/Exon Expression Tracks

Gene Expression and *Exon* tracks can be added in the *Chromosome View* to visualize up and down regulation of genes, alternative splice events, expression values by categorical attributes, difference profiles between categorical attributes, fold change, p-values, and more (Figure 6. 252).

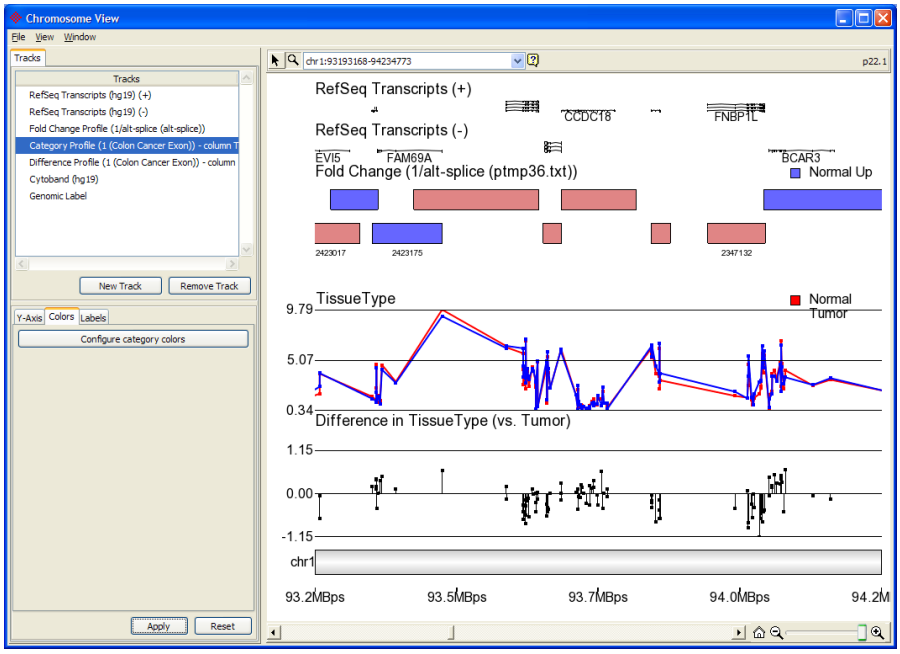


Figure 6. 252: Viewing the Gene/Exon Expression tacks in the Chromosome View

Category Profile Track

A *Category Profile* can be created to display the average expression values across samples at a given probeset to look for possible up/down regulation of genes and alternative splice events. The category profile in Figure 6. 253 shows average expression values of Normal vs. Tumor Tissue.

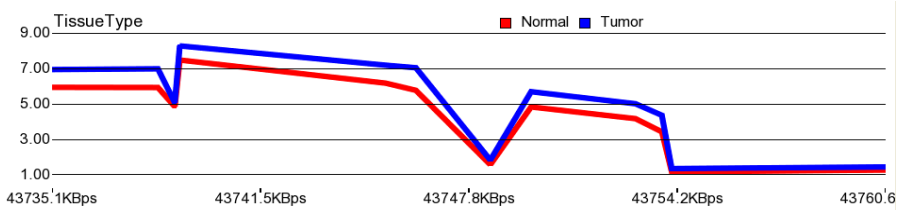


Figure 6. 253: Viewing the Category Profile of Normal vs. Tumor samples

The *Y-Axis* tab is used to configure the track plotting properties (Figure 6. 254). The *Column* drop down menu allows you to choose the column to display the average value of the samples. The *Min* & *Max* variables set the scale of the track. The position of points is determined by the average value of samples in the same level of a categorical variable. If Min and Max are blank then the y-axis range is automatically set to the range of points within the view. The range of the plot can be manually specified by entering Min and Max values. The grid line increment values will be determined by the difference in the Max & Min divided by the number of grid lines.

The *Smoothing window* option will change the way the probe sets for the selected sample at each location is displayed. This option specifies a window of probe sets to smooth. For every group of probe sets there will be one point drawn based on the

median of the probe sets in the window. There is no overlap between windows. The *Smoothed point size* and *Unsmoothed point size* determine how do display the individual probes. By default for better visualization, the smoothed probes will have a larger point size than the unsmoothed point size. The width of the lines can be increased or decreased using the *Line Width* slider.

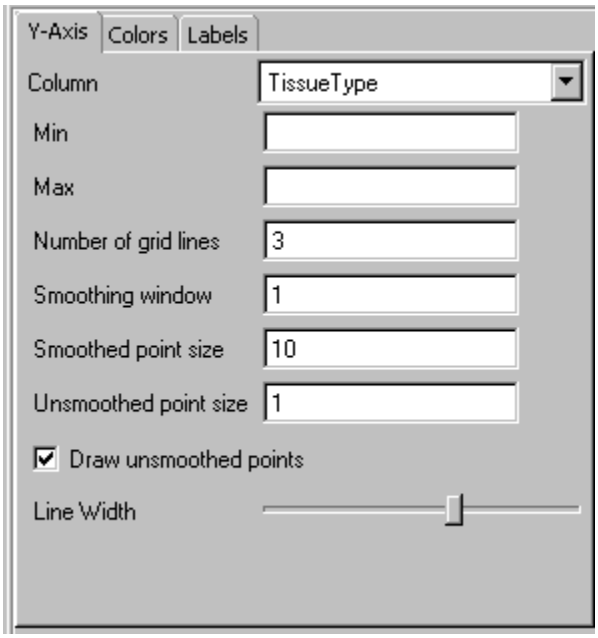


Figure 6. 254: Editing the Y-axis tab of the Category Profile

The *Colors* tab (Figure 6. 255) can be used to configure the color of the *Column* values in Figure 6. 254.

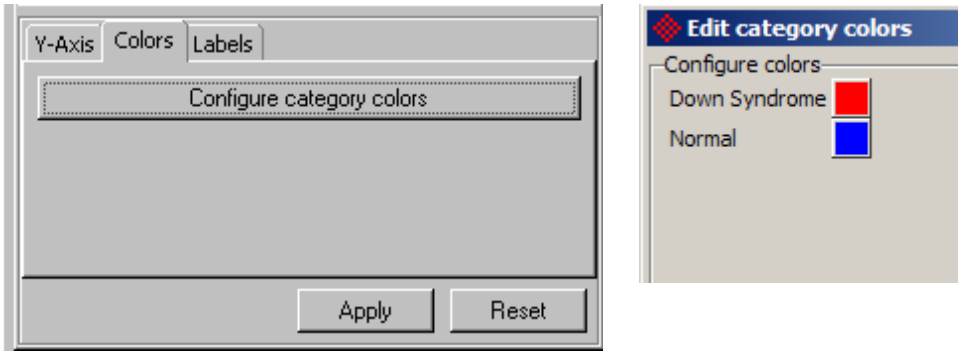


Figure 6. 255: Configuring the category colors from Colors tab

The *Labels* tab lets you edit the *Track Title* and *Title Size* (Figure 6. 256).

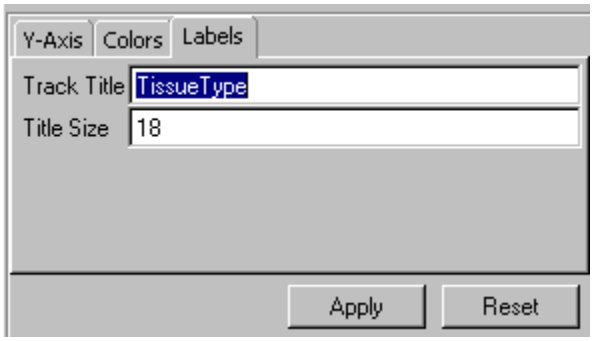


Figure 6. 256: Editing Track Title and Track Title size of the Category Profile

Difference Profile Track

The *Difference Profile* track displays the difference of the average expression values between a selected attribute category. Figure 6. 257 shows the *Difference Profile* between Normal and Tumor average expression values. It can be added by selecting *New Track > Other(Advanced) > Profile of the difference between two levels of a sample attribute(Difference Profile).*

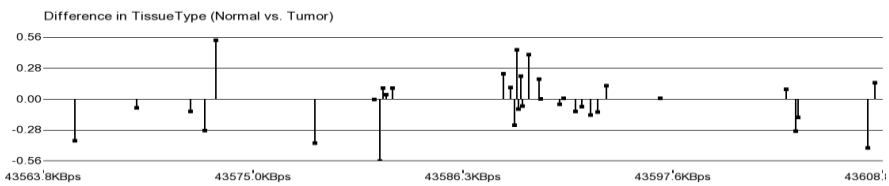


Figure 6. 257: Viewing the Difference Profile between Normal vs. Tumor

Configure the plot properties of the *Difference Profile* using the *Y-Axis* tab (Figure 6. 258). Check the *Split by* checkbox to view differences grouped by levels of a categorical variable. The length is determined by the average value (*Baseline level*) samples subtracted from the average value of the other samples (non-*Baseline level*). The *Compare attribute* drop down menu gives options for which categorical variables to display. The *Baseline level* is the level that comes from the categorical variable specified in *Compare attribute*. The *Min* & *Max* variables set the scale of the track. If Min and Max are blank then the y-axis range is automatically set to the range of points within the view. The grid line increment values will be determined by:

$$\frac{(Max - Min)}{Number\ of\ grid\ lines}$$

The *Smoothing window* option will change the way the probe sets for the selected sample at each location are displayed. This option specifies a window of probe sets to smooth. For every group of probe sets there will be one point drawn based on the <u>median</u> of the probe sets in the window. There is no overlap between windows.

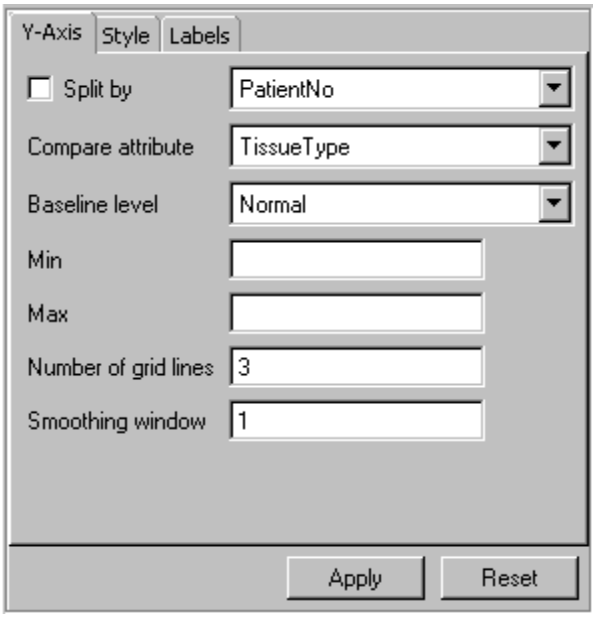


Figure 6. 258: Editing the Y-axis tab of the Difference Profile

Use the *Split by* option to view differences grouped by levels of a categorical variable, determining which *Compare attribute*, and which of the attributes to be chosen as the *Baseline level* (Figure 6. 259).

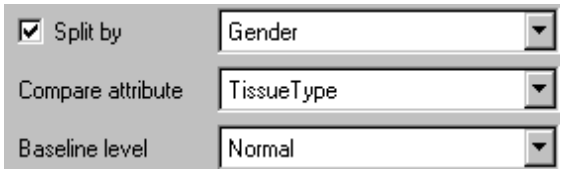


Figure 6. 259: Configuring the Split by option for the Difference Profile track

Figure 6. 260 shows a *Difference* profile *Split by* Gender, comparing *Normal vs. Tumor TissueType*.

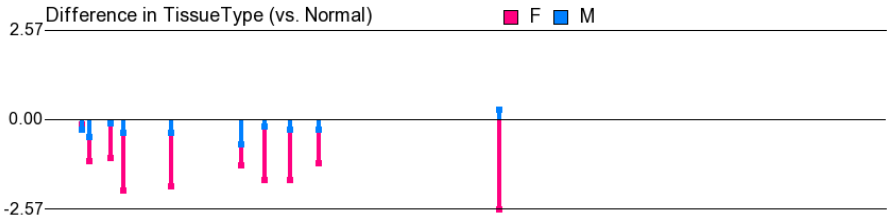


Figure 6. 260: Viewing the Difference profile Split by Gender comparing Normal vs Tumor TissueType. One line for [Avg(Female Tumor) – Avg(Female Normal)]-pink and one line for [Avg(Male Tumor) – Avg(Male Normal)]-blue

Use the *Style* tab to change the way the points are plotted (Figure 6. 261). By default, *Draw line to zero* is checked. Uncheck to view just the points. Use the *Point Size* slider to increase or decrease the size of the points. Check *Connect points*

to draw a line connecting every difference point. Use the *Line Width* slider to increase or decrease the width of the lines.

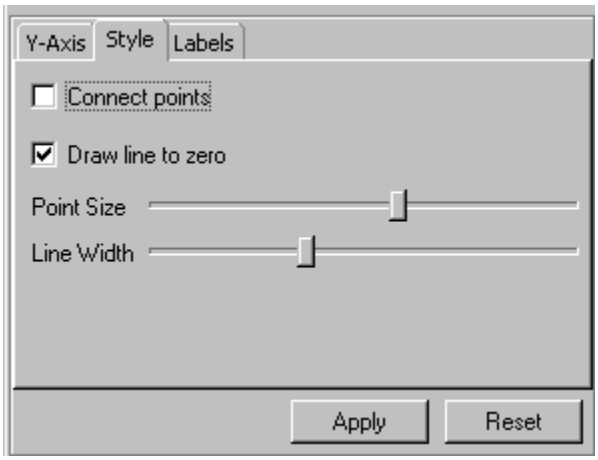


Figure 6. 261: Adjusting the Style tab of the Difference profile

The *Labels* tab lets you edit the *Track Title* and *Title Size* (Figure 6. 262).

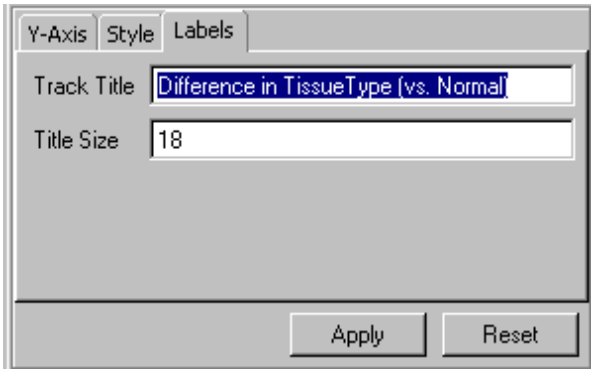


Figure 6. 262: Editing the Track Title and Track Title size of the Difference Profile

Fold Change Profile Track

The *Fold Change Profile* track displays gene or exon regions of up/down & significantly up/down differential expression as displayed by fold change value (Figure 6. 263).

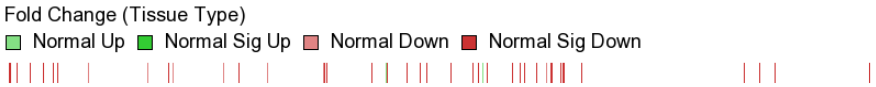


Figure 6. 263: Viewing the Fold change profile track of Normal sample up/down differential expression

Select the *Profile* tab to select the *Factor of interest* to display (Figure 6. 264). All columns with a “FoldChange (Factor)” label format will be listed as available options. Use the *Track Height* slider to increase or decrease the height of the track in the view.

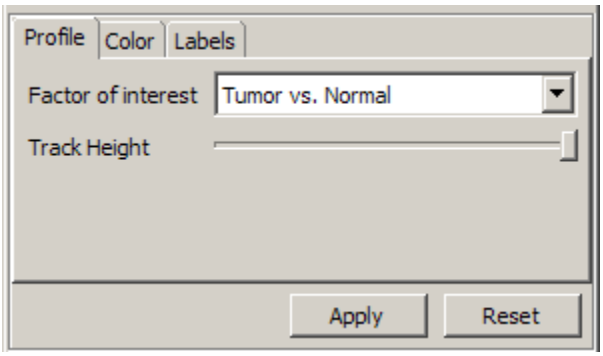


Figure 6. 264: Changing the Fold Change Profile display Factor and Track Height

Select the *Color* tab to choose the color for the positive and negative fold change values. The track can be configured to draw markers a darker color that pass a certain threshold (Figure 6. 265).

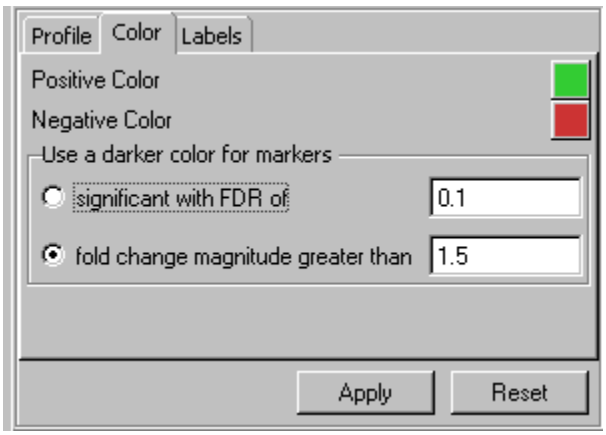


Figure 6. 265: Changing the default colors and by threshold

Edit the *Track Title* and *Title Size* properties in the *Labels* tab (Figure 6. 266).

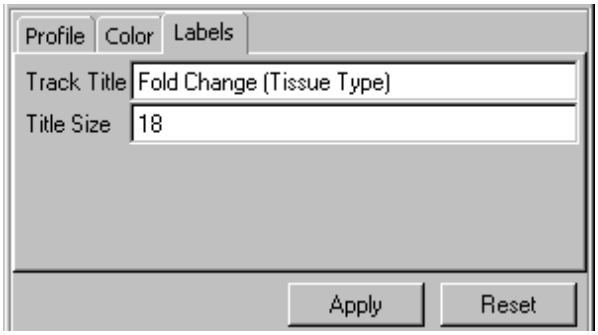


Figure 6. 266: Editing the Track Title label and Title Size of the Fold Change Profile

Smoothed Fold Change Track

The *Smoothed Fold Change* track provides visualization for viewing the distribution of nearest markers with positive or negative fold change in each direction (Figure 6. 267).

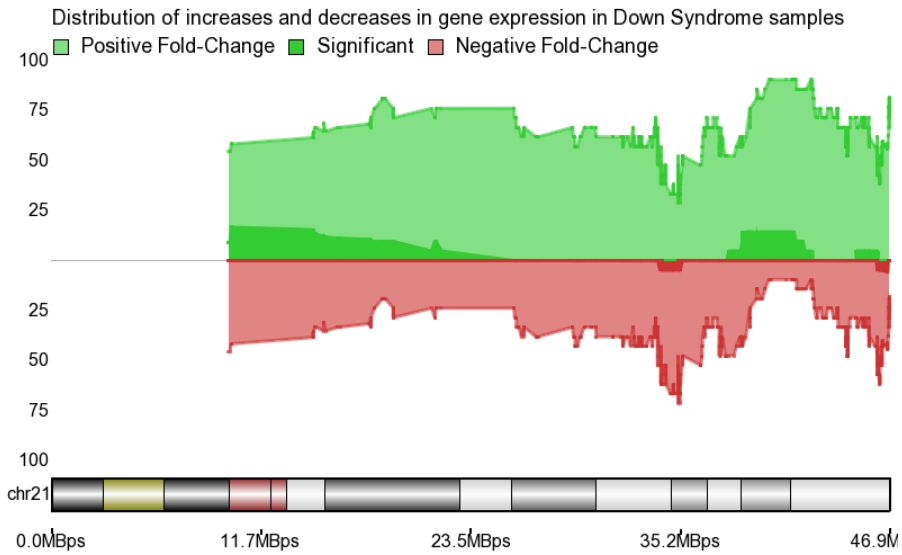


Figure 6.267: Viewing the Smoothed Fold Change plot of Down Syndrome samples

Use the *Y-Axis* tab to configure the display of the *Smooth Fold Change* track (Figure 6. 268). All columns with a "FoldChange (Factor)" format will be listed as available options. The input box for the *Threshold* will determine the cutoff of significant fold change. The *Min* & *Max* will set the scale for plot of distribution values. By default, the *Min* & *Max* values are set to -100 and 100. The *Nearest markers/Base Pairs* values can be chosen to extend or shorten the nearest marker or base pair length. The default setting is the 10 nearest markers in each direction.

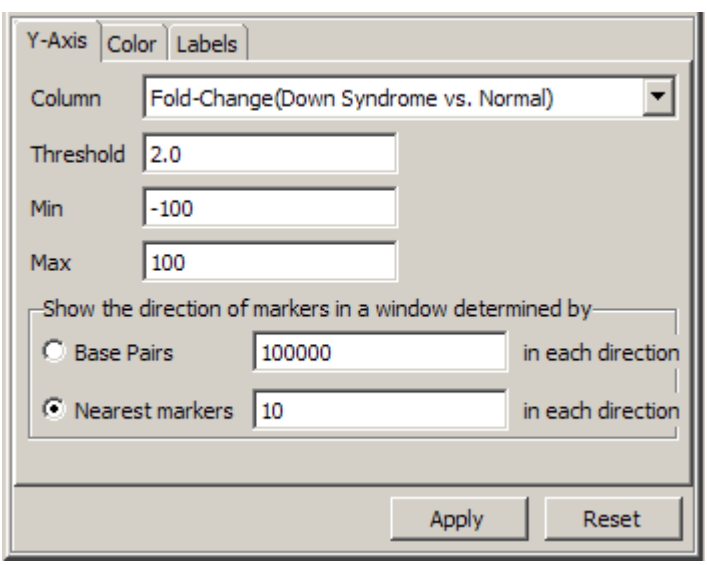


Figure 6. 268: Adjusting plot properties of the Smoothed Fold Change track

Figure 6. 269 shows the *Color* tab for changing the colors used to display the *Positive* and *Negative* fold change distributions.

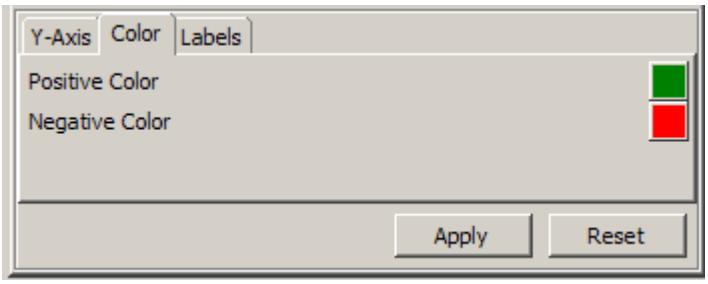


Figure 6. 269: Changing the default colors for Positive/Negative fold change

Edit the *Track Title* and *Title Size* properties in the *Labels* tab (Figure 6. 270).

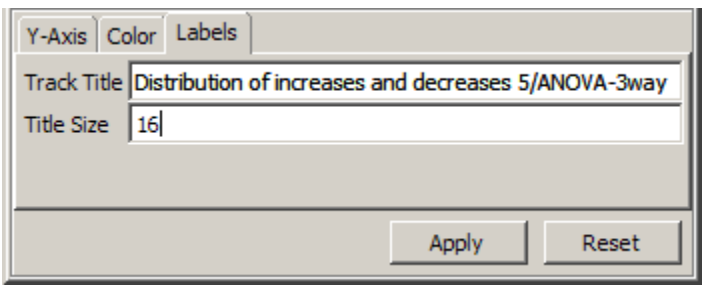


Figure 6. 270: Editing the Track Title and Track Title Size of Smoothed Fold Change track

The smooth value at each marker is displayed as a percentage and calculated by:

$$\left[\frac{g}{2m+1}\right] * 100$$

where *g* is the number of nearest markers with a) positive fold change or b) negative fold change; *m* is the number of nearest markers in each direction (default 10); 1 includes the current marker; *100 to turn into percentage.

The significant positive or negative fold change is calculated by dividing the number of significantly a) positive or significantly b) negative markers by the total number of nearest markers (including current marker), as such:

$$\left[\frac{s}{2m+1}\right] * 100$$

where *s* is the number of nearest markers with <u>significantly</u> a) positive fold change or significantly b) negative fold change; *m* is the number of nearest markers in each direction (default 10); 1 includes the current marker; *100 to turn into percentage.

The significant marker frequency is determined by the *Threshold* on the *Y-Axis* tab and is indicated by middle darker section of *Smoothed Fold Change* track.

p-value Profile Track

The *p-value Profile* track displays the p-values at markers for a specified category (Figure 6. 271).

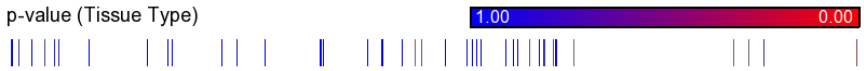


Figure 6. 271: Viewing the p-value Profile track by TissueType

All columns with a "p-value(Factor)" label format will be listed as available options for the *Factor of interest*. Adjust the *Track Height* using the track slider (Figure 6. 272).

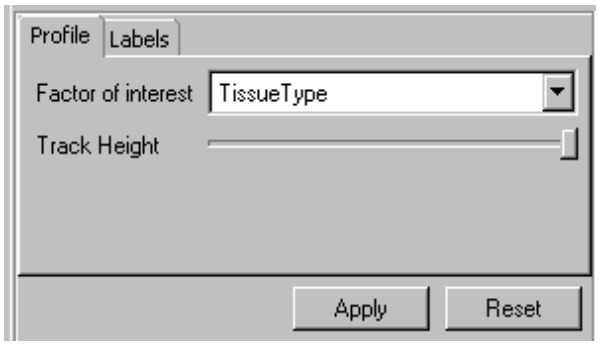


Figure 6. 272: Adjusting the Profile properties of the p-value Profile

Edit the *Track Title* and *Title Size* properties in the *Labels* tab (Figure 6. 273).

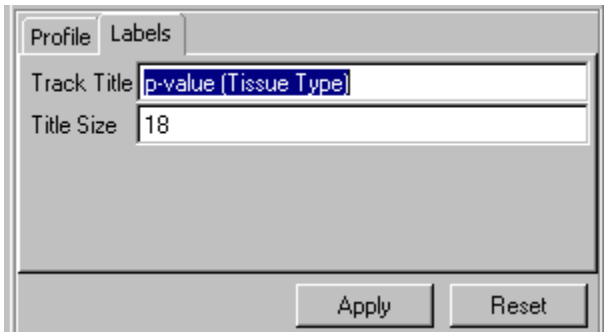


Figure 6. 273: Editing Track Title and Track Title Size of P-value Profile

Correlation Profile Track

The *Correlation Profile* track provides visualization for viewing the distribution of nearest markers with positive or negative correlation in each direction (Figure 6. 274).

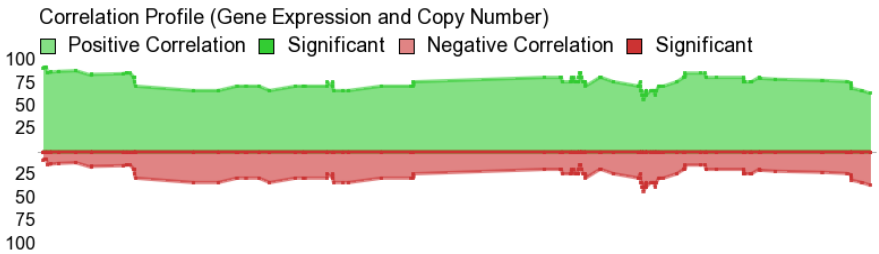


Figure 6. 274: Viewing the Correlation Profile track

The result of the *Correlate Copy Number with Gene Expression* workflow feature plots the *Correlation Profile* track. This plot has similar window parameters as the *Smoothed Fold Change* plot, but instead of fold change, it displays positive and negative correlation. The *Y-Axis* tab is used to change the plotting parameters (Figure 6. 275). The *Column* options permit the display of the linear or rank correlation. The *Threshold* value sets the cutoff for significant positive or negative correlation. The *Min* & *Max* will set the scale for plot of distribution values. By default, the *Min* & *Max* values are set to -100 and 100. The *Nearest markers/Base Pairs* values can be chosen to extend or shorten the nearest marker or base pair length. The default setting is the 10 nearest markers in each direction.

See the *Smoothed Fold Change* track section for description of how distribution plot values are calculated.

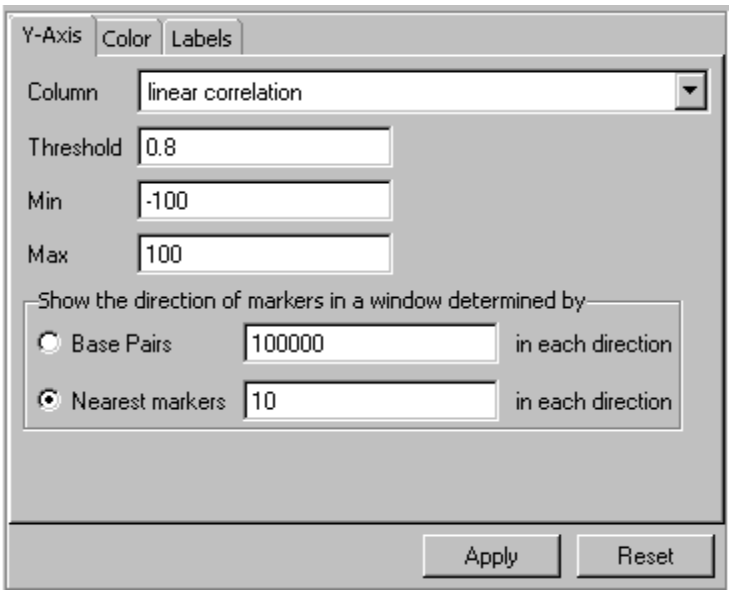


Figure 6. 275: Adjusting plot properties of the Correlation Profile track

The size of the dark section in the middle of the plot is determined by the percentage of correlation values that pass the *Threshold* parameter (Figure 6. 276).

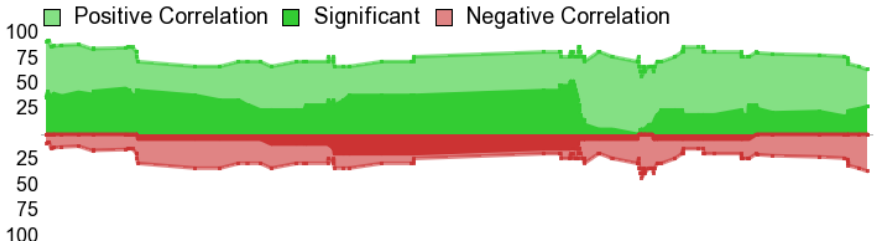


Figure 6. 276: Viewing the Correlation Profile track showing up/down and significant up/down correlation

Figure 6. 277 shows the *Color* tab for changing the colors used to display the *Positive* and *Negative* fold change distributions.

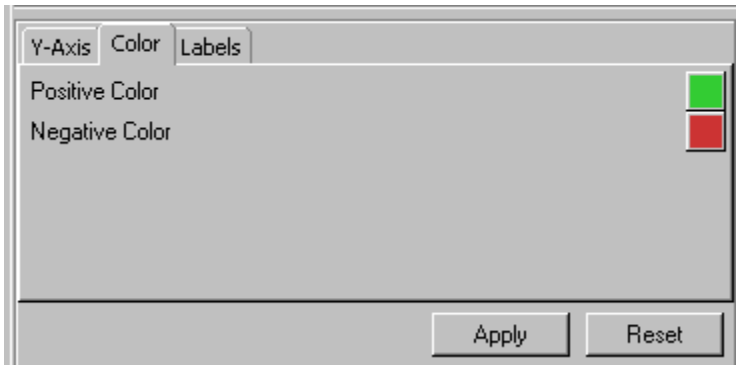


Figure 6. 277: Changing the default colors for Positive/Negative values of the Correlation Profile track

Edit the *Track Title* and *Title Size* properties in the *Labels* tab (Figure 6. 278).

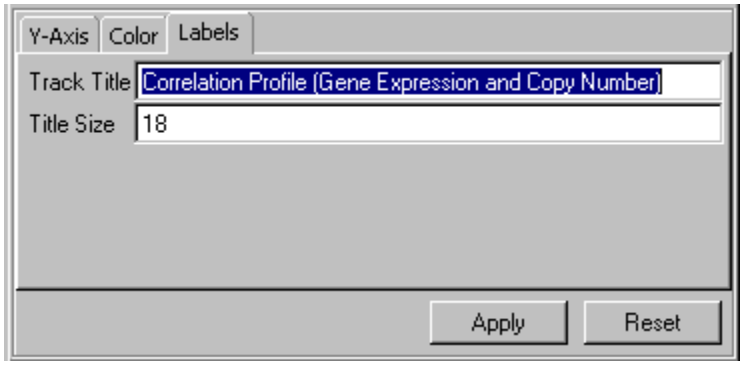


Figure 6. 278: Editing the Track Title and track Title Size of Correlation Profile track

Next Generation Sequencing Tracks

RNA-Seq Tracks

RNA-Seq tracks can be added in the *Chromosome View* to visualize mapped read counts along with gene annotation information, cytobands, SNP proportions to find

base pair changes, and isoform proportions locations to look for alternative splicing of genes across the transcriptome (Figure 6. 279).

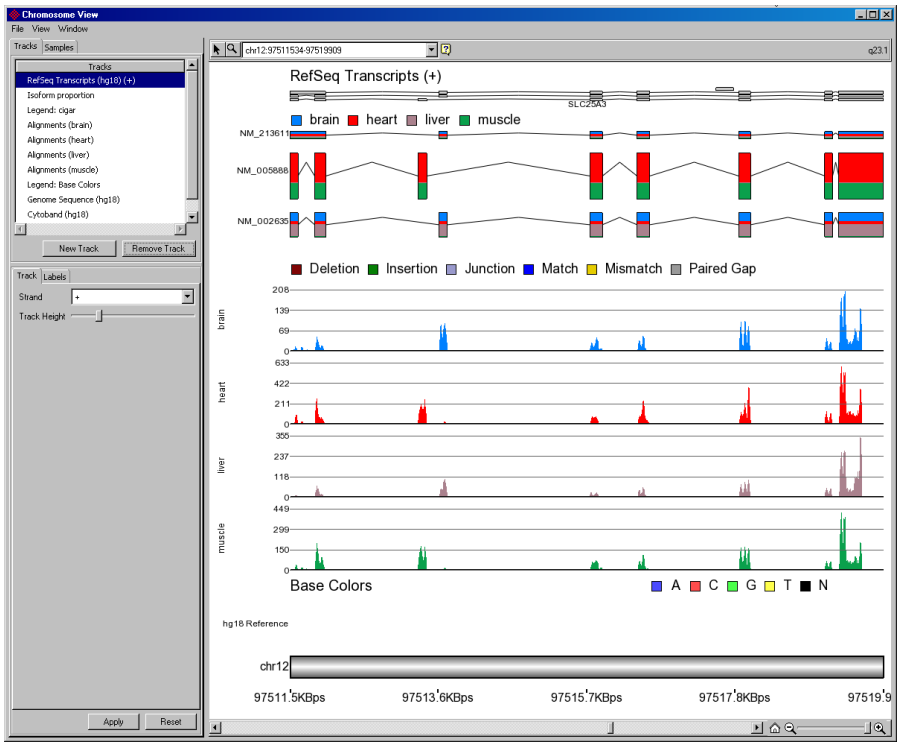


Figure 6. 279: Viewing the RNA-seq tracks in Chromosome View

Isoform Proportion Track

The *Isoform Proportion* track displays the mapped reads to transcripts and helps visualize differential expression and alternative splicing. The size of each transcript is proportional to the number of reads that map to the transcript. The color indicates the samples for which the reads belong. Figure 6. 280 shows heart and muscle primarily express in NM_005888. Brain and liver primarily express in NM_002635.

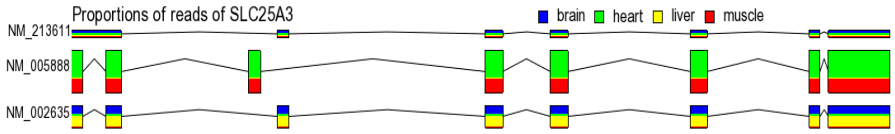


Figure 6. 280: Viewing the Isoform Proportion track showing reads mapped to transcripts

The gene symbol can be manually set on the *Profile* tab (Figure 6. 281).

Figure 6. 281: Manually setting the gene symbol on the Profile tab of the Isoform Proportion track

Configure the colors of the *Isoform Proportion* track in the *Color* tab (Figure 6. 282).

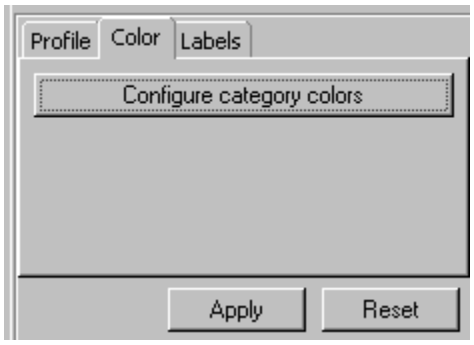


Figure 6. 282: Configuring the category colors from Colors tab

Edit the *Track Title*, *Track Title Size* and *Label Size* of the *Isoform Proportion* track using the *Labels* tab (Figure 6. 283).

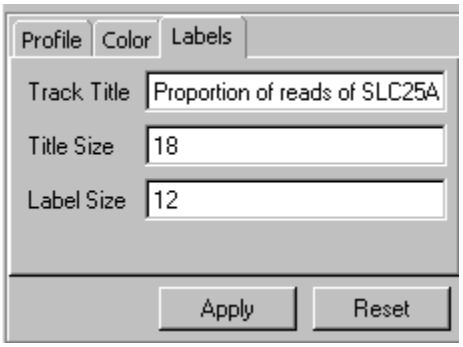


Figure 6. 283: Editing the Track Title, Track Title Size and Label Size of Isoform Proportion track

Alignment Track

The *Alignment* track displays a view of the number of alignments per read (Figure 6. 284). Each alignment track is displayed as an individual sample. By default, the *Histogram* view is displayed with the *alignments colored by sample*.

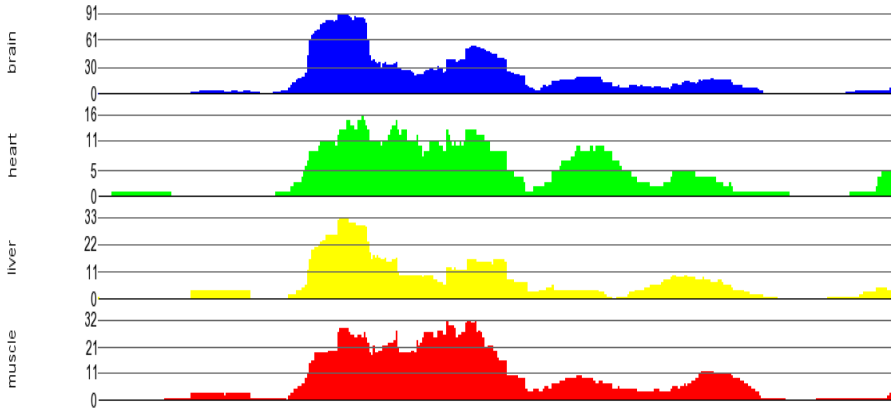


Figure 6. 284: Viewing the Alignment track of RNA-Seq data with Histogram view. The Histogram view is most appropriate for dense data

The *Cigar* track gives a legend of how matches are colored (Figure 6. 285). These include locations where Deletions, *Insertions, Junction, Matches, Mismatches* and *Paired Gaps* occur.

Deletion Insertion Junction Match Mismatch Paired Gap

Figure 6. 285: Viewing the Cigar track displayed by default with colors

The *Style* tab of the *Alignment* track controls the way the alignments are displayed (Figure 6. 286). The *Track Height* slider adjusts the height of the track. Moving the slider to the right will increase the track height, moving the slider to the left will decrease the track height. The reads can be displayed as *One per row*, *Fewest number of rows*, and *Histogram*.

The *Histogram* display can be adjusted to have a Maximum Y-Axis scale using the input box for the *Y-Axis Maximum.* Leave this blank to have the maximum automatically determined by the range of the data. From the *Style* tab, the color options for the alignment track can be changed.

The *Histogram* view is useful for viewing regions with the greatest number of reads.

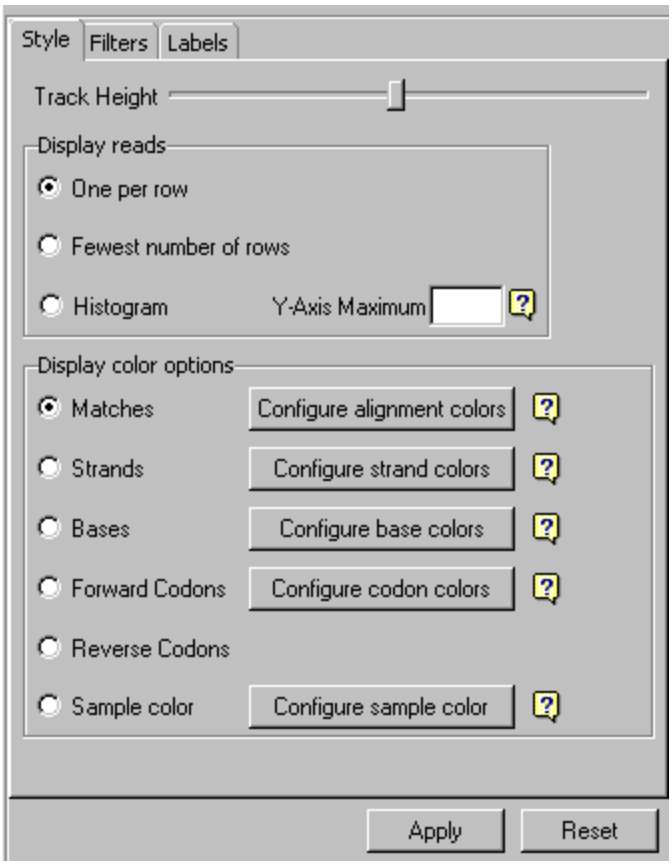


Figure 6. 286: Configuring the Style tab to configure Alignment tracks

Figure 6. 287 shows the *Alignment* track with the *Fewest number of rows* option.

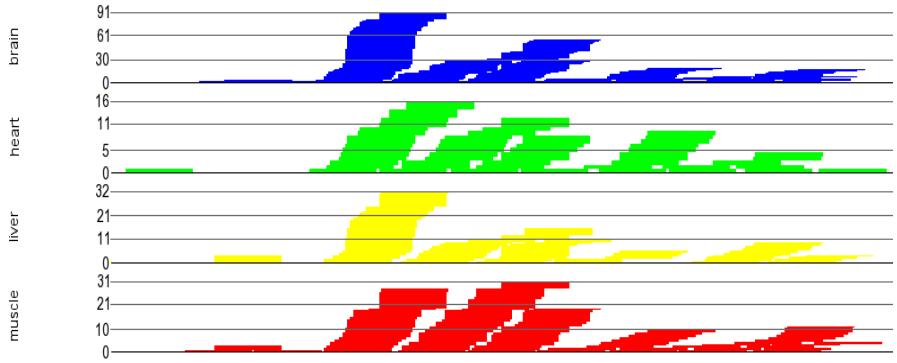


Figure 6. 287: Viewing the Alignment tracks displaying Fewest number of rows option

Figure 6. 288 shows the *Alignment* track with the reads at *One per row* option.

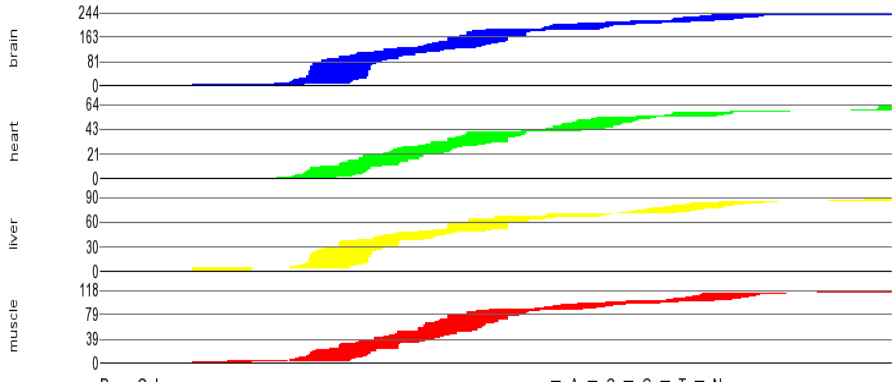


Figure 6. 288: Viewing the Alignment tracks displaying One per row option

The *Strands* color options will color the results according to the direction of either the forward or reverse read (Figure 6. 289).

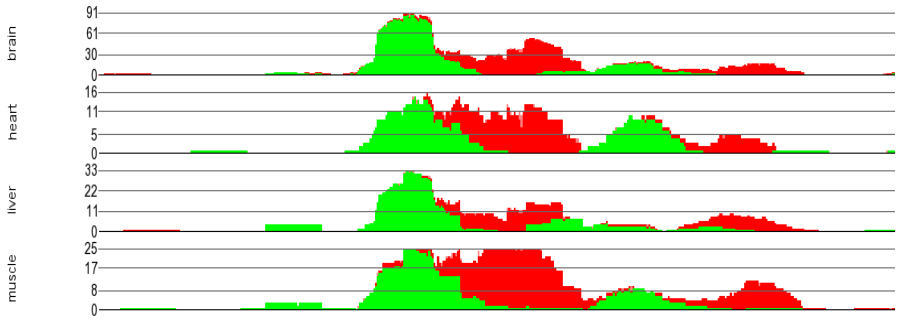


Figure 6. 289: Viewing the Alignment tracks colored by Forward and Reverse strands

The *Bases* color option will color the results according to the base (GATCN) of the read (Figure 6. 290). When the color is set to bases the view must be zoomed in far enough to distinguish base pairs – otherwise the plot will by colored by matches.

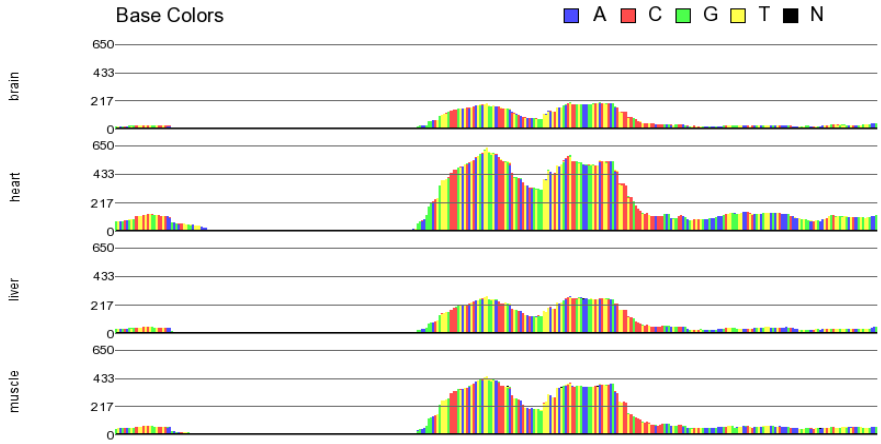


Figure 6. 290: Viewing the Alignment tracks colored by Base Colors

The *Forward Codons* and *Reverse Codons* color options color will be colored according to the codon of the read. It can be used to determine if a mutation causes a change in amino acid. Figure 6. 291 shows the *Alignment* track color by the *Forward Codons*.

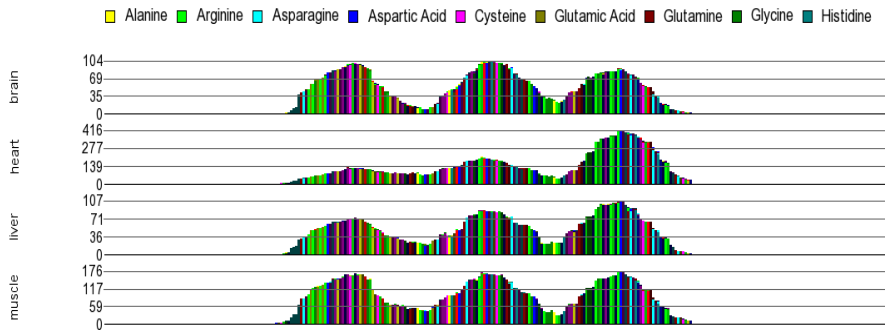


Figure 6. 291: Viewing the Alignment tracks colored by Forward codons

The *Sample color* options will color the results according to the color of the samples in the dataset (Figure 6. 292).

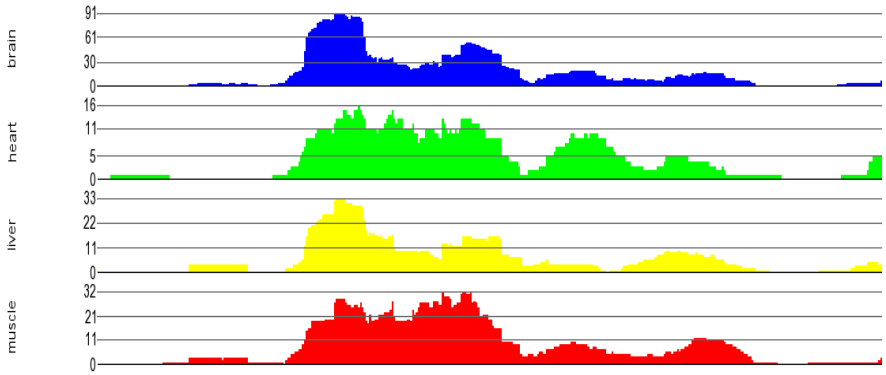


Figure 6. 292: Coloring by sample to match the colors in the isoform proportion view

Classes of reads to display forwards and reverse reads can be configured separately using the *Filter* tab (Figure 6. 293). This includes by *single and forward reads, single end reverse reads, paired end forward-forward reads, paired end forward-reverse reads, paired end reverse-forward reads* and *paired end reverse-reverse reads*.

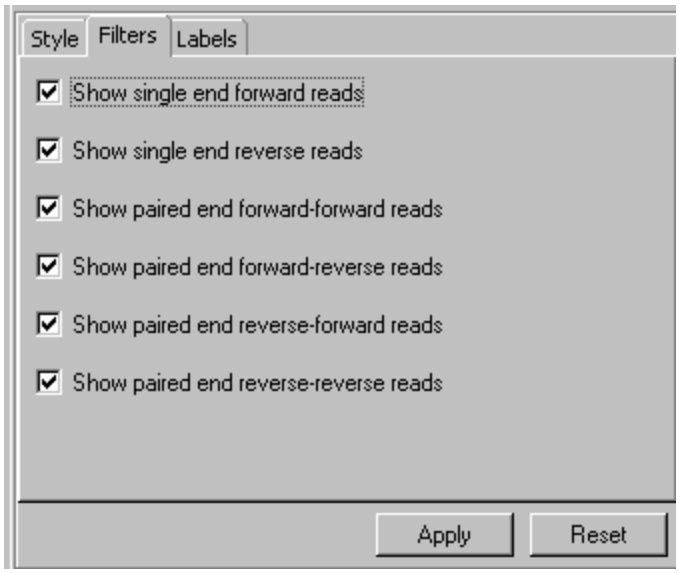


Figure 6. 293: Configuring the Filter tab to display forward and reverse reads

The *Labels* tab can edit the *Track Title*, *Track Title Size* and *Label Size* of the *Alignment* track (Figure 6. 294).

Figure 6. 294: Editing the Label properties of the Alignment track

SNP Proportion Track

The *SNP Proportion* track (Figure 6. 295) gives a graphical representation of the relative SNP abundance for each sample at each genomic location where one is found. The SNP list is created using the *Variations across Samples* option of the RNA-Seq workflow.

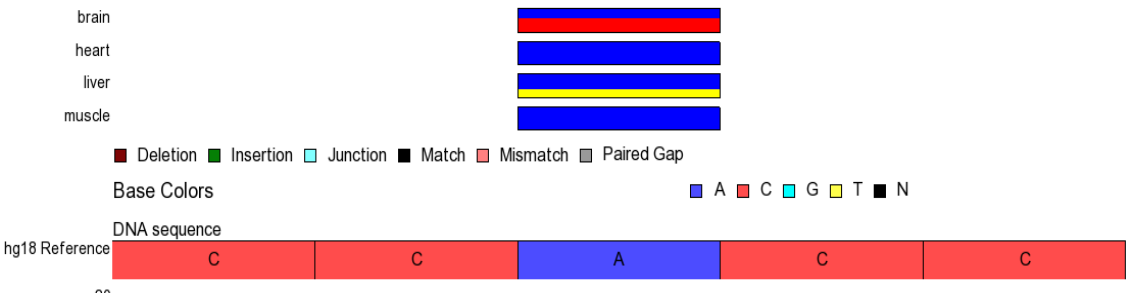


Figure 6. 295: Viewing the SNP Proportion track

Configure the colors of the *SNP Proportion* track from the *Color* tab using the *Configure base colors button* (Figure 6. 296).

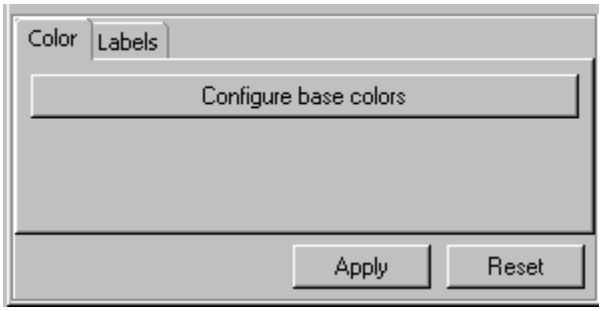


Figure 6. 296: Configuring the base colors of the SNP Proportion track

Edit the *Track Title*, *Title Size* and *Label Size* of the *SNP Proportion* track under the *Labels* tab (Figure 6. 297).

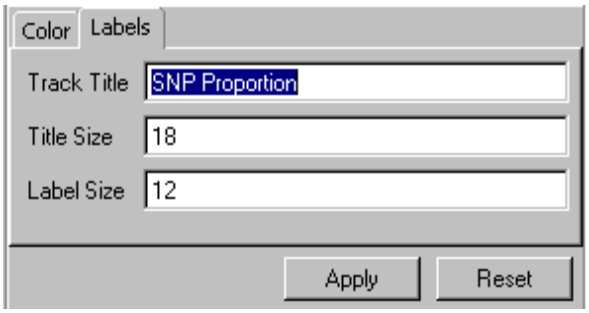


Figure 6. 297: Editing the Track Title, Title Size, and Label Size of the SNP Proportion track

The *SNP Proportion Legend* displays the color configuration of the *SNP Proportion* track (Figure 6. 298). Editing the legend colors will not edit the colors on the *SNP Proportion* track.

Figure 6. 298: Viewing the SNP Proportion Legend of the SNP Proportion track

The *Color* tab allows you to configure the colors of the *SNP Proportion Legend* track using the *Configure colors* button (Figure 6. 299).

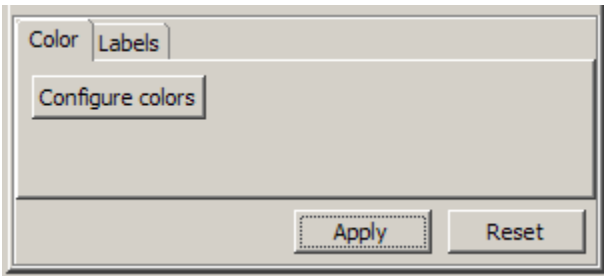


Figure 6. 299: Configuring the base colors of the SNP Proportion Legend track

Edit the *Track Title*, *Title Size* and *Label Size* of the *Labels* tab (Figure 6. 300).

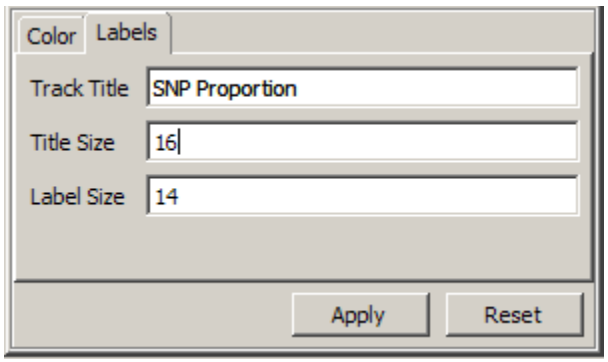


Figure 6. 300: Editing the Track Title, Title Size, and Label Size of the SNP Proportion Legend track

ChIP-Seq Tracks

The ChIP-Seq tracks are used to identify *in vivo* transcription factor binding sites across the entire genome including motif binding sites and enriched regions (Figure 6. 301).

Figure 6. 301: Viewing the ChIP-Seq tracks in the Chromosome View

Motif Binding Sites Track

The *Motif Binding Site* track displays the locations of the instances of detected known or de novo motifs (Figure 6. 302).

Figure 6. 302: Viewing the Motif Binding Site track showing instances motifs

From the *Track* tab you can adjust the height of the track using the *Track Height* slider (Figure 6. 303). Moving the slider to the right will increase the height, moving the slider to the left will decrease the height.

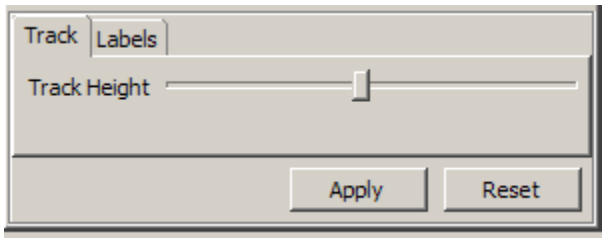


Figure 6. 303: Adjusting the height of the Motif Binding Site track

The *Labels* tab allows you to edit the *Track Title* and *Title Size*, or turn on/off labels per stack or per region (Figure 6. 304). The *Label Column* to display from the spreadsheet attributes can be specified. These will only be visible if you are zoomed in far enough. The labels can be changed to *Show one label per stack*, *Label every region*, or *Turn off labels*.

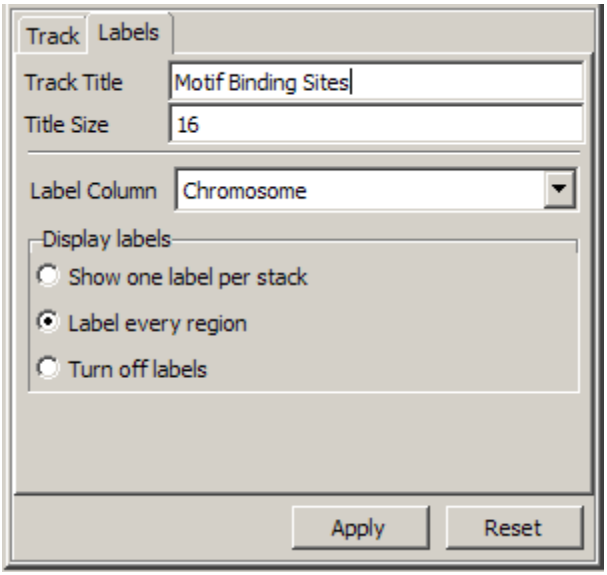


Figure 6. 304: Editing the Track Title, Title Size, and label properties of the motif region

Region Track

The *Region* track in Figure 6. 305 gives the enriched regions detected using the *Create List of Enriched Regions* step from the ChIP-Seq workflow. Regions that contain a binding site for the transcription factor of interest will have many sequence reads mapped to it. Lists of regions are created by looking at the Peaks and identifying regions in or not in a sample and in or not in a control sample.

Figure 6. 305: Viewing the Region track showing regions in chip sample but not in mock sample

The *Profile* tab of the *Region Track* allows you to separate the peaks by options in the drop down menu (Figure 6. 306). If a spreadsheet has genomic features on rows and a sample ID column then, by default, there will be one row per sample. The *Track Height* slider is used to adjust the height of the track in the view.

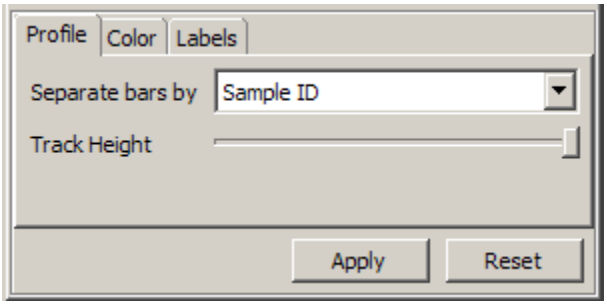


Figure 6. 306: Setting properties of the Profile tab of the Region Track

The *Color* tab is used to adjust how the bars are to be colored (Figure 6. 307). The *Min* and *Max* inputs control the color scale of how the *Color bars by* attribute are drawn. Select the *Configure category colors* button to configure the colors of the attribute selected in the *Color bars by* dropdown menu.

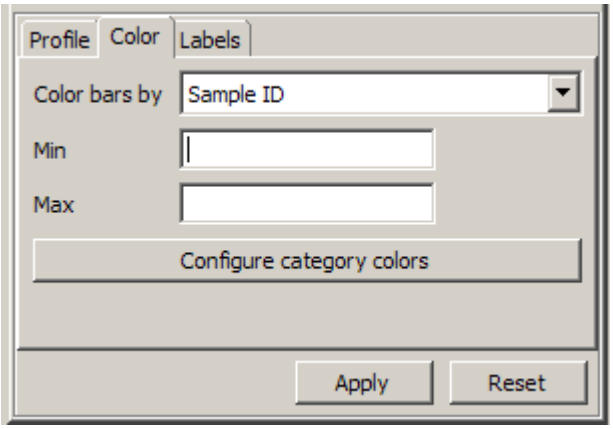


Figure 6. 307: Configuring the colors of the Color tab of the Region Track

The *Labels* tab is used to edit how the labels of the track are displayed, and whether or not to display all or only the selected sample (Figure 6. 308).

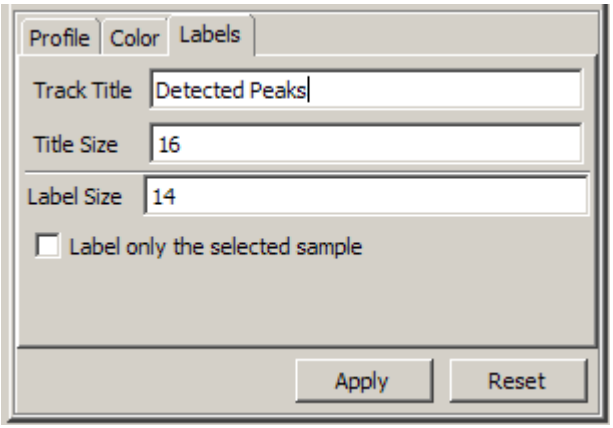


Figure 6. 308: Editing the Track Title, Title Size and Label only the selected sample of the Region track

Region Bar Profile Track

The *Region bar profile* by *New Track > Other (Advanced) > Region bar profile* adds a track which displays the mapped reads coverage and allows you to configure the ways the coverage is displayed (Figure 6. 309).

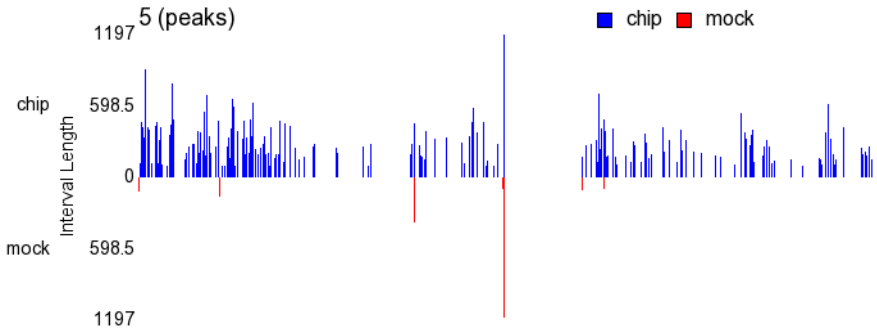


Figure 6. 309: Viewing the Region bar profile track showing chip vs. mock samples with height as interval length separated by Sample ID

From the *Profile* tab, check *Separate bars by* to separate the bars by the available attributes in the spreadsheet (Figure 6. 310). The histogram bar height is determined by the selected attribute in the *bar height by* drop down menu. The *Min* and *Max* values set the Y-axis scale of the track. The *Bars come from* feature sets the baseline from which to extend the histogram height.

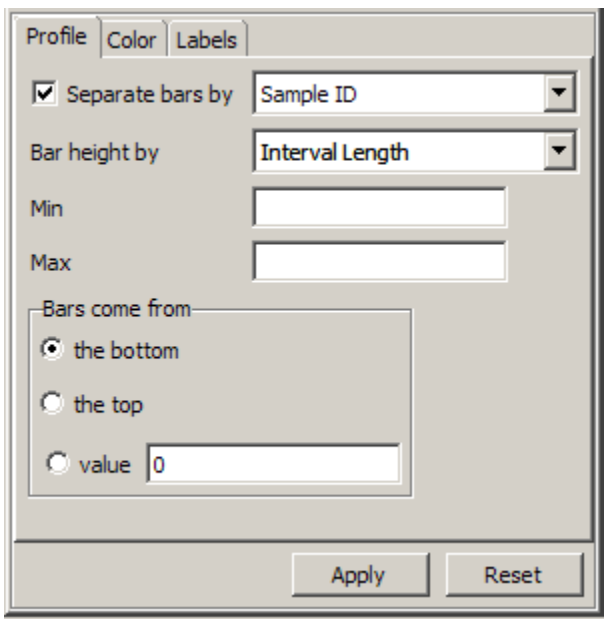


Figure 6. 310: Configuring the Profile tab of the Region bar profile tab

Color bars by will determine which attribute is used to color the histogram bars. Select *Configure category colors* to change the colors of the histogram bars by attribute (Figure 6. 311). The *Min* and *Max* values set the color range intensity values but will only be noticeable if specific attributes are selected for the *Color bars by* drop down menu.

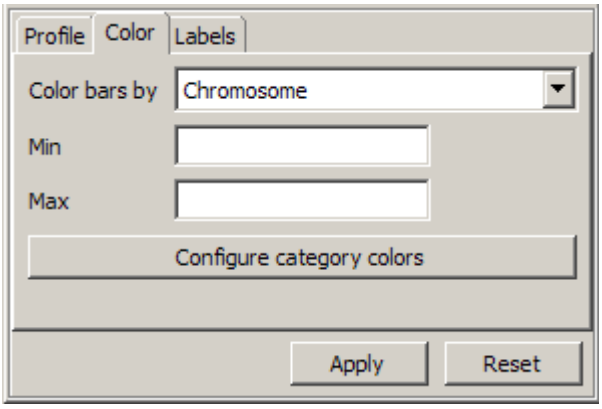


Figure 6. 311: Configuring the Color tab of the Region bar profile tab

The *Track Title*, *Title Size* and *Label Size* can be changed from the *Labels* tab (Figure 6. 312).

Figure 6. 312: Editing the Track Title, Title Size and Label only the selected sample of the Region track

Sequence Heat Map Track

Sequence Heat Map will add a heat map track & alignment track (Figure 6. 313). This option is specifically designed to display a heat map of sequence data. The color of cells is based on reads per kb per million reads (RPKM). Initially the alignment track is empty. Select a sample in the *Sequence Heat Map* to populate the alignment track.

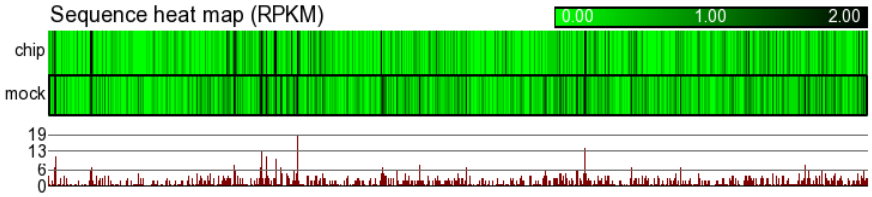


Figure 6. 313: Viewing the Sequence Heat Map with mock sample selected. Each row corresponds to one sample. Alignment track below is populated with alignments from sample as samples are selected in Sequence Heat map

Figure 6. 314 shows how to adjust the height of the *Sequence Heat Map* track using the *Track Height* slider. Moving the slider to the right will increase the height of the track. Moving the slider to the left will decrease the height of the track.

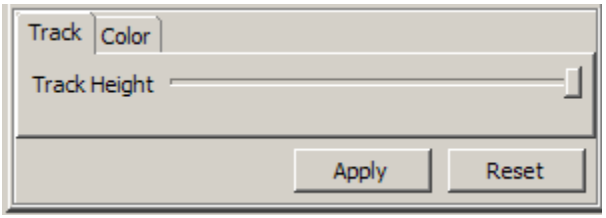


Figure 6. 314: Adjusting the height of Sequence Heat Map track

The color of the *Sequence Heat Map* can be changed using the *Color* tab (Figure 6. 315). The *Min* and *Max* colors can be set, and the heat map *Max color* intensity can be changed.

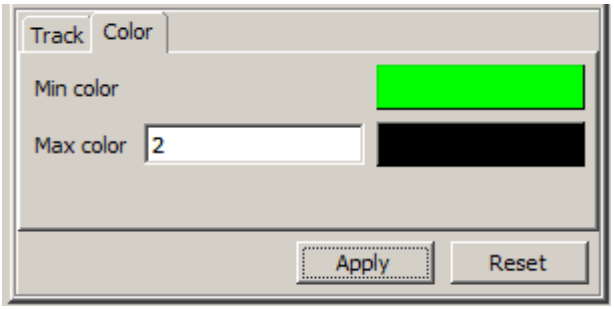


Figure 6. 315: Adjusting the Color tab of the Sequence Heat Map track

DNA-Seq Tracks

DNA-Seq is useful for looking at Mendelian inconsistencies, SNP, and inheritance information.

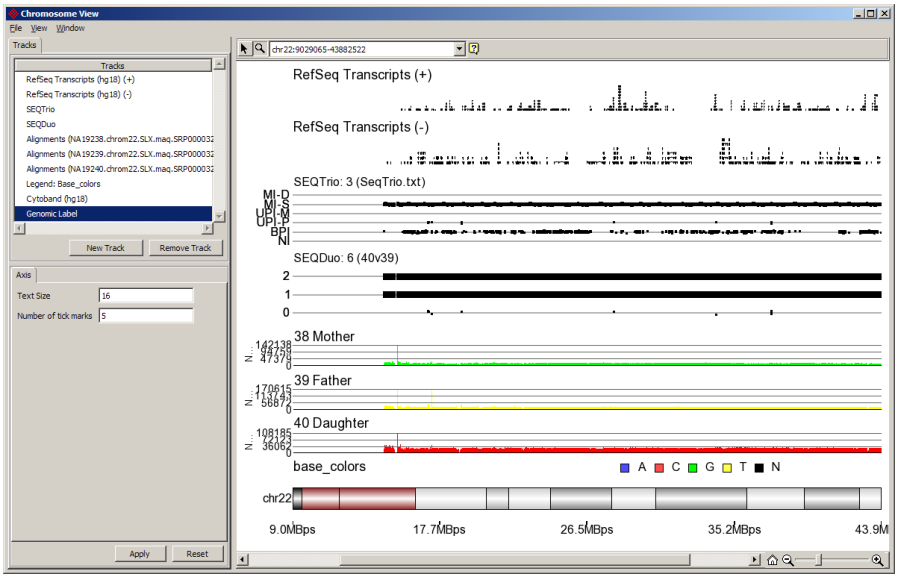


Figure 6. 316: Viewing the DNA-Seq tracks in the Chromosome View

SeqDuo Track

The *SeqDuo* track displays the number of concordant alleles between two samples' genotypes for each SNP. Figure 6. 316 shows a *SeqDuo* track containing an *Idenity by State of 1*.

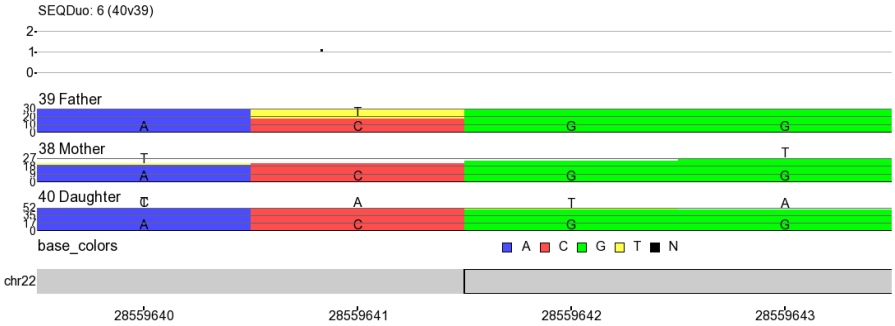


Figure 6. 316: Viewing the SeqDuo track with Identity by state

SNPTrio Track

The *SeqTrio* track displays information about Mendelian allele consistency and inheritance. Figure 6. 317 shows an example of a *SeqTrio* track with *Uniparental Inheritance* from the father's side.

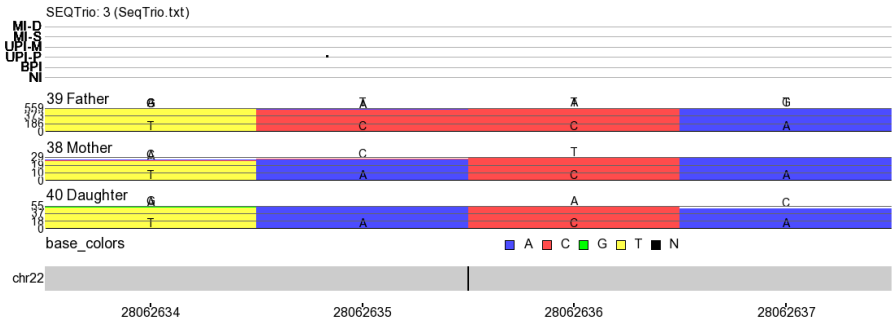


Figure 6. 317: Viewing the SeqTrio track with Uniparental Inheritance-Paternal

The *List 1* gives meanings for the *SeqTrio Scale* parameters.

MI-D	Mendelian Inconsistency – Double Alleles
MI-S	Mendelian Inconsistency – Single Allele
UPI-M	Uniparental Inheritance – Maternal
UPI-P	Uniparental Inheritance – Paternal
BPI	Biparental Inheritance
NI	Not Informative

List 1: Viewing the SeqTrio track scale parameters

More information regarding SNP Trio is available from http://pevsnerlab.kennedykrieger.org/SNPtrio04.htm.

Principal Components Analysis (PCA)

Partek's Pattern Visualization System® combines statistical analysis with visualization to help you see more of your data. In Partek, any time you try to plot more than 3-D data in a scatter plot, PCA is automatically used to show more of the information content in the high-dimensional data. In this case, about 55% of the information content in the entire 80-dimensional space was able to be visualized with a statistically driven 3-D visualization. (You can verify the 55% by adding the contributions of the PC's on the X, Y, and Z axes.) Remember when looking at a PCA mapped plot when two points are close together in the scatter plot; they are similar in 80-dimensional space. Likewise, if two points are very different in 80-dimensional space, they will be far apart in the PCA-mapped visualization of the data. The same cannot be said when simply plotting three of many variables in the data.

Multidimensional Scaling (MDS)

There are other ways to use statistical analysis combined with visualization to see more of high-dimensional data. Another useful technique that is related to PCA is called *Multidimensional Scaling* or *MDS* for short. PCA is a linear mapping that relays exactly how much of the information content is being displayed in the scatter plot. MDS, on the other hand, is a non-linear mapping of the data to a lower dimensionality for visualization (usually 2 or 3-D). While it does not have the advantage of giving exact numbers for information content revealed, it can outperform PCA in terms of preserving interpoint distance. The MDS option is found under **Tools > Discover > Multidimensional Scaling.**

Advanced Dimensional Reduction

Principal Components Analysis

Principal Components Analysis (PCA) is an exploratory technique that is used to describe the structure of high dimensional data by reducing its dimensionality (Jolliffe, 1986). It is a linear transformation that converts *n* original variables into *n* new variables ("PC's"), which have three important properties:

- The new variables (PC's) are ordered by the amount of variance explained
- The new variables (PC's) are uncorrelated
- The new variables (PC's) explain all variation in the data

PCA is a *Principal Axis Rotation* of the original variables that preserves the variation in the data. Therefore, the total variance of the original variables is equal to the total variance of the principal components. The *eigenvectors* and *eigenvalues* define the rotation and variation and are described as follows:

- The *eigenvalues* are the variances of the principal components.
- The *eigenvectors* are the direction cosines of the new axes (PCs) relative to the old (original variables), thus they define the rotations of the original axes

The method of PCA dates back to Harold Hotelling's 1933 paper "Analysis of a complex of statistical variables into principal components".

Configuring the PCA Dialog

The *Principal Components Analysis* (PCA) main dialog is shown in Figure 7. 1. By default, the current active spreadsheet is assigned as the spreadsheet to be analyzed, but any existing spreadsheet can be selected for analysis. To the immediate right of the *Data Source* are four colored accelerator buttons; these buttons will be described in detail later. The PCA dialog is invoked by going to **Tools > Discover > Principal Components Analysis** from the Partek main window.

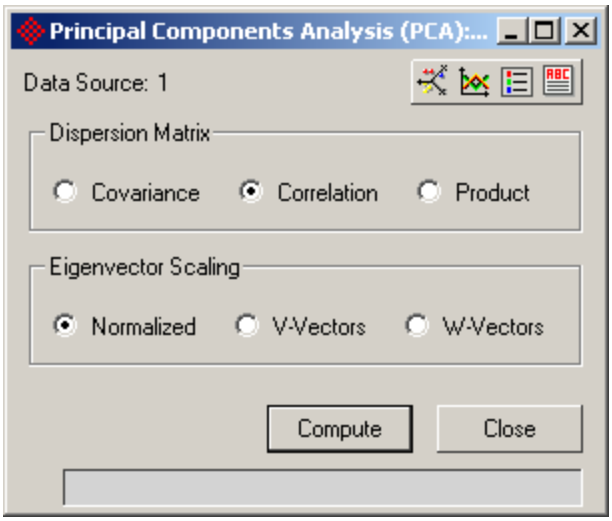


Figure 7. 1: Configuring the Principal Components dialog

Dispersion Matrix

The *Dispersion Matrix* module provides the choice of three dispersion matrices: covariance, correlation, and product.

Covariance

The covariance method operates on mean-centered data. During the computation of the covariance matrix, the data is automatically mean-centered. This adjustment is performed during the computation and does not modify the original data. Use the covariance method when your variables are measured in the same units and have similar variances.

Correlation

The correlation method adjusts the data to be standardized to a mean of zero (mean-centered) and a standard deviation of one. This adjustment is performed during the computation and does not modify the original data. Use the correlation method when your variables are measured in different units and/or have widely differing variances.

Product

The product matrix (or second moment matrix) is not adjusted by the mean or standard deviation. If the data is mean centered, the product matrix method, and the covariance method will yield the same eigenvectors. This dispersion method is rarely used.

Eigenvector Scaling

The eigenvectors are the direction cosines of the new axes (PCs) relative to the old (original variables), thus they define the rotations of the original axes. The eigenvectors are typically scaled using three methods.

Normalized:

- Orthogonal and scaled to unity
- PCs are uncorrelated
- PCs have variance equal to their eigenvalues

V-vectors:

- Scaled to characteristic roots (eigenvalues)
- PCs have the same units as the original variables
- PCs have variances equal to the squares of their eigenvalues

W-vector:

- Scaled to the reciprocal of their characteristic roots

Viewing the PCA Bi-Plot

If you click the Bi-plot button ( ) before clicking **Compute** or any other accelerator button in the dialog, then only the first 3 PCs will be computed (unless there are missing values or the Product matrix is chosen). This is also true when invoking PCA from the scatter plot ( ) accelerator button on the spreadsheet. This allows the initial PCA plot to open quickly.

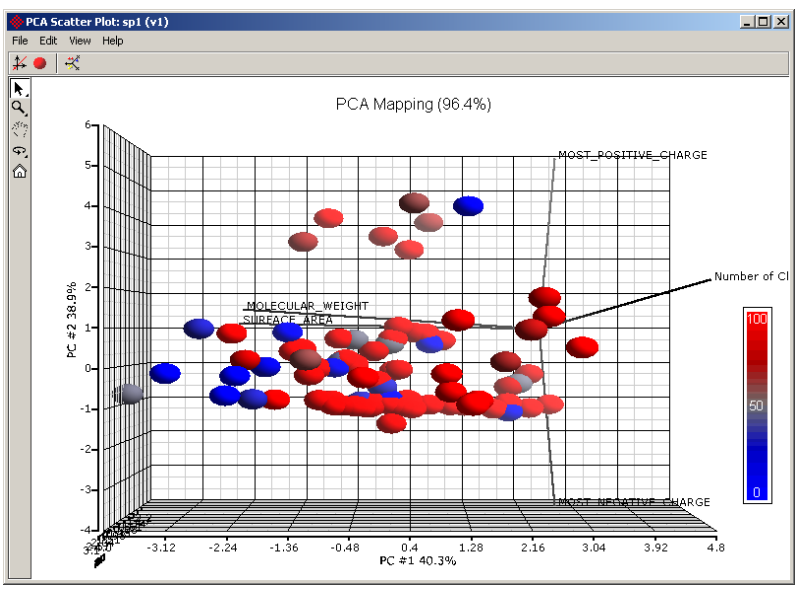


Figure 7. 2: Viewing the PCA Bi-plot

R-Analysis & Q-Analysis

If you have *p* variables measured on *n* observations, the Principal Components Analysis described above is referred to as *R-Analysis*. An R-Analysis refers to a pxp dispersion matrix (covariance or correlation) of variables where each transformed observation is represented by its PCs.

Similarly, *Q-Analysis* refers to an *nxn* dispersion matrix of observations where each transformed variable is represented by its PCs.

You can overlay these two representations in a scatter plot to obtain what is called a bi-plot.

Similarity among Observations

The observations (e.g. compounds) are represented by the points in the bi-plot.

- The distance between any pair of points is related to the similarity between the two observations in high dimensional space
- Observations that are near each other in the biplot are similar in a large number of variables
- Observations that are far apart in the biplot are different in a large number of the variables

Similarity among Variables

- Correlations between variables are related to the angles between the vectors (more specifically, to the cosine of the angles)
- Variables which have acute angles (<90°) between them indicate positive correlation; the smaller the angle, the stronger the correlation
- Variables that have obtuse angles (> 90°) are negatively correlated
- Variables whose angles are orthogonal (=90°) are uncorrelated
- Variables that project in the same direction (0°) have perfect positive correlation
- Variables that project in the opposite direction (180°) have perfect negative correlation

Reviewing the PCA Accelerator Buttons

The colored buttons at the top of the main PCA dialog are referred to as the *Accelerator buttons.* Accelerator buttons are used to invoke commonly used tasks. The four accelerator buttons are listed and described in Table 7. 1 below.

Accelerator Button	Action
	Invoke a PCA scatter plot
	Invoke a SCREE plot of the non-zero eigenvalues
	Dump PCA results to a new spreadsheet
	Create an HTML report of the PCA analysis

Table 7. 1: PCA accelerator buttons

Reviewing the Bi-plot Accelerator Button

The *Bi-plot accelerator* button is used to invoke a PCA scatter plot. By default, only the PCA samples are displayed. This is due to the fact that for very high dimensional data, overlaying the original variables is usually uninformative. The bi-plot parameters can be configured after invoking a PCA mapped scatter plot by clicking on the *Bi-plot Properties* button ( ) within the scatter plot viewer.

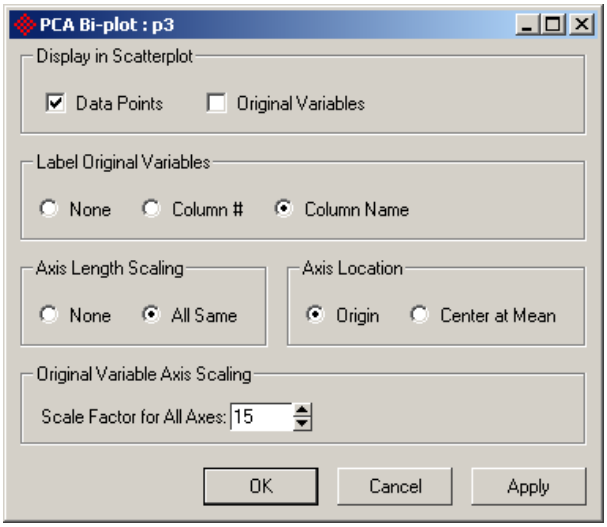


Figure 7. 3: Configuring the PCA Bi-plot Properties dialog

Display in Scatter Plot

This panel defines what is displayed in the PCA Bi-plot. By default, the *Data Points* are displayed and the *Original Variables* are not.

Original Variable Labeling

Labels for the original variables can be turned off or the original variables can be labeled by *Column Number* or *Column Name*.

Axis Length Scaling

By default, the length of each axis is scaled to be the same for all original variables. The length of each axis representing an original variable can also be scaled proportionally to the amount of variance for that particular variable. Thus, variables with long axes will have more variance than variables with smaller variances.

Note: the correlation method automatically scales the original data to a mean of zero and a standard deviation of one.

Axis Location

The location of the original variable axes can be centered at the origin of the PCA plot (0,0,0) or centered at the point representing the projected means of the original variables.

Original Variable Axis Scaling

Use this to apply a scaling factor to all variables.

Reviewing the SCREE Plot Accelerator Button

A SCREE plot (visually analyzing the eigenvalues) can be invoked by clicking on the **SCREE Plot** ( ) accelerator button.

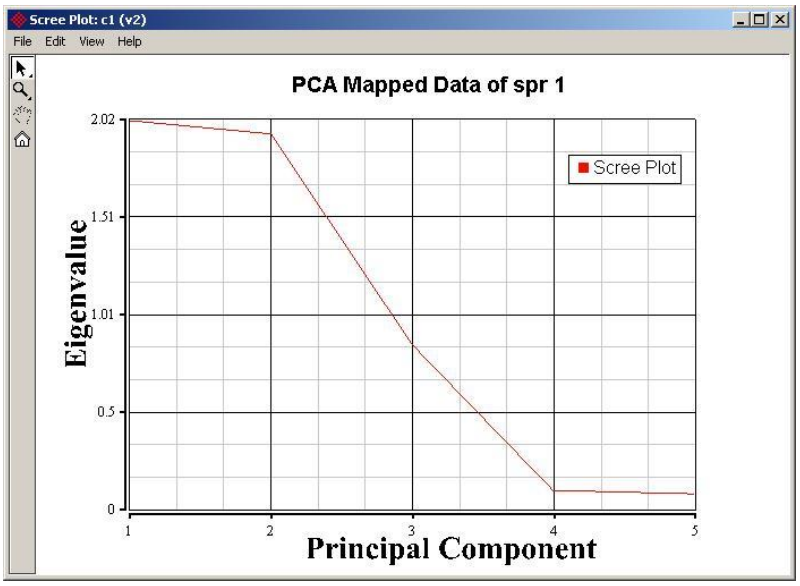


Figure 7. 4: Viewing the SCREE Plot

Reviewing the Dump Results to Spreadsheet Accelerator Button

The results available for dumping to a spreadsheet for further inspection are listed below. Clicking on the *Dump Results to Spreadsheet* ( ) button will dump the results.

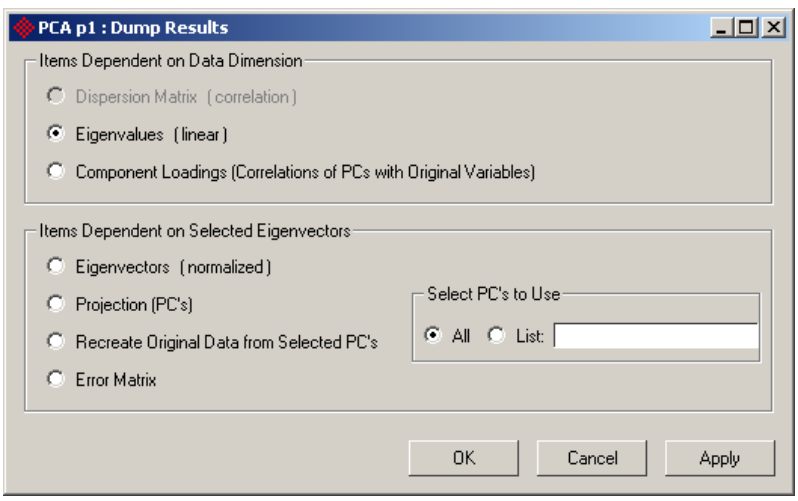


Figure 7. 5: Dumping the PCA results to the Analytical Spreadsheet®

Dispersion Matrix (type)

The selected dispersion matrix can be exported to a spreadsheet only when the number of rows is greater than or equal to the number of columns.

Eigenvalues

Exports the non-zero eigenvalues of the selected dispersion matrix.

Component Loadings (Correlations of PCs with Original Variables)

This spreadsheet will contain a row for each of the original variables and a column for each non-zero eigenvalue that contains the correlation with the eigenvectors.

Selected Eigenvectors

The following four items can be configured to use all of the eigenvectors corresponding to nonzero eigenvalues or the selected eigenvalues specified in *List*.

Eigenvectors (normalized)

This results in a spreadsheet containing the projection matrix with one row for each of the original variables.

Projection (PCs)

This spreadsheet will contain the sample information from the original spreadsheet and one column for each principal component.

Recreate Original Data from Selected PC's

Recreates the original data using all or only the selected eigenvectors.

Error Matrix

Computes the mapping error using all or the selected PCs

Viewing the PCA HTML Report

Clicking on the *HTML Report* button ( ) will invoke a dialog (Figure 7. 6) that gives options to specify the PCA HTML report (Figure 7. 7).

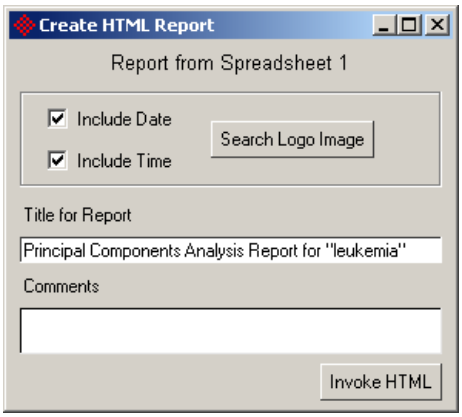
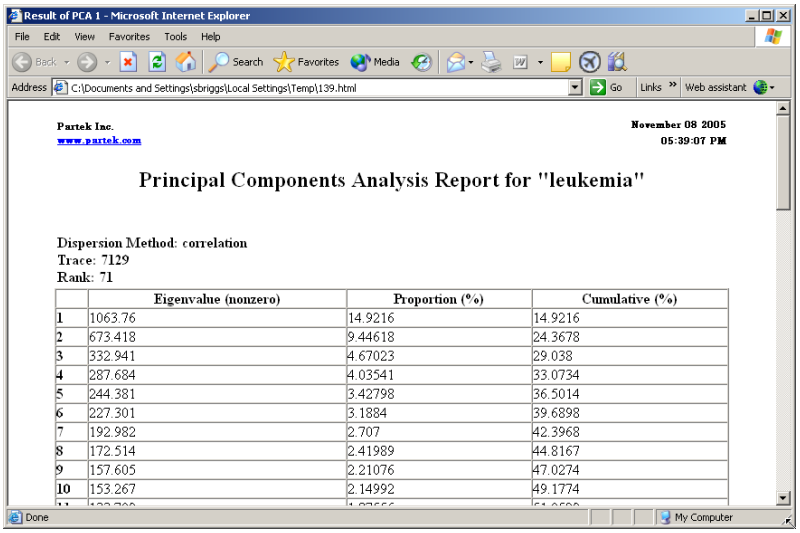


Figure 7. 6: Creating a HTML Report



Partek Inc.
www.partek.com

November 08 2005
05:39:07 PM

Principal Components Analysis Report for "leukemia"

Dispersion Method: correlation
Trace: 7129
Rank: 71

	Eigenvalue (nonzero)	Proportion (%)	Cumulative (%)
1	1063.76	14.9216	14.9216
2	673.418	9.44618	24.3678
3	332.941	4.67023	29.038
4	287.684	4.03541	33.0734
5	244.381	3.42798	36.5014
6	227.301	3.1884	39.6898
7	192.982	2.707	42.3968
8	172.514	2.41989	44.8167
9	157.605	2.21076	47.0274
10	153.267	2.14992	49.1774

Figure 7. 7: Viewing the PCA HTML report

PCA & Missing Data

Missing data in the computation of the PCA scatter plot is as follows: The covariance/correlation matrices are built using "available pairs" for each cell in their respective dispersion matrix. For example, the correlation between variables 1 & 2 is computed excluding all rows that have a missing value in either variable 1 or variable 2.

When projecting the data using the covariance/correlation eigenvalues, missing data values are replaced with the mean of its column. Because the covariance/correlation methods both mean-center the data prior to projecting it, this insures that the missing cell will have no contribution to the projected point.

This allows for plotting rows with missing data without having to completely remove the rows with missing data (case-wise deletion) or force the user to perform some sort of missing data imputation. The effect of the missing data is obviously data dependent and depends heavily on the proportion of missing cells to non-missing cells. It works very well for high dimensional data such as microarray or proteomic data.

PCA & Zero Variance Variables

For the purpose of robustness and the prevalence of zero variance variables in many types of scientific data, the computation of the dispersion method is set up to automatically deal with zero variance variables by effectively ignoring them. If a particular variable (column) has no variance, the eigenvector elements for that variable are set to zero.

Zero variance variables can be automatically detected using **Stat > Descriptive Statistics > Find Zero Variance Variables**.

PCA & Multidimensional Scaling

Multidimensional scaling (MDS) is a non-linear cousin of PCA. A brief comparison of the two methods is described in Table 7. 2 below.

PCA	MDS
Linear projection	Nonlinear projection
% information content known	% information content not known
Computationally efficient for large number of samples	Computationally inefficient for large number of samples (order n2 algorithm)
Meaningful orientation	Arbitrary orientation
Meaningful variables	Variables have no meaning
Preserves large dissimilarities better	Preserves small dissimilarities better
Performed on covariance or correlation similarities	Performed on any type of (dis)similarity

Table 7. 2: Brief comparison of PCA vs. MDS

Multidimensional Scaling

Multidimensional scaling (MDS) is mapping from high-dimensional space to a lower dimension. The purpose of multidimensional scaling is to provide a visual representation of the pattern of proximities (similarities, dissimilarities, or distance) among a set of objects. MDS plots the objects on a map such that objects that are very similar to each other are placed near each other on the map and objects that are very different from each other are placed far away from each other on the map.

Implementation Details

Classical MDS

Classical Scaling treats dissimilarities directly as Euclidean distances, and then makes use of the spectral decomposition of a doubly centered matrix of dissimilarities. Classical MDS and Principle Components Analysis (PCA) are equivalent when the dissimilarities for classical scaling are chosen to be Euclidean distances. This technique is often referred to as Principal Coordinates Analysis (PCO). Classical MDS preserves large dissimilarities well.

Nonlinear MDS

Nonlinear MDS methods minimize a cost function that describes how well the pairwise distances in a data set are preserved. The most well known method of Nonlinear MDS is Sammon's Method. Nonlinear MDS models will preserve small dissimilarities well.

Multidimensional Scaling Dialog

- Open the *Multidimensional Scaling* dialog by selecting **Tools > Discover > Multidimensional Scaling…** from the Partek main window (Figure 7. 8)

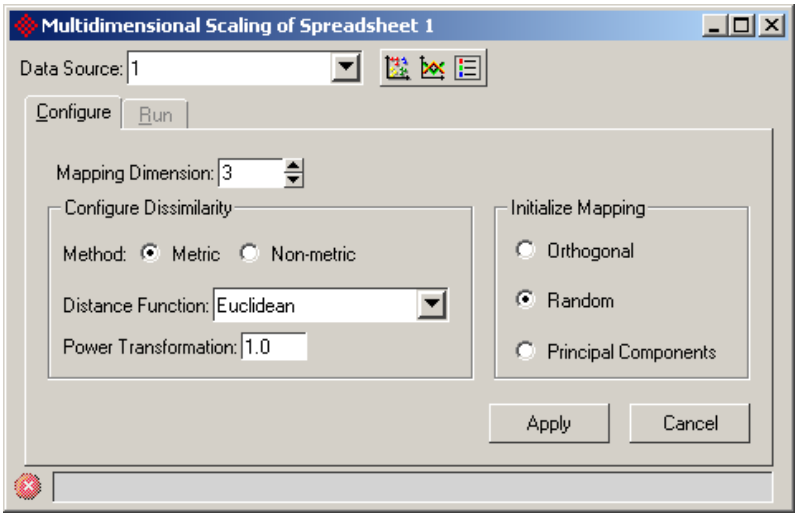


Figure 7. 8: Configuring the Configure tab in the Multidimensional Scaling dialog

You will use this dialog to specify Mapping Dimensions, Methods of Dissimilarity, Distance Functions, Power Transformations, and to Initialize Mapping.

Metric MDS

Metric MDS is performed on measured proximity data (interval or ratio). Classical and Nonlinear MDS are examples of Metric MDS.

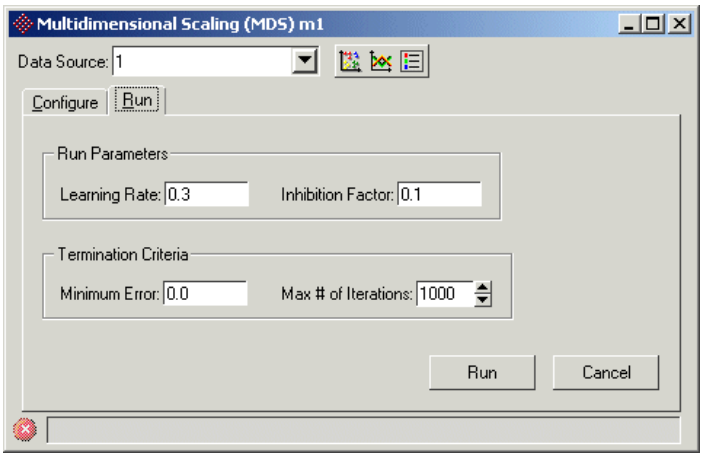


Figure 7. 9: Configuring the Run tab in the Multidimensional Scaling dialog

References

Hotelling, H. “Analysis of a complex of statistical variables into principal components” J. Educ. Psych 1933, 26:417–441.

Jolliffe, I.T. *Principal Component Analysis*, Springer-Verlag, New York, 1986.

Hierarchical & Partitioning Clustering

Introduction

Hierarchical and partitioning clustering methods are described in this chapter.

Hierarchical Clustering

Hierarchical clustering is used to group similar objects into "clusters." In the beginning, each row and/or column is considered a cluster. In hierarchical clustering, the two most similar clusters are combined and continue to combine until all objects are in the same cluster. Hierarchical clustering produces a tree (called a dendrogram) that shows the hierarchy of the clusters.

Creating Clusters

To invoke the *Creating Hierarchical Clusters* dialog from the Partek main window select **Tools > Discover > Hierarchical Clustering…**

What to Cluster and Normalization

The cluster will be performed on either rows or columns or both or null by checking/unchecking the *Row* and *Column* boxes (Figure 8.1). The computation includes all the numeric response variables. Before the clustering calculation, you can choose to normalize the data, either standardize or shift the data, which will only be performed on spreadsheet columns. Standardize will make each column mean as zero, std. dev as 1, this operation is making all the columns have equal weight. Shift will make each column mean as zero. Choose *none* will perform clustering on the values in the spreadsheet.

If both *Rows* and *Columns* are unchecked, the heatmap will be in spreadsheet order

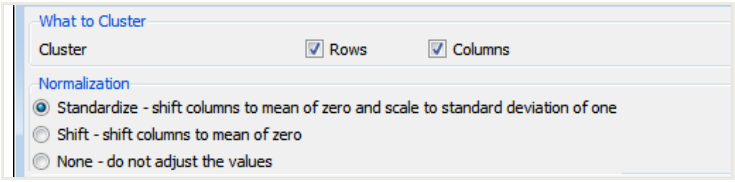


Figure 8.1: Configuring the what to Cluster panel

How to Cluster

There are about twenty different method on calculate the row/column distance and five methods on how to calculate cluster (Figure 8.2)

Figure 8.2: Configuring the How to Cluster panel

Row / Column dissimilarity

Row/Column dissimilarity is used to determine the distance between two rows or columns. See *Chapter 9 Descriptive Statistics, Correlation, & Measures of Similarity & Dissimilarity* for details on each of the distance measures.

When clustering genotype data, the only option is *Genotype Distance Measure*. The distance between two elements is the number of different alleles in the genotype. The distance between AA and AB is one, and the distance between AA and BB is two. NoCalls (NC) are ignored. The distance between two vectors is the square root of the sum of the squared differences.

Row / Column method

Row/Column method is used to determine how the distance between two clusters is calculated.

Single linkage: The distance between two clusters is determined by the distance of the two closest objects ("nearest neighbors") in the two clusters (Figure 8.3). Single linkage tends to produce clusters that form long "chains" or "strings." This typically results in a smaller amount of variance in the height of clusters. Figure 8.3 shows single linkage on 1, 2, and 3. First, 1 and 2 are combined with a distance of 1. Next, the clusters (1, 2) and (3) are combined with a distance of 1 (the distance from 3 to 2). The two clusters appear to be merged since they have the same distance.

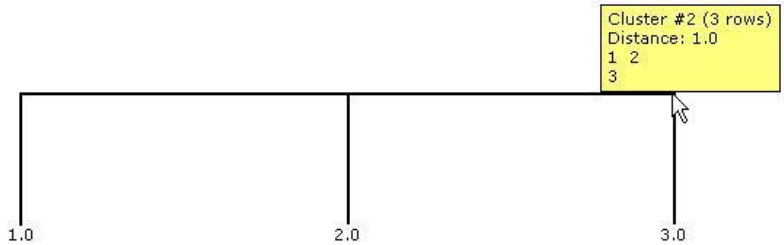


Figure 8.3: Viewing a single linkage

Complete linkage: The distance between two clusters is equal to the distance between the two furthest members of those clusters (Figure 8.4). The distance between two clusters is determined by the largest distance between any two objects in the two clusters ("furthest neighbors"). Complete linkage tends to produce clusters that are spherical and compact. This method usually performs well when the objects actually form naturally distinct "clumps." Figure 8.4 shows complete linkage on 1, 2, and 3. First, 1 and 2 are combined with a distance of 1. Next, the clusters (1, 2) and (3) are combined with a distance of 2 (the distance from 3 to 1).

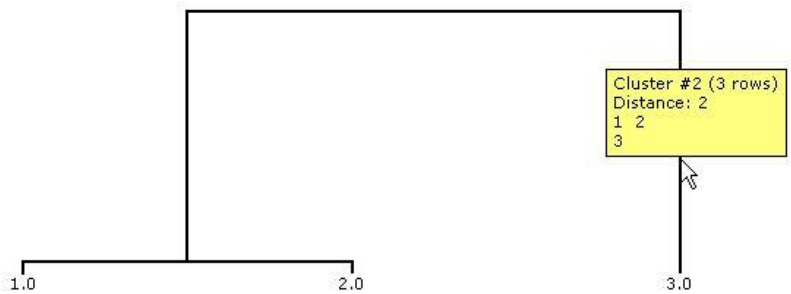


Figure 8.4: Viewing a complete linkage

Average linkage: The average distance between all pairs of objects in the two different clusters is used as the measure of distance between the two clusters (Figure 8.5). This method is effective when the objects form natural distinct "clumps," as well as when the data form elongated "chain" type clusters. This method is commonly referred to as "UPGMA", or "un-weighted pair-group method using arithmetic averages" (Sneath & Sokal, 1973). Figure 8.5 shows average linkage on 1, 2, and 3. First, 1 and 2 are combined with a distance of 1. Next, the clusters (1, 2) and (3) are combined with a distance of 1.5 (the distance from 3 to 2 plus the distance from 3 to 1, divided by 2).

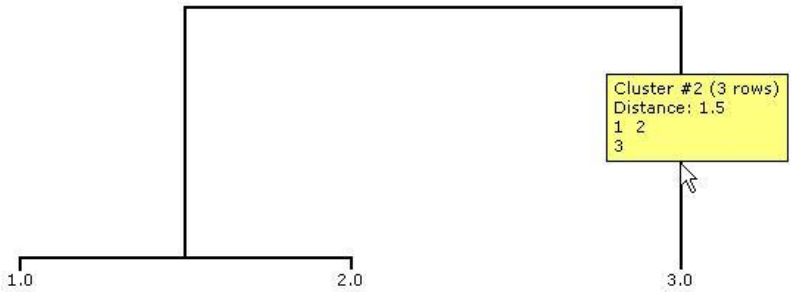


Figure 8.5:Viewing an average linkage

Centroid method: The distance between two clusters is equal to the distance between the centroids of the clusters. The distance is calculated using Gower's formula so that the centroid for a parent cluster lies geometrically on the line between the two children. This method is also called "UPGMC" or "unweighted pair-group method using the centroid approach" The centroid of a cluster is defined as its mean vector. Figure 8.6 demonstrates centroid method applied to 1, 2, and 3.

First, 1 and 2 are combined with a distance of 1. Next, the clusters (1, 2) and (3) are combined with a distance of 1.25.

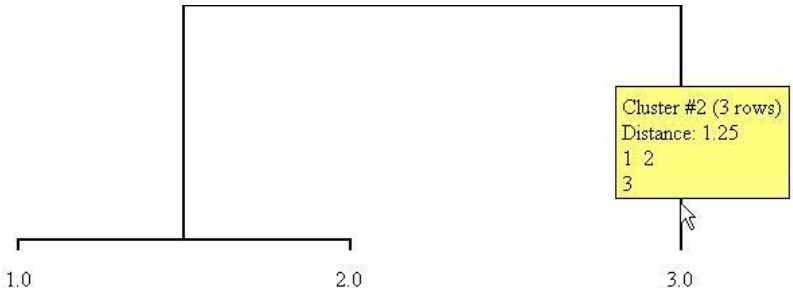


Figure 8.6: Viewing the centroid method

This distance function can lead to reversals. That is to say, that cluster distances do not necessarily always increase; it is possible for a parent cluster to have a lower distance than its children. For example, in you see that cluster 4 (joined after cluster 3) has a lower distance.

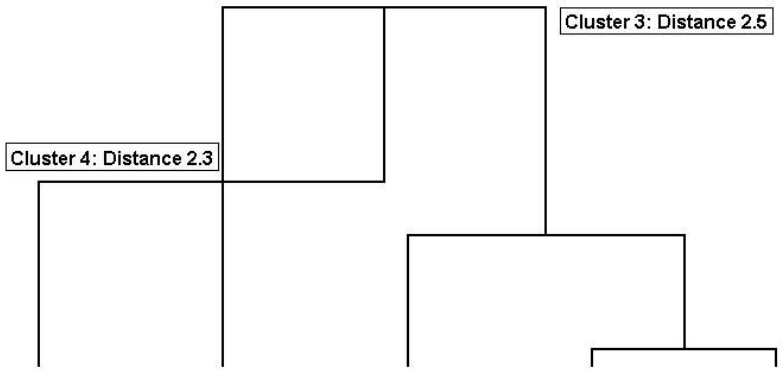


Figure 8.7: Viewing a reversal

Ward's method: The distance between two clusters is designed to minimize the size of an error measure based on the sum of squares. This method tends to result in spherical clusters. Figure 8.8 shows Ward's method applied to 1, 2, and 3.

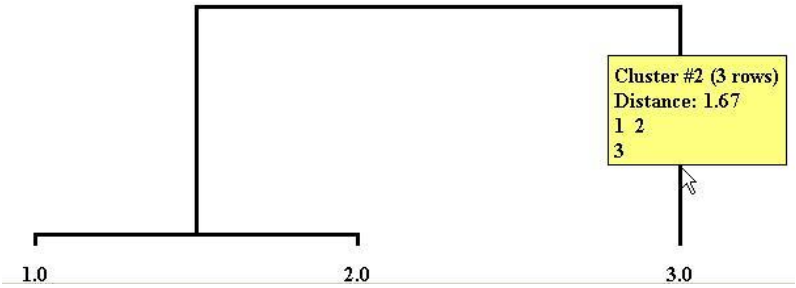


Figure 8.8: Viewing Ward's method

Running Cluster Analysis

Click **OK** or **Apply** to begin clustering. Hierarchical clustering requires a significant amount of memory and processing time on large data sets. It is recommended that any unnecessary programs be closed. If both rows and columns are selected for clustering, columns are clustered first. If the clustering is aborted while grouping the columns, it will proceed to cluster the rows.

Clustering Stages

The progress bar indicates the progress of the computations. It shows the progress of the following stages while clustering:

Stage 1: Reading record i of n: Loading the data from chosen rows and columns

Stage 2: Calculating dissimilarity i of n: Calculating the interpoint dissimilarity between all the rows or columns

Stage 3: Finding neighbor i of n: For each row or column, finding the row or column that is nearest

Stage 4: Creating cluster i of n: Grouping the two most similar entities (row/column or cluster), and updating neighbor distance, if necessary

If the *Abort* button is pressed before "Creating cluster i of n" then the result will be the same as if "Cluster rows" or "Cluster columns" was not selected.

Post-clustering messages include:

Ordering record i of n: Ensures that the left and right branches of each cluster are in accordance with the chosen order

Organizing record i of n: Ordering the intensity plot so that clusters do not overlap

Viewing the results

Heat Map

The heat map (also called intensity plot) can be shown even if no clusters are created. Each value from the specified rows and columns is drawn as a cell. The color of each cell is determined by the continuous color map (Figure 8.9).

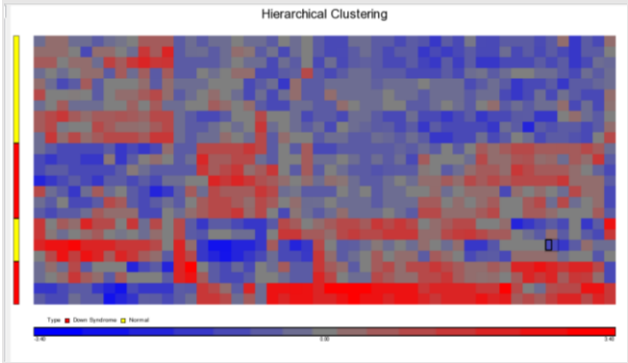


Figure 8.9: Viewing Heat Map

Select HeatMap tab on the left configuration panel to render the plot (Figure 8.10)

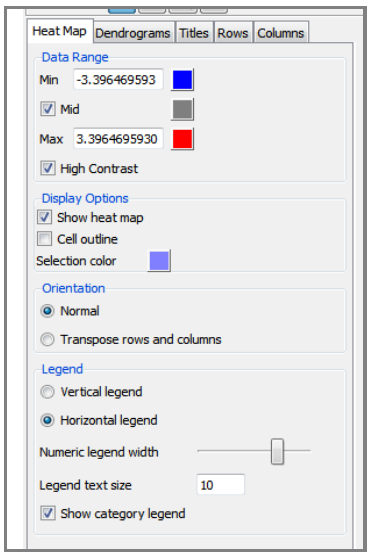


Figure 8.10: Heatmap configuration dialog

The default color map is blue-grey-red representing small-middle-larger values respectively, you can click on the color square to pick a different color.

Check *High Contrast* option will have better contrast between min and max color by using discrete color steps in the color map.

Display option allows you to show/not show heat map, show/not show outline on the cells by selecting/deselecting the check boxes. When a row/column is selected, they are highlighted in a color, you can change the color by clicking the *Selection color* square.

There are two modes in orientation, default is normal which means the row and column in heatmap is the same orientation as the spreadsheet it invoked from, *Transpose rows and columns* will have rows in heatmap represent columns on the spreadsheet and columns in heatmap represent rows on the spreadsheet, in other words, you pivot the plot in 90 degrees.

Legend of the heatmap color can be displayed horizontally at the bottom of the plot, or vertically at the right of the plot by selecting *Horizontal legend* and *Vertical legend* respectively. You can change the width of the legend bar by sliding *Numeric legend width* option.

Dendrograms

If rows or columns are clustered, then the results of each clustering will be shown as a dendrogram (Figure 8.11).

Figure 8.11: Viewing a dendrogram

Click on Dendrograms tab to configure the row/column dendrogram (Figure 8.12)

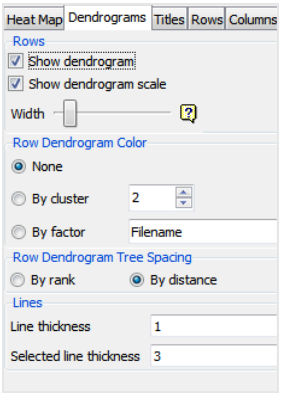


Figure 8.12 Configuring dendrogram dialog

The clusters that contain only a row or a column are called “leaves”. All other clusters contain exactly two objects. A cluster may contain two leaves, a leaf and a cluster, or two clusters. The last cluster created contains all rows/columns and is called the “root”.

You can check/uncheck to show or hide dendrogram and dendrogram scale. Draging the *Width* sliding bar will change the proportion of the dendrogram section vs heatmap section, e.g. if you want to show more detailed information on row dendrogrm, slide the bar to the right, after click **Apply**, the row dendrogram will takes more horizontal space, the heatmap section will decrease.

By default, *dendrogram color* is set to **None**, which means it is in black. Specify the number of colors to use when coloring *By cluster*. The first color is assigned to the top cluster and its children (all clusters) (Figure 8.13). The next color is applied to the next highest cluster and its children. This continues until all colors are assigned. If the number of colors is equal to or exceeds the number of clusters then each cluster will be colored uniquely.

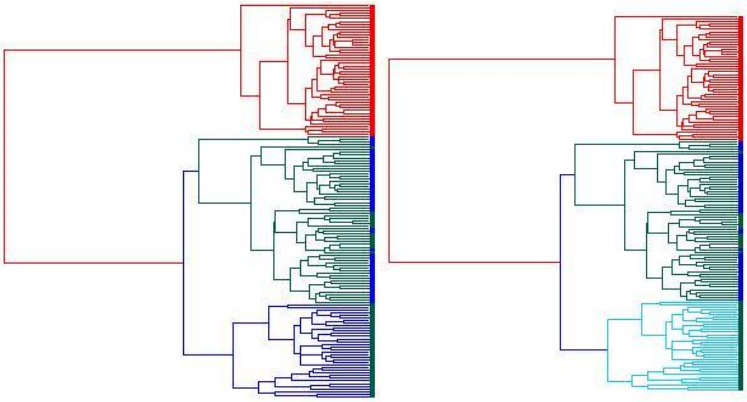


Figure 8.13: Viewing the dendrogram colored by cluster: 3 colors (left), 4 colors (right)

When coloring by a categorical column, the top of a cluster is drawn using the axis color if the two members are not the same color (Figure 8.14).

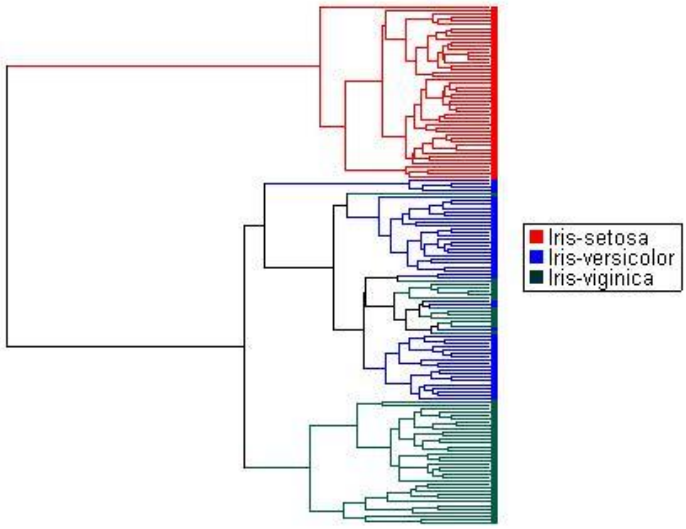


Figure 8.14: Viewing the dendrogram colored by categorical columns

The mouseover of a cluster tells distance of the two members of the cluster. The distance between the two members of the cluster determines its height (Figure 8.15). Groups of rows or columns that are similar will be combined with short clusters while tall clusters will separate dissimilar groups. The width of the cluster has no mathematical significance.

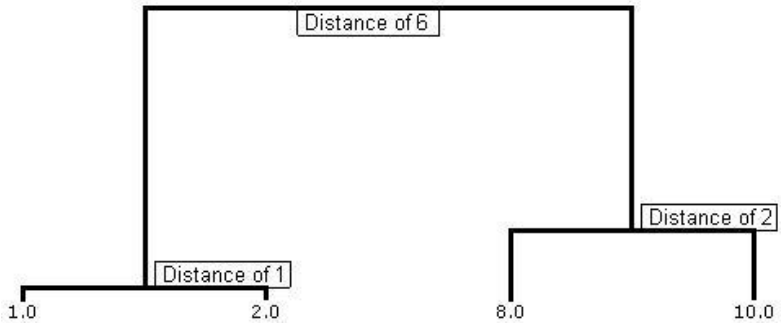


Figure 8.15: Viewing the distance and height

By default, dendrogram tree spacing is by distance. When there is a big range on the distances, it might be useful to change it to by rank.

Titles

You can edit titles of the plot or the axes, change the font of the titles (Figure 8.16).

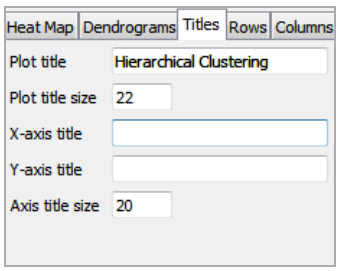


Figure 8.16. Configure font of the plot title and axes title

Rows

You can add more annotation on rows in spreadsheet in the plot, they can be color which is used to annotation categorical information, or text which is useful to label sample ID for instance (Figure 8.17).

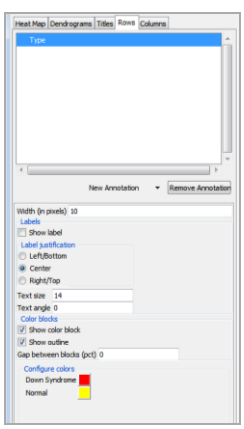


Figure 8.17. Configure row annotation dialog

Select a factor column from the *New Annotation* drop-down, the column will appear on the top panel, you can add multiple annotation columns on the top panel. You can add the same factor column multiple times also, can make them different configuration, e.g. one is color, one is text. You can select multiple annotations on the top panel to remove them by clicking **Remove Annotation** button.

When one factor is highlighted, we can render the highlighted annotation using the bottom section of the left panel.

Width option determines the how wide the annotation will take on the plot, if the annotation is text, make sure the section is wide enough to fit all the characters.

Select the check box of *Show label* will display the text of the annotation, you can select the label justification to align the text to the *Left* text, or center or *Right*. If the view is in transpose row and column mode, left will become bottom, right will become top. Type the *Text size* to change the font size, *Text angle* will change the orientation of the text.

You can select *Show color block* to display categorical annotation information, for just labeling purpose, you might uncheck this option.

Configure colors is only needed when you check the *Show color block*, click on each color square to change the color of the corresponding category.

Columns

There is only column label or gene symbol available for column annotation. You can adjust the label alignment, font size and angle of the text (Figure 8.18).

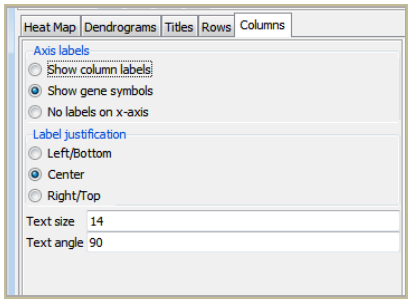


Figure 8.18: Column annotation

Mouse Mode

There are four mode options: selection mode , flip mode , zoom mode , pan mode

When you click on selection mode (), mouse over a cell in the intensity plot displays the value of the cell as well as the row and column of the cell. Clicking on a cluster will select that cluster and all of its children. When a cluster is selected, it is drawn with a thicker line. If a bounding box is used to select multiple clusters,

then the result will be the same as clicking the highest numbered cluster. The <**Shift**> keyboard key selects multiple clusters (and their children).

When you select some clusters, right click on any white space in the viewer, you will have more options (Figure 8.19).

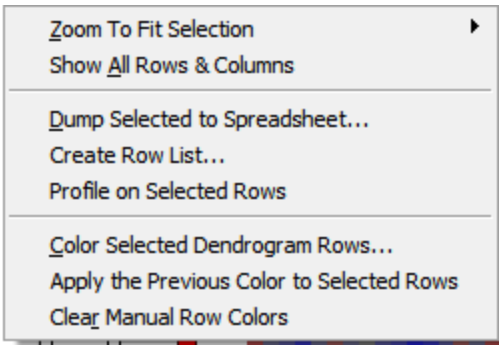


Figure 8.19: Right click on selected dendrogram options

- Zoom to Fit Select: can zoom the view to selected cluster in just row or column or both.
- Show All Rows & Columns: quick way to zoom out after zoom in
- Dump Selected to Spreadsheet: you can export the selected cluster to a spreadsheet, the saved data can be heap map values with the row and column in the viewer order; distances of the clusters on row/column and cluster membership of row/column (Figure 8.20). We will discuss this function in more details later.
- Create a list of the selected rows/columns
- Profile of selected rows/columns: you can verify the profile of the selected items in a cluster
- Color selected row/column: you can manually color any clusters, this option is useful when you want to define how many clusters you want to partition the data, and you can select different color for different cluster and export the cluster labeled rows/columns

When you in *flip mode* (), selecting a cluster will swap the left and right branches.

In zoom mode (), left click will zoom in to where you click, or you can draw a bounding box to zoom in certain regions. You can also scroll mouse wheel to zoom in/out. Click on the home button () at the lower right corner of the viewer, you can go back to the default for horizontal/vertical.

After you zoom in, you can switch to pan mode () to examine difference sections of the viewer.

Exporting to a Spreadsheet

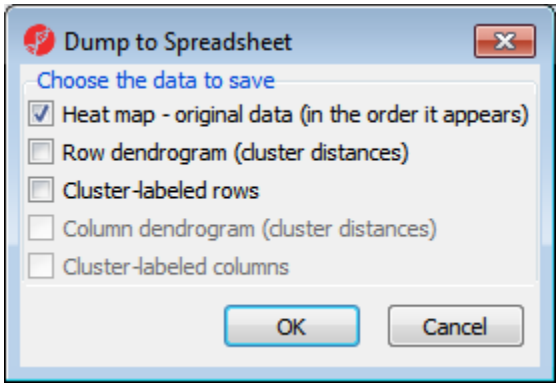


Figure 8.20: Configuring the Dump to Spreadsheet dialog

Heat map – original data (in the order it appears)

This option will export the intensity values to a new spreadsheet , the row and column will be in the order as the clustering viewer

Row/Column dendrogram (cluster distances)

Note: Negative number ids are the cluster id. Positive numbers correspond to the row/column number in this example (Figure 8.21).

	1. id	2. distance	3. left child	4. right child
1.	-1	3.24147	17	23
2.	-2	3.79675	-1	21
3.	-3	4.4114	-2	19
4.	-4	3.50534	16	15
5.	-5	3.99323	14	-4
6.	-6	4.75887	-3	-5
7.	-7	3.74994	22	18
8.	-8	3.88747	20	-7
9.	-9	5.09048	-6	-8
10.	-10	4.02079	8	7
11.	-11	4.27071	6	9
12.	-12	4.64445	-10	-11

Figure 8.21: Viewing the cluster distances

Cluster Distances dump the **id** of each cluster, the distance between its two children, and the id of its children. Cluster ids are negative; the positive numbers are row/column numbers.

Cluster-labeled rows/columns

Selecting *Cluster-labeled rows/columns* will export the data to a new spreadsheet with an extra column that contains the cluster assignments based on the color of the leaves of the dendrogram.

The dendrogram should be colored *By Cluster* or *Manually*.

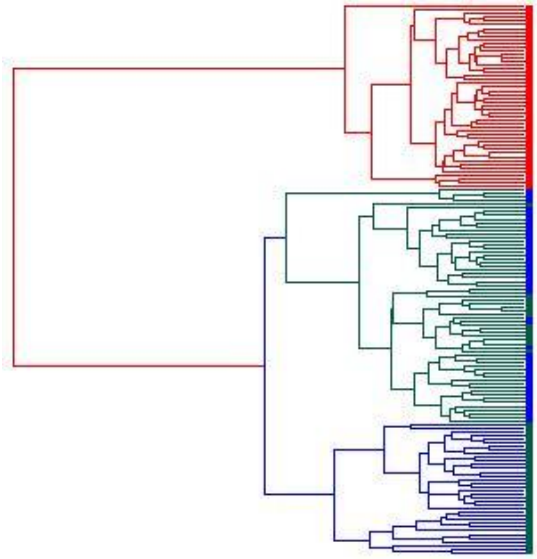


Figure 8.22: Viewing the colored dendrogram and resulting group profile

Partitioning Clustering

Partitioning Clustering provides a quick way to find groups in data. Partek offers two types of Partitioning Clustering algorithms: K-Means and Fuzzy C-Means.

Cluster Analysis Dialog

To open the *Cluster Analysis* dialog, select **Tools > Discover > Partitioning Clustering…**from the Partek main window.

The dialog shown in Figure 8. 1 is used to select the data file and distance function.

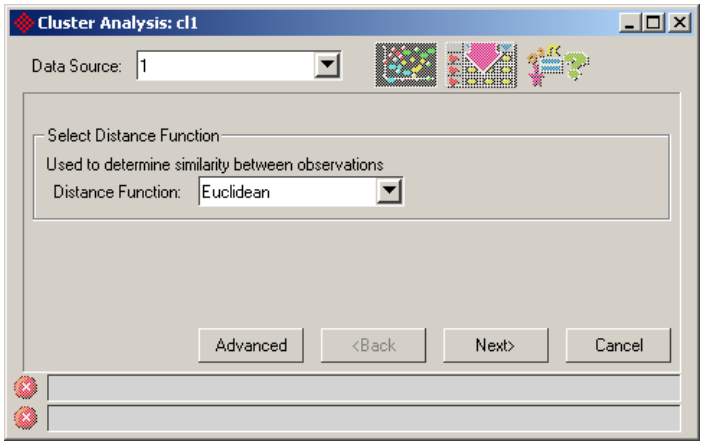


Figure 8. 1: Configuring the Cluster Analysis dialog

By selecting the **Advanced** button on the *Cluster Analysis* dialog, the analysis can be configured in more detail, such as *Clustering Method*, *Centroid Updating*, *Fuzzification*, *Initial Centers*, and *Termination Criteria* (Figure 8. 2).

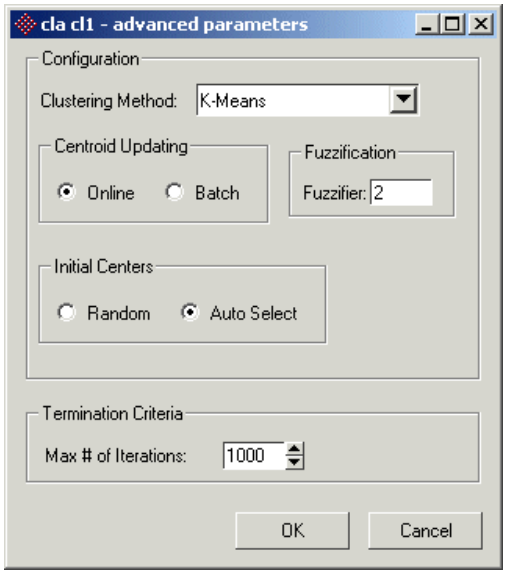


Figure 8. 2: Advanced configuration of the Cluster Analysis dialog

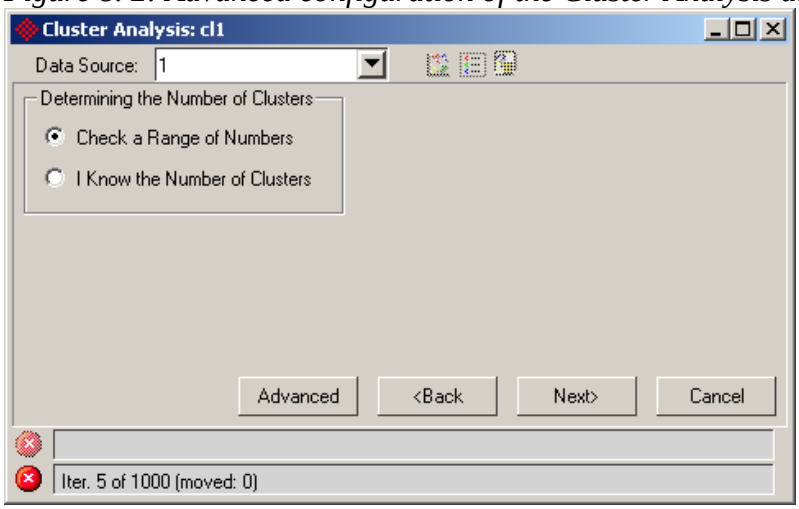


Figure 8. 3: Determining the number of clusters

After choosing the range of numbers, click **Next** (Figure 8. 4), and a curve set will pop up plotting the Davies-Bouldin score for each number of clusters (Figure 8. 5).

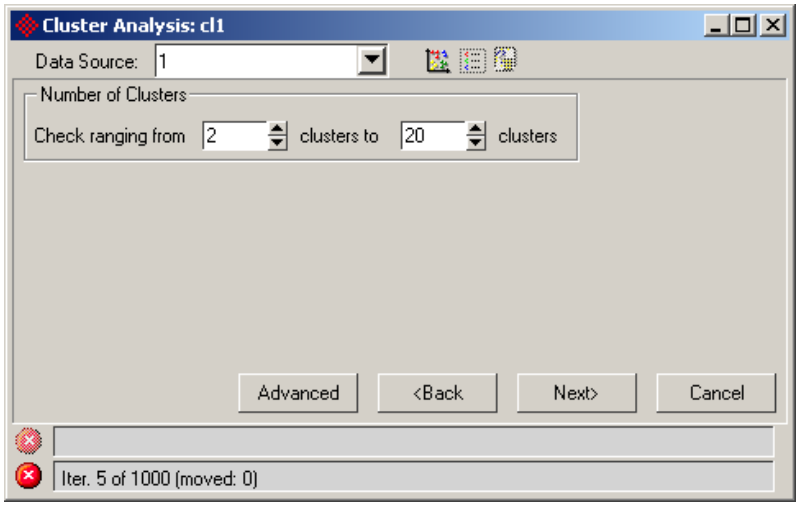


Figure 8. 4: Checking a range of numbers

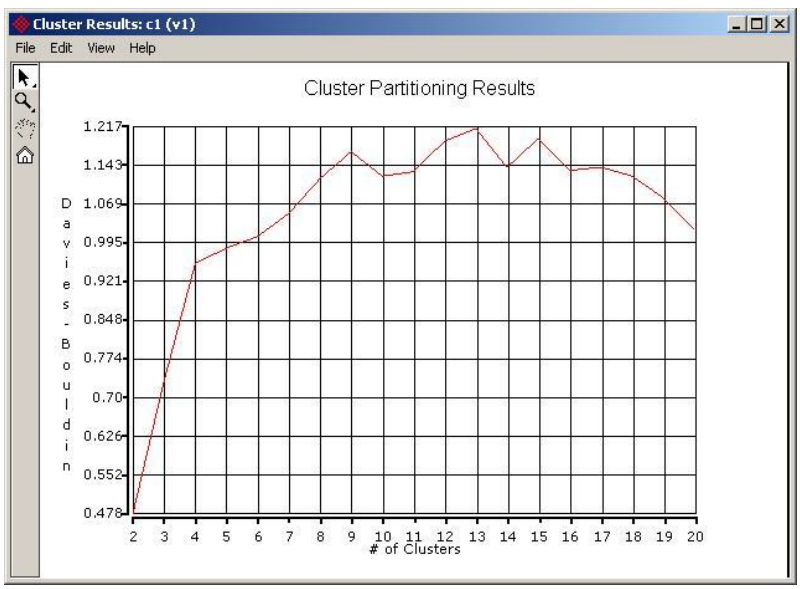


Figure 8. 5: Viewing the cluster partitioning results

If a range of numbers has been tested, the spin box will be set to the number of clusters with the best (lowest) Davies-Bouldin score.

Confirm the number of clusters and click **Next** (*Figure 8. 6*).

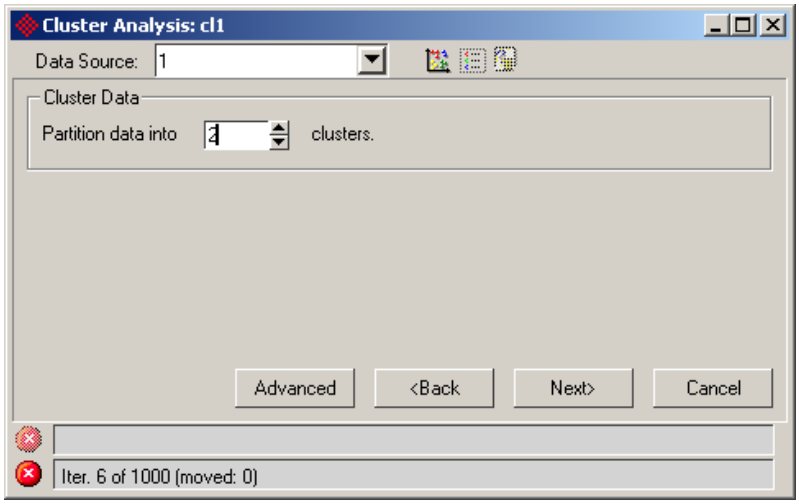


Figure 8. 6: Configuring the cluster data

The results may be dumped to the spreadsheet by clicking the appropriate button or by clicking **Next** with **Dump to Spreadsheet** checked (Figure 8. 7).

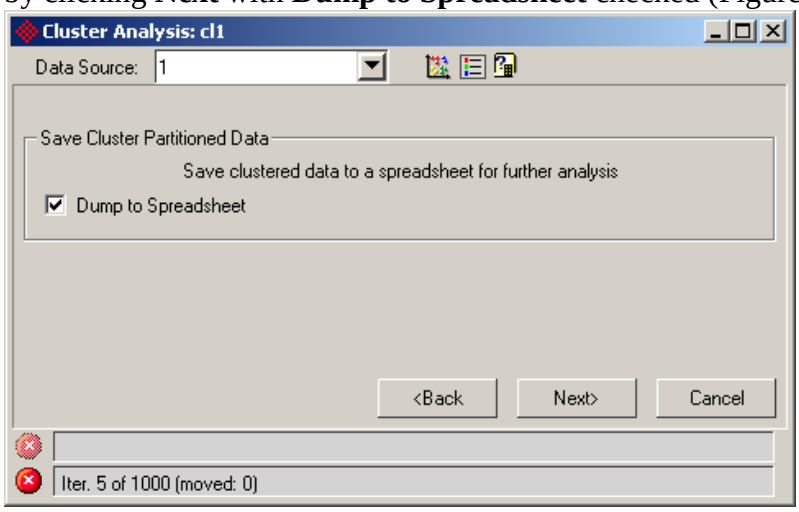


Figure 8. 7: Clustering done

Clicking on the **Configure Scatter Plot** action button (Figure 8. 8) will display a dialog that will create a scatter plot to show the clustered data (Figure 8. 9).

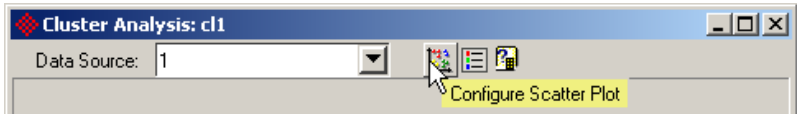


Figure 8. 8: Creating a scatter plot based on the clustering

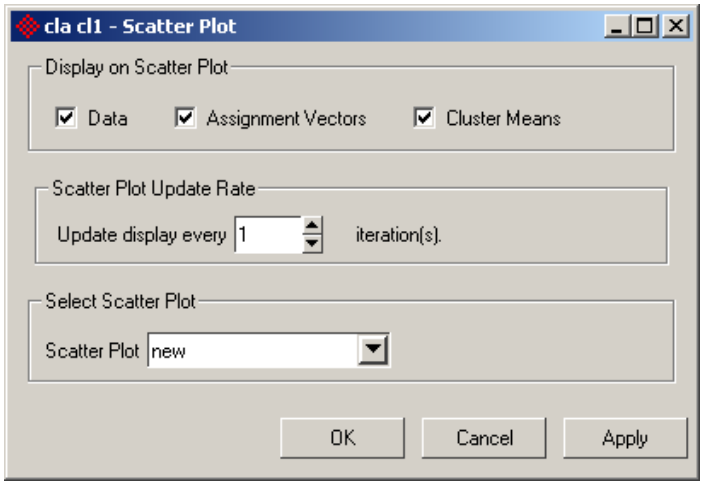


Figure 8. 9: Configuring the scatter plot

In addition to the normal scatter plot configuration options, this scatter plot can be colored by cluster assignment (Figure 8. 10).

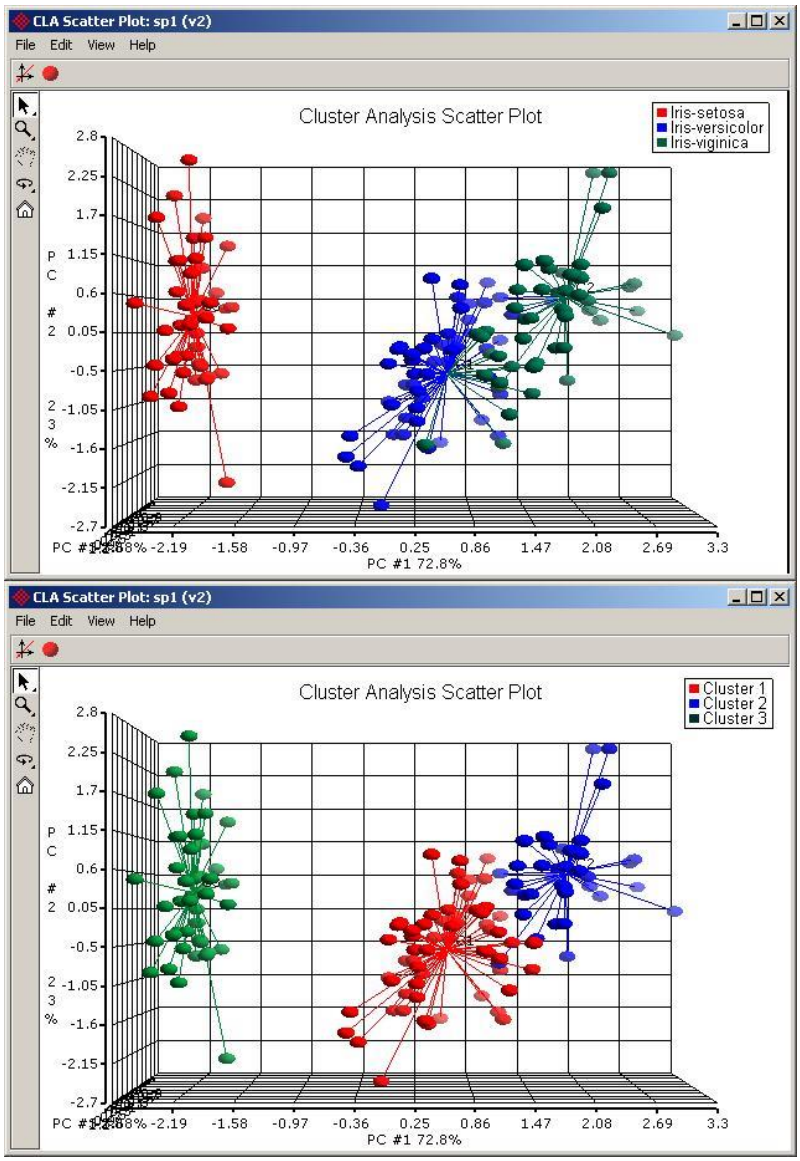


Figure 8. 10: Viewing the cluster analysis scatter plot colored by Type (top) and Cluster (bottom)

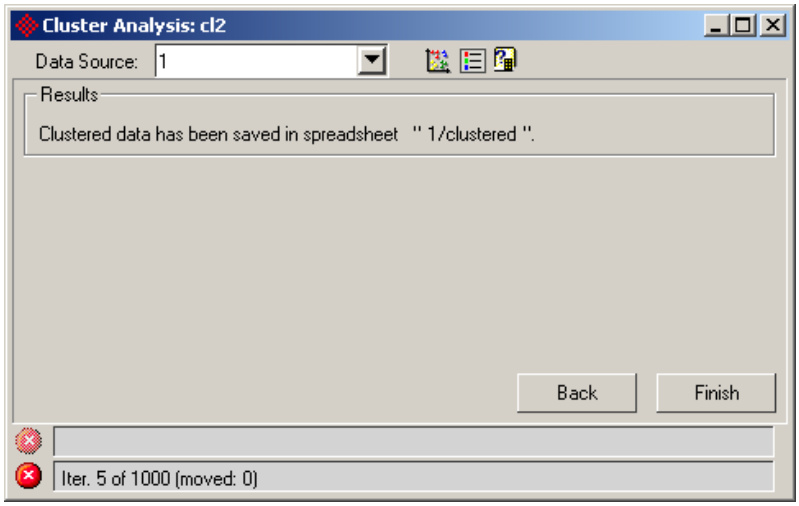


Figure 8. 11: Viewing the Results panel of the cluster analysis

The added cluster column is a good candidate for a group profile (**View > Profiles > Group Profile...** from the Partek main window) (Figure 8. 12).

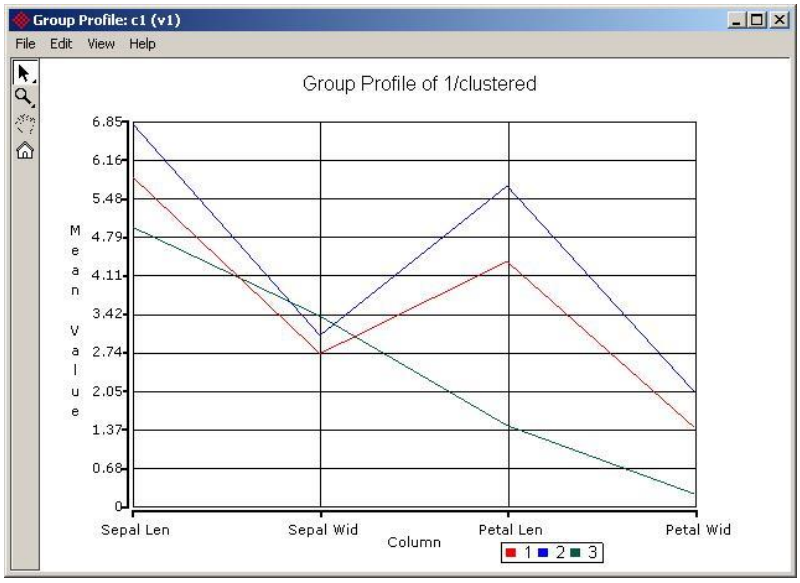


Figure 8. 12: Viewing the Group Profile of the clustered column

This plot reveals the mean value of each cluster across all columns and provides an easy way to select all members of a given cluster. The mouse-over of a curve will reveal how many rows are in the cluster.

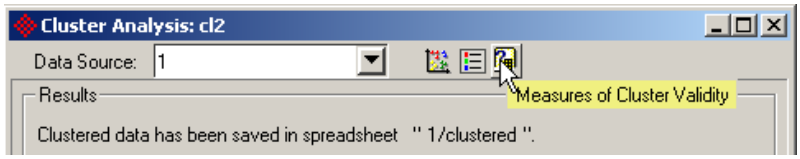


Figure 8. 13: Measures of Cluster Validity accelerator button

The validity of the resulting clusters may be verified by using several measures. The measures using internal criteria are *Davies-Bouldin* and *Modified Hubert*. The measures using external criteria are *Rand* and *Jaccard*. Click **Compute** to calculate the selected measures (Figure 8. 14).

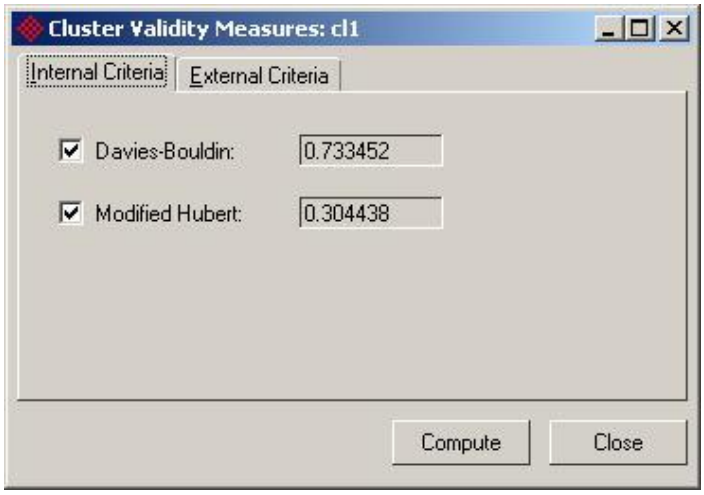


Figure 8. 14: Configuring the Measures of Cluster Validity dialog

References

Cho, R. J., et. al. "A Genome-Wide Transcriptional Analysis of the Mitotic Cell Cycle". Molecular Cell **2**, 65-73 (1998). http://genomics.stanford.edu/yeast_cell_cycle/

Descriptive Statistics, Correlation, & Measures of Similarity & Dissimilarity

Introduction

The descriptive and correlative statistics tests in Partek are explained in this chapter. Descriptive statistics in Partek include calculating column statistics, row statistics, and grand statistics, and finding zero variance variables. Correlative statistics in Partek include associative measurements, finding correlated variables, many to one, similarity matrix, dissimilarity matrix, and finding duplicate patterns.

Descriptive (Univariate) Statistics

Explanation of Descriptive Statistics

Each descriptive statistic available in Partek will be defined below. For this discussion, $\{x_1,...,x_n\}$ refers to an array of numbers (e.g., a column or a row in a spreadsheet).

Avg. Dev.	Mean	Skewness
CV (%)	Median	Std. Dev.
Geometric Mean	Min	Sum
Harmonic Mean	Norm	Trimmed Mean
Kurtosis	Range	Variance
Max	Root MS	Winsorized Mean

Table 9. 1: Descriptive statistics available in Partek

For explanations of Median Polish, see Tukey, J. (1977) *Exploratory Data Analysis*. Addison-Wesley, Reading, MA.

For explanations of Tukey's Bi-weight, see the Statistical Algorithms Description Document, Affymetrix, Inc. Technical documentation – white papers, http://www.affymetrix.com/support/technical/whitepapers.affx (2002).

Measures of Location

Simple measures of the "middle" and "extent" of the distribution include the *mean*, *median*, and *sum*:

- Mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

- Harmonic Mean $H = \dfrac{n}{\sum_{i=1}^{n} \dfrac{1}{x_i}}$

- Geometric Mean $G = \sqrt[n]{\prod_{i=1}^{n} x_i}$

- Root Mean Square $RMS = \sqrt{\dfrac{\sum_{i=1}^{n} x_i^2}{n}}$

- Median, for computation of the median, sort the vector $\{x_1,...,x_n\}$ after which:

$Median = x_{\frac{n+1}{2}}$, n odd or

$Median = \frac{1}{2} x_{\frac{n}{2}} + \frac{1}{2} x_{\frac{n}{2}+1}$, n even

- Sum $Sum = \sum_{i=1}^{n} x_i$

Measures of Dispersion

- Min $x_{\min} = \min_i (x_i)$

- Max $x_{\min} = \max_i (x_i)$

- Range $range = x_{\max} - x_{\min}$

Variance and *standard deviation* are closely related and are common measures of the "variability" of a set of measurements. The formulas for each statistic depend on whether the actual mean of the data is known or whether the mean is an estimate. If the data is comprised of the entire population, then the mean is known exactly. Otherwise, we have to estimate the mean from a sample of the entire population.

Note: In Partek, the computation of variance and standard deviation depend on a global parameter that decides whether to use *population* or *sample* statistics. By default, *sample statistics* are computed. Note also that calculations based on the variance or standard deviation are also affected by this global parameter. The global parameter can be set on the *Other Settings* page of the **Edit > Preferences.**

- Population Variance $\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$

- Sample Variance $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
- Population Standard Deviation $\sigma = \sqrt{\sigma^2}$
- Sample Standard Deviation $s = \sqrt{s^2}$
- Average Deviation $avgdev = \frac{1}{n}\sum_{i=1}^{n}| x_i - \bar{x} |$
- Coefficient of Variation (CV) $cv = \frac{s}{\bar{x}}$

Measures of Distribution

Skewness and *kurtosis* measure how much a distribution varies from a normal distribution. Skewness measures symmetry about the mean of a distribution. If a distribution is symmetric about its mean, the skewness is equal to zero. If the distribution (histogram) of the variable has a longer tail on the left than on the right, it has a negative skewness. If the distribution of the variable has a longer tail on the right than on the left, it has a positive skewness. The kurtosis is a measure of a distribution's peak relative to a normal distribution. A distribution with a point (like the tip of an arrow head) will have a positive kurtosis where as a distribution, which is somewhat flat (like the profile of a thimble) will have a negative kurtosis.

- Population Skewness $skewness = \frac{1}{n}\sum_{i=1}^{n}[\frac{x_i - \bar{x}}{\sigma}]^3$
- Sample Skewness $skewness = \frac{n}{(n-1)(n-2)}\sum_{i=1}^{n}[\frac{x_i - \bar{x}}{s}]^3$
- Population Kurtosis $kurtosis = \frac{1}{n}\sum_{i=1}^{n}[\frac{x_i - \bar{x}}{\sigma}]^4 - 3$
- Sample Kurtosis

$$kurtosis = \frac{n(n+1)}{(n-1)(n-2)(n-3)}\sum_{i=1}^{n}[\frac{x_i - \bar{x}}{s}]^4 - 3\frac{(n-1)(n-1)}{(n-2)(n-3)}$$

Column Statistics

- To compute descriptive statistics on columns, select **Stat** > **Descriptive** > **Column Statistics** to invoke the *Column Statistics* dialog (Figure 9. 1)

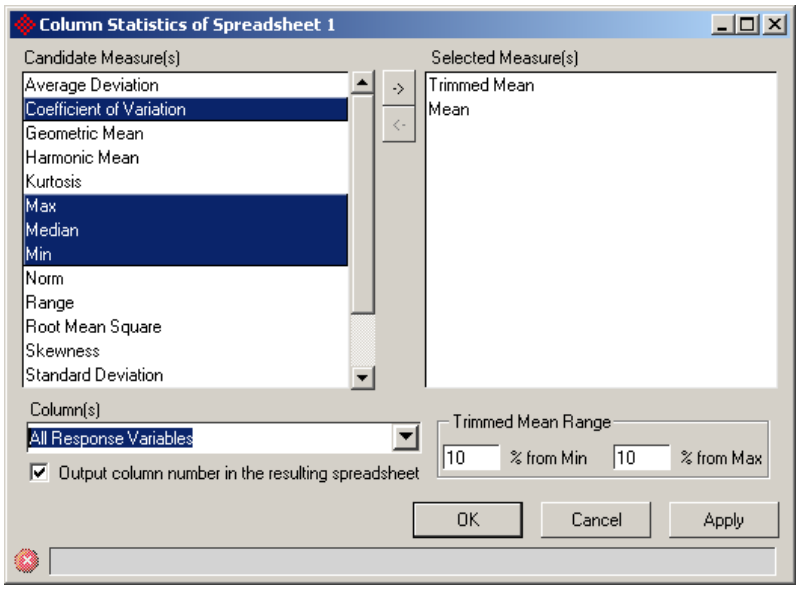


Figure 9. 1: Configuring the Column Statistics dialog

The descriptive statistics that can be computed on columns and are listed on the *Candidate Measure(s)* panel on the left side of the dialog box.

- Double click on a measure to copy it to the *Selected Measure(s)* panel

To select multiple items, click and drag or hold the **<Ctrl>** key down and left click. Click the **->** button to move the selected items to the *Selected Measure(s)* panel. The dialog in Figure 9. 1 shows the **Trimmed Mean** and **Mean** as selected measures. The **CV, Max, Median,** and **Min** have been selected in the *Candidate Measure(s)* panel and the *% from Min* and *% from Max* entries have been enabled because the **Trimmed Mean** is one of the two statistics shown in the *Selected Measure(s)* panel. You can compute the measurements on one column at a time by selecting the column from the *Column(s)* drop-down list (Figure 9. 2)

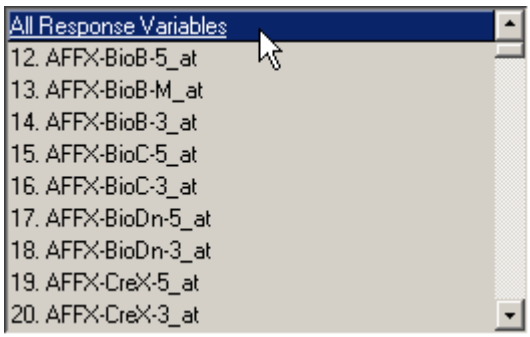


Figure 9. 2: Column(s) drop down list

You can also compute the statistics on all numeric response variables by choosing **All Response Variables** in the drop-down list. When you choose **All Variables**, the computation will be performed on all numeric variables in the spreadsheet regardless of the column attribute.

Trimmed Mean

The trimmed mean is computed by first removing a certain percentage of the lowest and highest values and then computing the mean. Partek allows you to specify different percentages for the *Min* and *Max* value. To compute the *Trimmed Mean*, copy the **Trimmed Mean** list item to the *Selected Measure(s)* panel and specify the percentages from the low and high ends of the data that will be used to trim the data.

Winsorized Mean

The winsorized mean is computed by first setting a certain percentage of the lowest values to the closest value above the certain percentage. The same action is done for the highest values, but instead the closest value below the certain percentage will be used, and then the mean will be computed. Partek allows you to specify different percentages for the *Min* and *Max* value. To compute the *Winsorized Mean*, copy the **Winsorized Mean** list item to the *Selected Measure(s)* panel and specify the percentages from the low and high ends of the data that will be used to trim the data.

For example, to take the Winsorized mean of the following data {1,2,3,4,5,6,7,8,9,10} with Min being 10% and Max being 20%, would take the mean of {2,2,3,4,5,6,7,8,8,8}.

Computing the Statistics

Click **OK** or **Apply** to compute the selected statistics. If a single column is selected, the results will be displayed in an HTML report, otherwise the results are displayed in a child spreadsheet (Figure 9. 3).

Descriptive:column

	1.Column #	2.Column ID	3.Avg Dev.	4.Coefficient of	5.Geome
1.	11	200000_s_at	136.335680	0.235196	678.1974
2.	12	200001_at	366.593920	0.439789	882.601C
3.	13	200002_at	467.733760	0.166286	3360.404
4.	14	200003_s_at	619.310080	0.184289	4213.417
5.	15	200004_at	533.920	0.206746	3288.161
6.	16	200005_at	133.158720	0.265426	562.8743

Figure 9. 3: Viewing the result spreadsheet of column descriptive statistics on numeric variables

In the results spreadsheet, each row represents a column in the parent spreadsheet. The first two columns contain the column number and column name from the original (parent) spreadsheet. The selected statistics begin in column 3. When you right click on any row header, you can invoke an HTML report of the measurements of that specific variable (Figure 9. 4, Figure 9. 5).

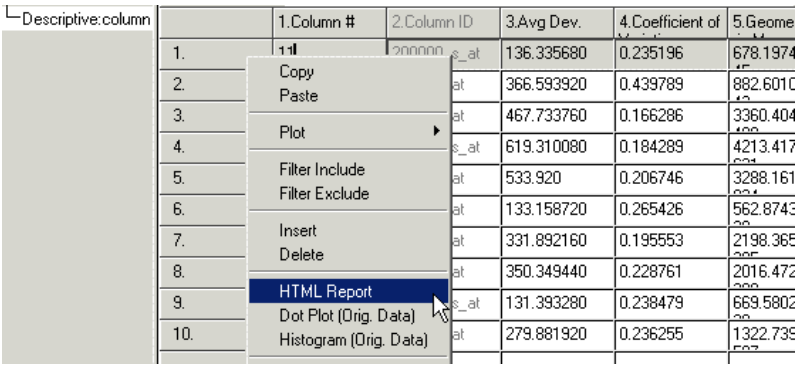


Figure 9. 4: Invoking an HTML report of descriptive statistics on a single variable

Descriptive Statistics of
200000_s_at

Avg Dev.	136.335680
Coefficient of Variation	0.235196
Geometric Mean	678.197445
Harmonic Mean	658.021514
Kurtosis	-0.861518
Max	1017.10

Figure 9. 5: HTML report of descriptive statistics on a single variable

Row Statistics

Row statistics can be computed by using the *Row Statistics* dialog. To invoke the dialog, select **Stat > Descriptive > Row Statistics (**Figure 9. 6). Instructions for selecting *Candidate Measure(s)* are described in the **Column Statistics** section of this document and are not repeated here.

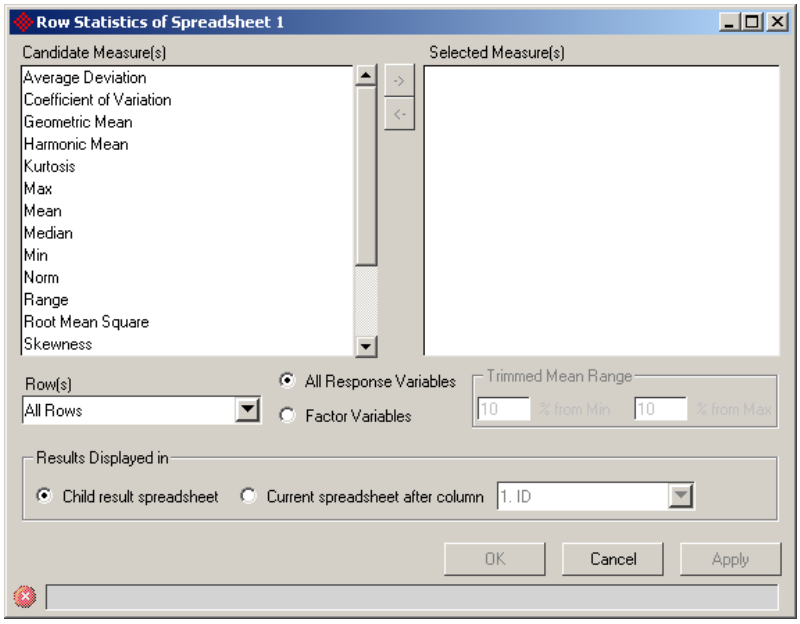


Figure 9. 6: Configuring the Row Statistics dialog

You can compute the statistics on one row at a time by selecting the row from the *Row(s)* drop-down list or you can compute statistics on all rows by choosing **All Rows** in the drop-down list. There are two types of numeric variables in Partek, but for most applications, you will want to compute the statistics on the *All Response Variables* (Figure 9. 7).

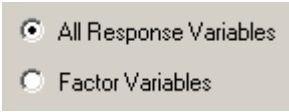


Figure 9. 7: Selecting the type of numeric variable for row statistics

Computing the Statistics

When the computation is performed on *All Rows*, you can choose to create a new spreadsheet to store the results or add the statistics to the current spreadsheet. When adding the statistics to the existing spreadsheet, specify where the new columns will be added (Figure 9. 8).

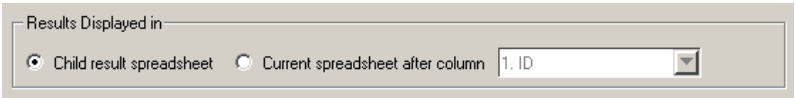


Figure 9. 8: Selecting destination spreadsheet for row statistics

Click **OK** or **Apply** to perform the selected computations. The results are displayed either in a child spreadsheet (Figure 9. 9) or added as factor numeric variables in the current spreadsheet.

Descriptive:row

	1.ID	2.Source	3.Type	4.Avg Dev.	5.Coefficient
1.	01	heart	TS21	1062.040	0.698175
2.	02	cerebrum	TS21	1099.460	0.648776
3.	03	cerebellum	TS21	998.180	0.676345
4.	04	cerebrum	TS21	1249.0920	0.722503
5.	05	cerebellum	TS21	1127.760	0.805396
6.	06	heart	control	1133.180	0.774371
7.	07	cerebellum	control	991.380	0.725824
8.	08	cerebrum	control	1237.0480	0.723722
9.	09	cerebrum	control	1440.7520	0.803295
10.	10	cerebellum	control	1180.730	0.722792

Figure 9. 9: Viewing the row statistics that are displayed in a separate spreadsheet

Creating a Separate Results Spreadsheet

In the results spreadsheet, each row corresponds to the same row in the parent spreadsheet; it contains all the non-numeric columns of the parent spreadsheet and a column for each of the selected statistics. In Figure 9. 9, columns 1-3 (*ID, Source,* and *Type)* are the non-numeric columns of the original data and the selected statistics for each row are stored in subsequent columns. When you right click on any row header, you can also get an HTML report for each selected row(s) (Figure 9. 10, Figure 9. 11).

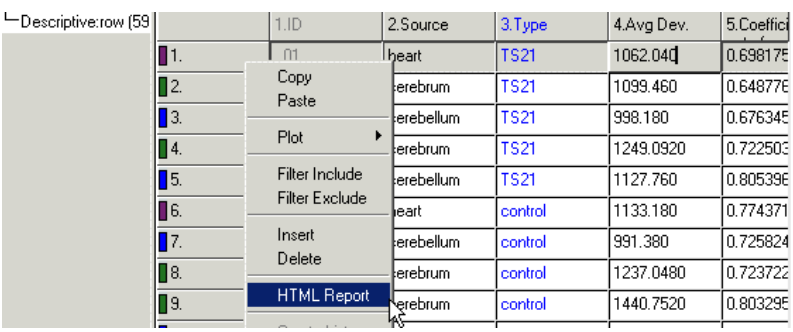


Figure 9. 10: Invoking an HTML report of descriptive statistics on a single row

Descriptive Statistics of
Row 1

Avg Dev.	1062.040
Coefficient of Variation	0.698175
Geometric Mean	1514.207243
Harmonic Mean	1179.739425
Kurtosis	-1.192714
Max	4355.10

Figure 9. 11: Viewing the HTML report of descriptive statistics on a single row

Grand Statistics

Statistics can be computed for the entire spreadsheet by using the *Grand Statistics* dialog.

- Select **Stat** > **Descriptive** > **Grand Statistics.** The *Grand Statistics* dialog will appear (shown in Figure 9. 12)
- Click **Compute** to compute the statistics

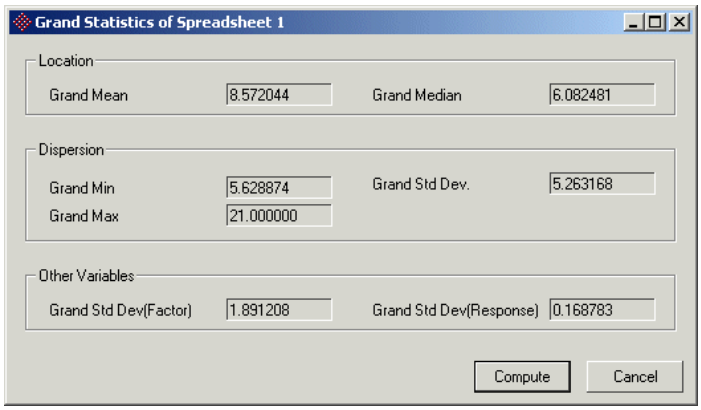


Figure 9. 12: Grand Statistics dialog

Zero Variance Variables

If a variable has no variance, its value is constant across all samples of data. It therefore offers no real information and will cause computational problems when, for example, computing the correlations between columns. Check the data for zero variance variables and consider removing them from the spreadsheet. Click the **Delete** button to delete the zero variance variables or the **Filter** button to filter exclude the variables. Zero Variance Variables can be found by going to **Stat** > **Descriptive** > **Find Zero Variance Variables**. The dialog, shown in Figure 9. 13, will appear.

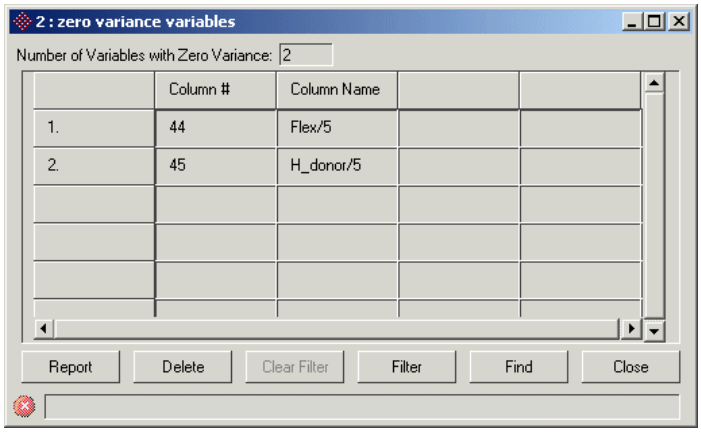


Figure 9. 13: Viewing the Zero Variance Variables dialog

Correlative (Bivariate) Statistics

Measures of Association

- Select **Stat** > **Correlate** > **Associative Measurements** to invoke the *Measures of Association* dialog, shown in 9. 14
- Click **Compute** to get the results

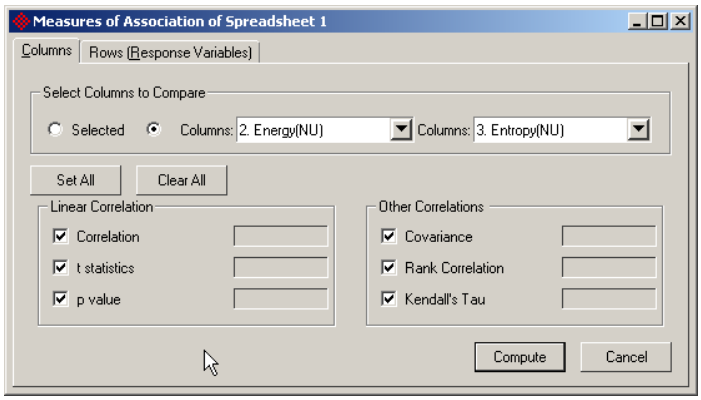


Figure 9. 14: Configuring the Measures of Association dialog

- Covariance $$r_{xy} = \sum_i (x_i - \bar{x})(y_i - \bar{y})$$

- Linear Correlation (Pearson's r) $$r_{xy} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (y_i - \bar{y})^2}}$$

- Non-Linear Correlation (Spearman's Rank coefficient)

$$r_s = \frac{\sum_i (R_i - \bar{R})(S_i - \bar{S})}{\sqrt{\sum_i (R_i - \bar{R})^2}\sqrt{\sum_i (S_i - \bar{S})^2}}$$

where R_i is the rank of x_i in the vector x, S_i is the rank of y_i in the vector y.

- Kendall's Tau

$$\tau = \frac{concordant - discordant}{\sqrt{concordant + discordant + extraY}\sqrt{concordant + discordant + extraX}}$$

Finding Correlated Variables

To check for highly correlated variables in Partek, select **Stat** > **Correlate** > **Find Correlated Variables** to invoke the dialog of Figure 9. 15. For example, the default *Absolute value of r* is **0.9**, therefore, it will include all variable pairs that have a linear correlation greater than or equal to **0.9** or less than or equal to **-0.9.** Click **Compute** to compute the correlations and click **Report** to get the result in HTML format. You can reduce the dimensionality by filtering or deleting redundant variables. Clicking **Delete** or **Filter** in Figure 9. 15 will filter delete or filter out the variables listed in the second *Column #* list, in this case columns *10*, *13*, *15* and *31*.

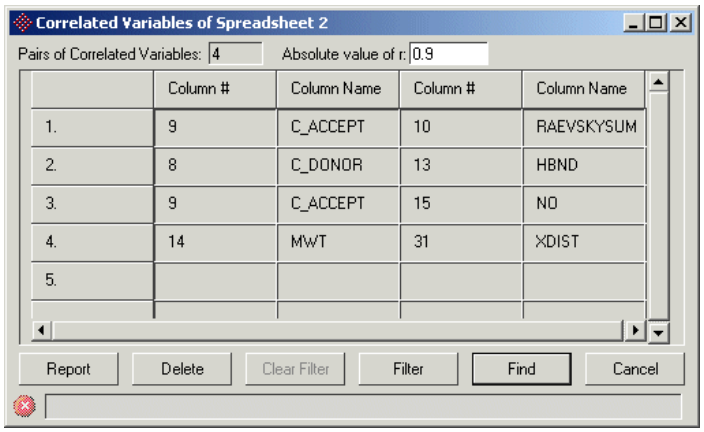


Figure 9. 15: Viewing the Correlated Variables dialog

Correlation of All Numeric Variables to One Numeric Variable

To invoke the *Correlation* dialog shown in Figure 9. 16, select **Stat** > **Correlate** > **Many to One**.

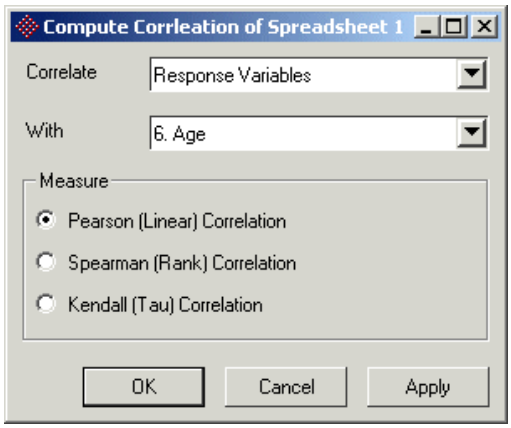


Figure 9. 16: Configuring the Compute Correlation dialog

The *Correlate* drop-down list has three options to group the numeric variables, *All Variables*, *Response Variables*, and *Factor Variables*. Select one of the groups to correlate with one variable from the second drop-down list labeled *With*, which

includes all the numeric variables in the spreadsheet. Make a selection on the measures and click **OK** or **Apply**. The correlation will be computed and the result will be stored in a child spreadsheet (Figure 9. 17).

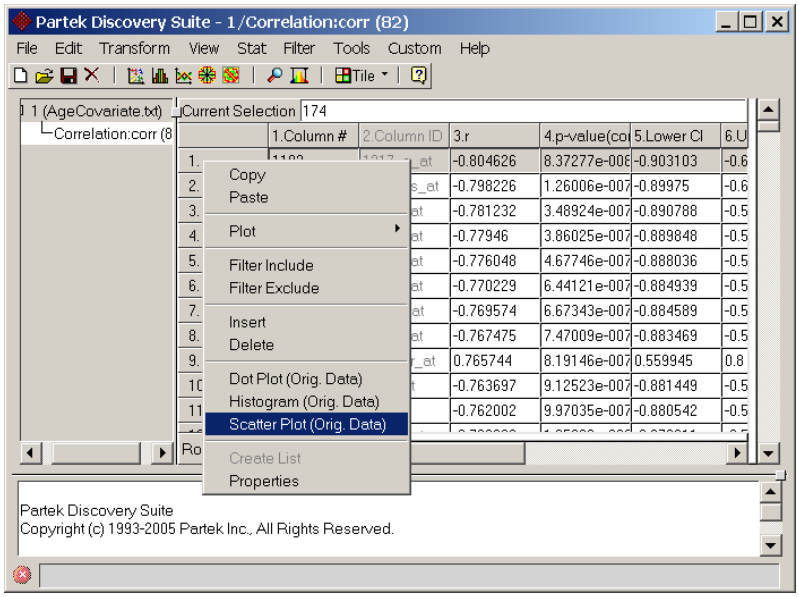


Figure 9. 17: Viewing the spreadsheet containing the Many to One correlation results

The results spreadsheet displays *r*, *p-value*, and *N*. *r* is the Coefficient of Correlation between the selected variable and each of the numeric variables. The coefficient value is between –1 and 1. The larger the absolute value of r, the stronger the correlation. The sign of the coefficient represents whether the correlation is positive or negative. The p-value is calculated based on the correlation *r* and the sample size *N*. If the p-value is infinitesimal for perfect positive and negative correlation, Partek will report a p-value of 0. N is the sample size that corresponds to the number of rows in the original spreadsheet. In the result of *Pearson (Linear) Correlation*, Partek also supplies the lower and upper 95% confidence interval.

Missing Data

If any of the rows or columns contains missing data, the correlation is computed using the available pairs of data and the sample size is adjusted accordingly. For the data in Table 9. 2, the correlation would be computed using rows **1, 4,** & **5** with a sample size of 3.

ID	Column 1	Column 2
1	9	10
2	5	?
3	?	7
4	3	3

5	14	10

Table 9. 2: Correlating columns with missing values

Right-clicking on a row label will invoke a pop-up menu with further options. Select **Scatter Plot (Orig. Data)** to plot the two variables. The plot can be configured to show the *regression line* and *confidence interval.*

Similarity Matrix

The similarity matrix is computed on the columns in the spreadsheet, but should be used with caution for very high dimensional data such as microarray data. Select **Stat > Correlate > Similarity Matrix** to invoke the *Variable Similarity Matrix* dialog (Figure 9. 18). Select the *Similarity Measure* and click **OK** to get the results in a child spreadsheet.

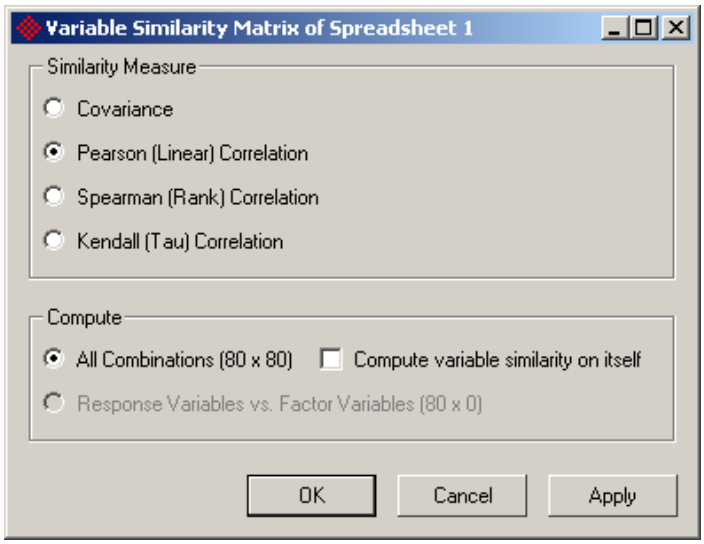


Figure 9. 18: Configuring the Variable Similarity Matrix dialog

Dissimilarity Matrix

The dissimilarity matrix is computed on the columns in the spreadsheet, but should be used with caution for very high dimensional data such as microarray data. Select **Stat > Correlate > Dissimilarity Matrix** to invoke the *Variable Dissimilarity Matrix* dialog (Figure 9. 19). Configure the *Distance Function* and click **OK** to get the result in a child spreadsheet.

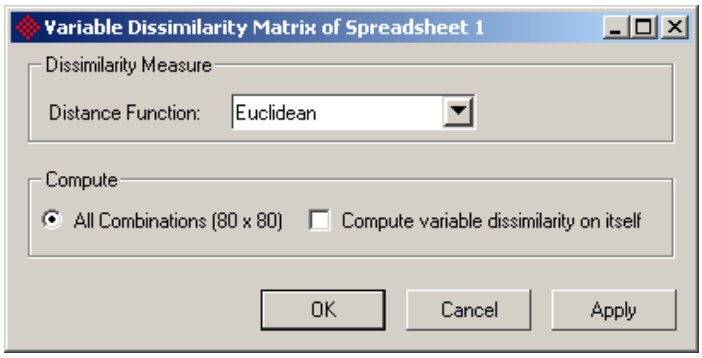


Figure 9. 19: Configuring the Variable Dissimilarity Matrix dialog

Duplicate Pattern

Select **Stat > Correlate > Find Duplicate Pattern** to invoke the *Find Duplicates* dialog (Figure 9. 20). Click **Compute** to get pairs of rows that are the same on all the numeric variables. The result is displayed in the command window. If the duplicates are from different subgroups of the class variable, the conflicts are also displayed in the dialog box. Click **Report** to get the result in HTML format.

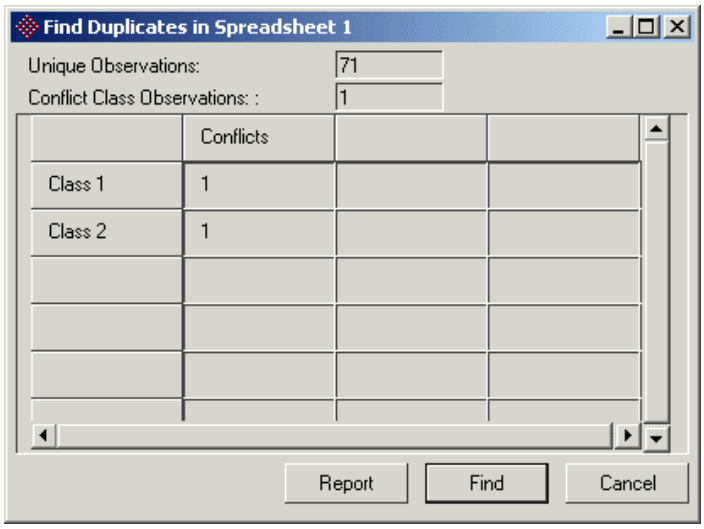


Figure 9. 20: Viewing the Find Duplicates dialog

Measures of Distance & Dissimilarity

Introduction

This section presents measures of distance and dissimilarity that can be used in analysis and modeling modules, which directly make use of the dissimilarity between objects such as the *Multi-Prototype Classifier* (MPC), *Cluster Analysis* (CLA), and *Multidimensional Scaling* (MDS).

Similarity measures increase as the similarity between objects increase, while *dissimilarity measures* decrease as the similarity increases. Since many pattern recognition algorithms traditionally use distance metrics (which measure dissimilarity between objects) Partek converts similarity measures into dissimilarity measures so that they can be interchanged with distance metrics without having to modify the algorithms that use them.

Distance Metrics

Distance Metrics tell how far apart two vectors are in n-dimensional space. Formal definitions of distance functions and distance metrics can be found in a variety of texts on cluster analysis and topology (including Anderberg 1973, Munkres 1975, Royden 1988, and Spath 1980).

Let x and y denote two real vectors $(x_1,...,x_n)^T$ and $(y_1,...,y_n)^T$ (Spath 1980). A real-valued function d(x, y) is said to be a distance function if, and only if, the following three conditions are satisfied:

$$d(x,y) \geq d_0$$
$$d(x,x) = d_0$$
$$d(x,y) = d(y,x)$$

The distance function d(x, y) can further be considered a *metric* **if and only if** in addition to the above three conditions, the following two conditions are also satisfied:

$d(x,y) = d_0$ if and only if $x = y$

$d(x,y) \leq d(x,z) + d(z,y)$ for all $x, y, z \in R^n$ where R^n is n-dimensional Euclidean space and d_0 is an arbitrary real number (usually 0).

Euclidean Distance

The Euclidean distance between vectors x and y is given by

$$d_{euc}(x,y) = \sqrt{\sum_i (x_i - y_i)^2} \ .$$

Euclidean distance is the default measure used in Partek. The Euclidean distance satisfies all conditions of a metric.

Average Euclidean Distance

The average Euclidean distance is the same as the Euclidean distance except that it is normalized by dividing by $\sqrt{n}$:

$$d_{avgEuc}(x,y) = \sqrt{\sum_i \frac{(x_i - y_i)^2}{n}}$$

Because d_{avgEuc} is a scaled version of d_{euc} it will give the same results as d_{euc} in many algorithms. The average Euclidean distance is preferred to the Euclidean distance when the data contains missing values because it does not tend to grow larger as the vector length grows and is better suited to measuring the distance between vectors, which may contain missing values (this assumes that the data has been standardized). The average Euclidean distance satisfies all conditions of a metric.

Squared Euclidean Distance

The squared Euclidean distance between vectors x and y is given by

$$d_{sqrEuc}(x, y) = \sum_i (x_i - y_i)^2$$

It is nearly identical to Euclidean distance. However since it does not compute square root, squared Euclidean is faster than Euclidean distance.

Minkowski Distance

The Minkowski distance is defined as the pth root of the sum of the absolute value of the differences of the vector elements raised to the power p and is therefore a generalization of the Euclidean distance:

$$d_{\min k}(x, y) = \sqrt[p]{\sum_i | x_i - y_i |^p}$$

Average Minkowski Distance

Since the Minkowski distance is a generalization of the Euclidean distance, it is natural that you also provide an average Minkowski distance for the same reasons that you include the average Euclidean distance. The average Minkowski distance is the same as the Minkowski distance except that it is normalized by dividing by $\sqrt[p]{n}$:

$$d_{avgMink}(x, y) = \sqrt[p]{\sum_i \frac{| x_i - y_i |^p}{n}}$$

Mahalanobis Distance

The Mahalanobis distance is used when you want to compensate for the fact that different variables may be measured on different scales:

$$d_{mahal}(x, y) = \sqrt{(x - y)^T C^{-1} (x - y)}$$

where C is the covariance matrix of the entire data set. When C^{-1} is the identity matrix, this metric is equivalent to the Euclidean distance. It should also be noted that models that make use of this distance must save C^{-1} as part of the saved model.

Maximum Value Distance

The maximum value distance metric can be used when you only care how close two vectors are at their farthest point. For example, it can be used to measure the maximum deviation between two observations of the same phenomena.

$$d_{\max}(x, y) = \max_i | x_i - y_i |$$

Minimum Value Distance

The minimum value distance function is used when you only care how close two vectors are at their closest point. For example, suppose two vectors contain measurements of altitude of the ground and a high power line. In this case you may only care how close the high power line is to the ground at its closest point.

$$d_{\min}(x, y) = \min_i | x_i - y_i |$$

Absolute Value Distance

Also known as the taxi cab distance, the absolute value distance metric is a special case of the Minkowski distance with *p*=1:

$$d_{abs}(x, y) = \sum_i | x_i - y_i |$$

You can compute an average absolute value distance by using the average Minkowski distance metric and specifying *p*=1.

Tanimoto Distance

The Tanimoto distance is used to see how similar two chemicals are. It does this by counting the number of chemical substructures or chemical groups they have in common:

$$d(x, y) = \frac{x^t y}{x^t x + y^t y - x^t y}$$

Where $x^t y$ is number of attributes possessed by both x and y

The distance is given by the ratio between the number of groups that are occur in both, divided by this plus the number in only one, plus the number only in the other. The number that occurs in neither is ignored.

Measures of Dissimilarity

In addition to the distance metrics described above, Partek provides measures of dissimilarity. These measures tell how similar the shapes of the data profiles are. The first three are simple transformations of three measures of correlation between the vectors. The cosine dissimilarity is the cosine of the angle between the two vectors. Finally, other measures that were specifically designed to measure dissimilarity are presented.

Pearson's Dissimilarity

Pearson's dissimilarity is a transformation of the linear (Pearson's *r*) correlation between two vectors. When used as a dissimilarity measure, it is rescaled to the interval [0,1] with 0 indicating perfect similarity (perfect positive correlation) and one indicating perfect dissimilarity (perfect negative correlation).

$$d_r(x, y) = \frac{(1 - r)}{2}$$

where r is the linear correlation.

Pearson's Absolute Value Dissimilarity

Pearson's Absolute Value dissimilarity is a slight modification of Pearson's dissimilarity. It is rescaled to the interval [0,1] with 0 indicating either maximum similarity or dissimilarity and 1 indicating uncorrelated.

$$d_{rabs}(x, y) = 1 - |r|$$

where r is the linear correlation.

Rank (Spearman) Dissimilarity

Rank dissimilarity is a transformation of Spearman's non-parametric r_s correlation between two vectors and is called for when the data is ordinal. When used as a dissimilarity measure, it is rescaled to the interval [0,1] with 0 indicating perfect similarity (perfect positive correlation) and 1 indicating perfect dissimilarity (perfect negative correlation).

$$d_{r_s}(x, y) = \frac{1 - r_s(x, y)}{2}$$

where r_s is Spearman's rank order coefficient.

Rank (Spearman) Absolute Value Dissimilarity

Rank absolute value dissimilarity is a slight modification of Rank dissimilarity. When used as a dissimilarity measure, it is rescaled to the interval [0,1] with 0 indicating either maximum similarity or dissimilarity and 1 indicating uncorrelated.

$$d_{rabs_s}(x, y) = 1 - |r_s(x, y)|$$

where r_s is Spearman's rank order coefficient.

Kendall's Dissimilarity

Kendall's dissimilarity is the third dissimilarity metric based on the correlation between the vectors and is computed by:

$$d_{\tau}(x, y) = \frac{(1 - \tau)}{2}$$

where τ is Kendall's Tau correlation. Like the two previous measure, it is rescaled to the interval [0,1] with 0 indicating perfect similarity (perfect positive correlation) and one indicating perfect dissimilarity (perfect negative correlation).

Kendall's Absolute Value Dissimilarity

Kendall's absolute value dissimilarity is a slight modification of Kendall's dissimilarity. When used as a dissimilarity measure, it is rescaled to the interval [0,1] with 0 indicating either maximum similarity or dissimilarity and 1 indicating uncorrelated.

$$d_{\tau abs}(x, y) = 1 - |\tau|$$

where τ is Kendall's Tau correlation.

Coefficient of Shape Difference

Created by Penrose, the coefficient of shape difference is defined in the range [0, ∞] and is a function of the average Euclidean distance. The shape difference ignores

additive displacement and therefore gives similar results to the cosine dissimilarity and measures based on correlation.

$$d_{shape}(x,y) = \sqrt{\frac{n}{n-1}(d_{avgEuc}(x,y)^2 - q(x,y)^2)}$$

where $d_{avgEuc}(x,y)$ is *Average Euclidean Distance* and q(x,y) is given by:

$$q(x,y) = \frac{\sum_i x_i - \sum_i y_i}{n}$$

Cosine Dissimilarity

The cosine dissimilarity is based on the cosine coefficient cos(x,y) (defined in the interval [-1,1]):

$$\cos(x,y) = \frac{\sum_i x_i y_i}{\sqrt{\sum_i x_i^2}\sqrt{\sum_i y_i^1}}$$

The cosine coefficient measures the cosine of the angle formed by the vectors x and y. Convert cos(x,y) to a measure of dissimilarity in the interval [0,1] as follows:

$$d_{\cos ine}(x,y) = \frac{(1-\cos(x,y))}{2}$$

Canberra Metric

The Canberra metric is a dissimilarity measure defined on the interval [0,1] and satisfies all four conditions of a metric.

$$d_{canberra}(x,y) = \frac{1}{n}\sum_i \frac{|x_i - y_i|}{(x_i + y_i)}$$

Bray-Curtis Coefficient

The Bray-Curtis coefficient is a dissimilarity measure defined on the interval [0,1] and satisfies all four conditions of a metric.

$$d_{bc}(x,y) = \frac{\sum_i |x_i - y_i|}{\sum_i (x_i + y_i)}$$

Distances Computed on Vectors with Missing Data

When there are missing values in the data vectors, the distance metrics will operate only on the data points that are not missing. This is best illustrated by example. Consider two vectors x and y:

x = {1, 3, 7, ?, 4, 8}

y = {?, 4, 5, 8, 9, 6}

Since each of these vectors contains missing data, the distance will be computed using a subset of each vector containing only the data elements for which each vector has a value:
x' = {3, 7, 4, 8}
y' = {4, 5, 9, 6}

Some dissimilarity measures are affected more than others by the missing data. The measures that will perform best under these conditions include:

- Average Euclidean Distance
- Average Minkowski Distance
- Pearson's r
- Rank Correlation
- Kendall'sτ
- Shape Dissimilarity
- Cosine Coefficient

Measures that tend to grow larger as the vector gets longer are not good candidates for use on data that has missing values. These include:

- Euclidean Distance
- Minkowski Distance
- Mahalanobis Distance
- Absolute Value Distance
- Canberra Metric
- Bray-Curtis Coefficient

The minimum and maximum value distances may or may not perform well with missing data - this is problem dependent.

Correspondence Analysis

Introduction

Correspondence analysis is designed to analyze the association between two categorical variables. It builds a contingency table to display data that can be classified by the two variables. One variable is arbitrarily assigned to the rows and the other variable is assigned to the columns.

Configuring the Correspondence Dialog

Open the *Correspondence Dialog* by selecting **Tools > Discover > Correspondence Analysis…** from the Partek main window (Figure 9. 21).

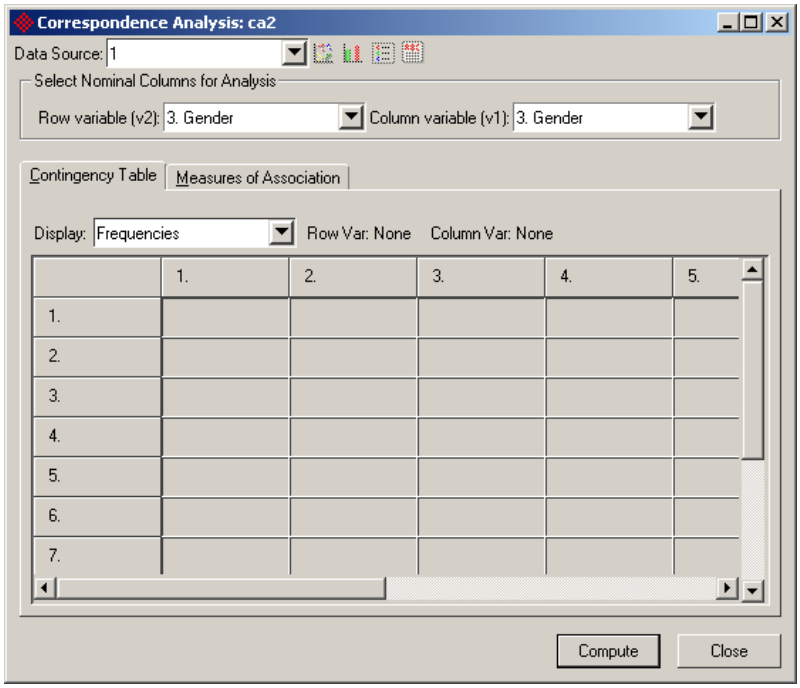


Figure 9. 21: Configuring the Correspondence Analysis dialog

Use the *Data Source* drop-down list to select the spreadsheet to do the correspondence analysis on; the current spreadsheet is the default. If the *Data Source Spreadsheet* is changed, the candidate row and column variables will change to reflect the newly chosen spreadsheet.

Select the *Nominal Columns for Analysis* for the **Row variable (v2)** and **Column variable (v1)** to build a contingency table from the drop-down list (Figure 9. 22).

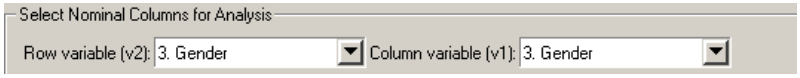


Figure 9. 22: Selecting the Row Variable and Column Variable

Select the values to display in the table from the *Display* drop-down list (Figure 9. 23).

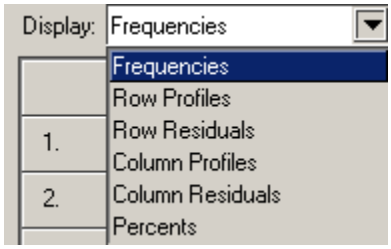


Figure 9. 23: Configuring the value to display in the table

When the *Display* option has been selected, click **Compute**, and the values will be displayed in the table, and the *Measures of Association* will be computed. Select the *Measures of Association* tab in the *Correspondence Analysis* dialog; the results will

be displayed for each measure if the checkboxes were selected (Figure 9. 24). To see a measure that may have not been originally checked, simply click the check button next to the measure and click the **Compute** button at the bottom of the dialog. The results will appear.

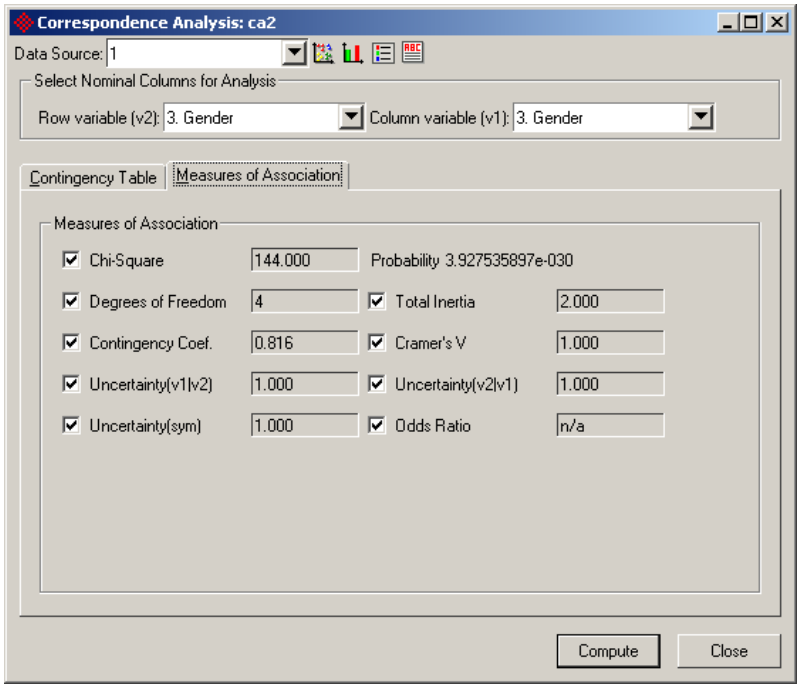


Figure 9. 24: Configuring the Measures of Association dialog

The *Odds Ratio* will only be computed when it is a 2X2 table, which means the row variable and column variable both have only two subgroups respectively.

The ratio=Value[row1, column1] * Value[row2, column 2]/ (Value[row 1, column 2] * Value[row 2, column 1]).

After the result is computed, the accelerator buttons on the upper right of the dialog will be enabled (Figure 9. 25).

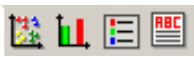


Figure 9. 25: Accelerator buttons

Click on the 2nd button from the left to **Plot Nominal Variables**. A histogram will be drawn on those two variables. The X-axis represents groups of the row variable in the contingency table, different colors represent different groups of the column variable, and the Y-axis represents the value computed in the contingency table (Figure 9. 26).

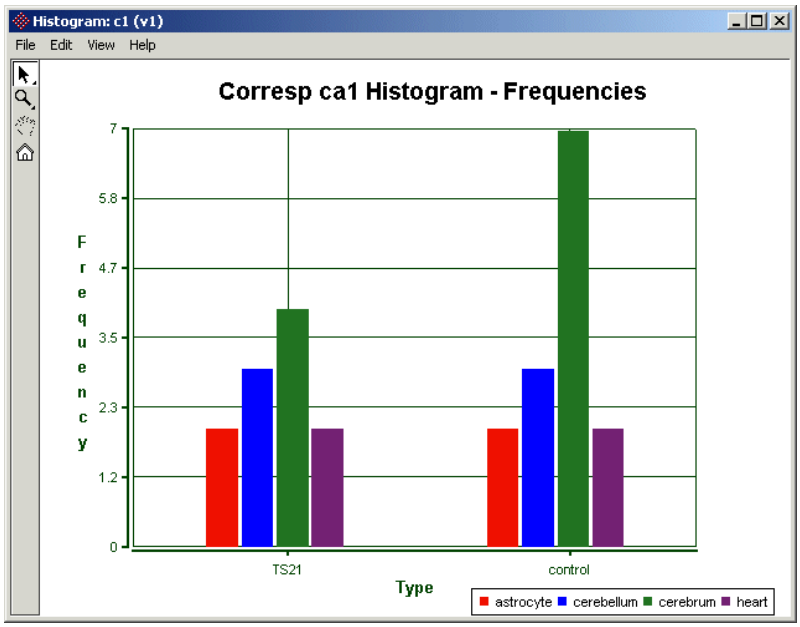


Figure 9. 26: Viewing a histogram of the Frequencies of Row & Column Variables

Click the **Dump Results to Spreadsheet** button (2nd button from the right), to output the computed values to a new child spreadsheet. Choose the value to dump and click **OK** or **Apply** (Figure 9. 27).

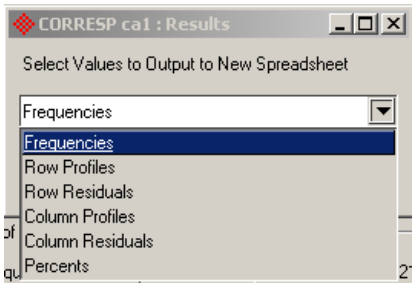


Figure 9. 27: Configuring the Dump Results to Spreadsheet dialog

To get a printable report of the contingency table and the measures of association, click the **Generate Report** button. A report like the one shown in Figure 9. 28 will appear.

Contingency Table Analysis Summary

Datafile: C:/Documents and Settings/www/Desktop/example1

Row Variable: #5 (Gender) Column Variable: #4 (Type)

	TS21	control	Totals
F	10	4	14
M	1	10	11
Totals	11	14	25

MEASURES OF ASSOCIATION

Degrees of Freedom: 1

Chi Square: 9.715 Probability: 0.001827740332

Total Inertia: 0.389

Contingency Coefficient: 0.529

Cramer's V: 0.623

Uncertainty(v5|v4): 0.316

Uncertainty(v4|v5): 0.316

Uncertainty(symmetric): 0.316

Odds Ratio: 25.000

Figure 9. 28: Viewing a report of the results

References

Canavos, G.C., and Miller, D.M., An Introduction to Modern Business Statistics, Belmont, CA: Duxbury Press, 1993.

Cliff, N., *Analyzing Multivariate Data*, San Diego: Harcourt Brace Jovanovich, Publishers, 1987.

Duda, R. and P. Hart, Pattern Classification and Scene Analysis, New York: John Wiley and Sons, 1973.

Dixon, John, K., “Pattern Recognition with Partly Missing Data,” IEEE Trans actions on Systems, Man, and Cybernetics, vol. 9, No. 10, October 1979.

Greenacre, M.J., “Theory and Applications of Correspondence Analysis”, London: Academic Press, 1984.

Greenacre, M.J., *Correspondence Analysis in Practice*, London: Academic Press, 1993.

Hogg, Robert V., and Tanis, Elliot A., *Probability and Statistical Inference*, New York: Macmillan Publishing Co., 1983.

Hand, D.J., and Taylor, C.C., Multivariate Analysis of Variance and Repeated

Measures, New York: Chapman and Hall Ltd., 1987.

Keller, G., Warrack, B., and Bartel, H., *Statistics for Management and Economics*, Belmont, CA: Duxbury Press, 3rd edition, 1994.

Kendall, M.K., *Multivariate Analysis*, New York: Macmillan Publishing Co., 2nd edition, 1980.

Lebart, L., Morineau, A., and Warwick, K., *Multivariate Descriptive Statistical Analysis*, New York: John Wiley & Sons, 1984.

Milton, J.S., and Arnold, J.C., Probability and Statistics in the Engineering and *Computing Sciences*, New York: McGraw-Hill Book Company, 1986.

Reynolds, H.T., *Analysis of Nominal Data*, Beverly Hills: Sage Publications, 1977.

Tabachnik, B., G., and Fidell, L., S., *Using Multivariate Statistics*, NY: Harper & Row, 2nd edition, 1989.

Tou, J.T. and Gonzalez, R.C., *Pattern Recognition Principles*, Reading MA: Addison-Wesley Publishing Company, 1974.

Data Transformation

In this chapter, you will learn how to plot the location of and impute missing values, normalize and scale your data, smooth-out time-series data, shift the values of the rows and columns, and create a transposed spreadsheet.

Missing Data

Missing data often occurs in empirically measured data in science and engineering. Proper handling of missing data is necessary to avoid biased analysis results and to help make better use of the data that is measured.

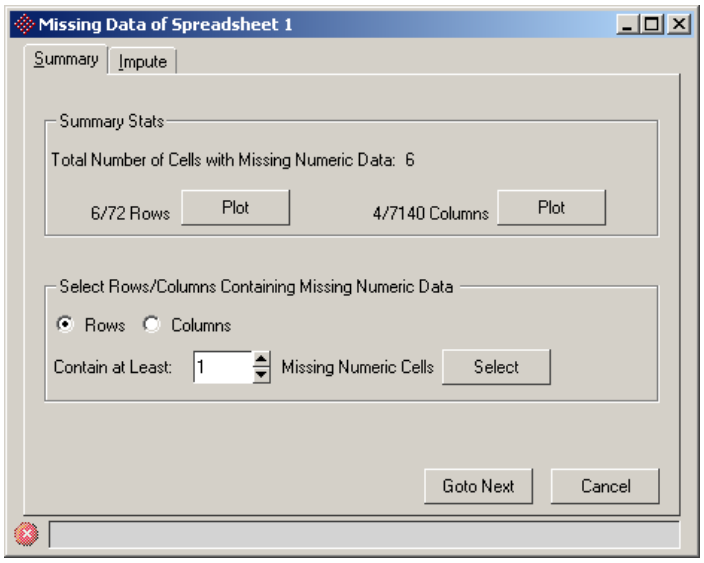


Figure 10. 1: Configuring the Missing Data dialog (Transform > Missing Data)

Missing Data Concepts

Missing data may occur in any or all variables measured and in any or all observations. Missing data is unavoidable in many situations where data is collected (e.g. medical data). The most common currently used solution to the missing data problem is to ignore all observations that contain missing data on any value (commonly called “case-wise deletion”). When the sample size is large relative to the amount of missing data, case-wise deletion may be appropriate because it ignores a large amount of data and may reduce the power of statistical analysis introduced when the missingness is not completely at random.

Another common approach to the missing data problem is to fill in, or “impute” the missing values with a value that is believed to be a good estimate for the missing value. One of the simplest approaches is to impute missing values by substituting the mean or median of the variable for the missing value. While this preserves the mean or median of the variable, it causes estimates of variance and covariance to be biased towards zero.
A third approach is to try to predict the missing values using the values for other variables for the observation that are not missing. Imputing values based on predictive models can be useful but also biases the correlations between variables to be higher than they truly are.

Causes of Missing Data

Before intelligently imputing or otherwise dealing with missing data, the nature of the missing-ness must be understood. Almost all statistical approaches to imputation of missing data rely on the assumption of "ignorability". Ignorability essentially means that the missingness of the data is not dependent on the distribution of the variable that is missing. For example, if in a survey, a person refused to answer a question about smoking habits, there might be a chance that this is not "ignorable". A possible reason might be that smokers are more likely to leave the question unanswered than non-smokers are. In this case, the model of missingness is not ignorable and not missing at random (MAR). In this example, "casewise deletion", in which all observations that contain missing data in any of the variables is omitted from analysis, will produce bias by under sampling the smokers.

Another common technique is complete case (CC) where missing values are replaced by the mean or median of a variable. While the technique may be regarded as an ignorable procedure because it is consistent with the belief that the data is missing completely at random (MCAR), it is not a general ignorable method because it discards the missing values completely.

A major advantage of general ignorable procedures over ad hoc procedures is that the general ignorable methods remove all of the non-response bias explainable by the observed values of the variable, whereas ad hoc procedures, such as case deletion and CC, may not.

Methods for Dealing with Missing Data

Casewise Deletion

The simplest and most common way of handling missing data is to ignore all observations that contain missing data on any of the variables and analyze only those records that have known values for all variables being considered. This strategy is the default behavior of most statistical software packages including Partek. It may be satisfactory if only a small portion of the observations contain missing values, but it can lead to serious biases in results if the pattern of missingness is not ignorable and not MAR.

Imputation

Imputation based procedures involve "filling in" the missing values with what are believed to be reasonable values and then proceeding to analyze the data as if it had no missing values. Common procedures for imputation include Mean/Median/Min/Max imputation (Figure 10. 2).

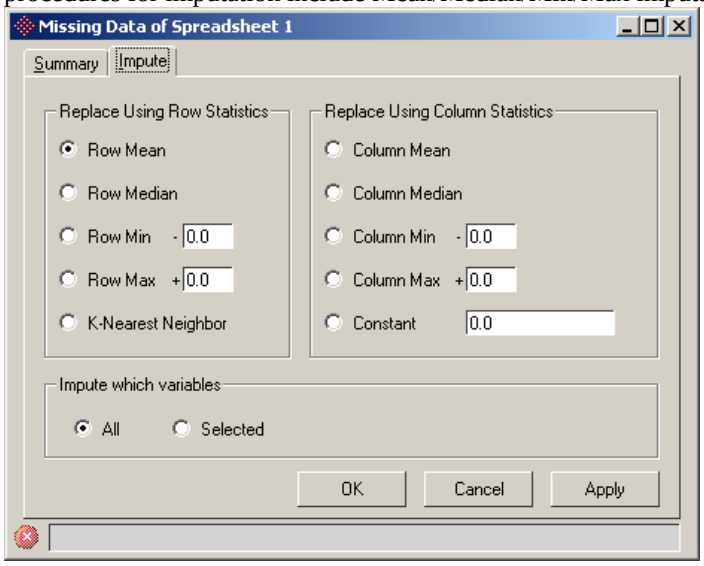


Figure 10. 2: Configuring the Impute tab of the Missing Data dialog

Missing values are filled with the mean/median/min/max respectively of the measured values for the variable being imputed.

- Imputation with a Constant: Imputing a constant value is appropriate when it is believed that the missingness of the data is due to the true value of the missing value being above or below the range of the device of measurement. It is sometimes reasonable that these missing values have a suitable high or low constant at or beyond the maximum or minimum of the measurable range for that variable

- K-Nearest Neighbor imputation: In "K-Nearest Neighbor" imputation (Figure 10. 3), a missing value is generated from several similar observations (to get the dialog in Figure 10. 2, select *K-Nearest Neighbor* in the *Impute* tab of the *Missing Data* dialog (Figure 10. 1), and press **OK**)

In Partek, all observations are compared to the observation containing the missing value by using the user-selection *Distance Measure*. Some *Distance Measures,* such as Average Euclidean and Pearson's Dissimilarity, are more robust to missing data. The values from k observations (*# of Neighbors*) that are most similar to the observation being imputed are used for imputation. When k or *# of Neighbors* equals 1, the algorithm is also known as "Hot Deck" imputation

Usually, *only samples with no missing data in response variables* can be neighbors; however, the alternative is *All samples* can be neighbors. In this case, those response variables that have missing values will not be used in distance calculation. For example, if sample 1 vector is (1.0 ? 3.0 4.0) and sample 2 vector is (? 5.0 6.0 7.0), the final vectors that are used in distance calculation would be (3.0 4.0) and (6.0 7.0); here "?" represents a missing value. If *Newly imputed samples can be neighbors*, the imputed values will no longer be considered missing.

When the missing value is numeric, it can be replaced by either the *Mean* or the *Median* of the k values in the corresponding variable from the k nearest neighbors. If the type of the variable is integer, only the integer part of the imputed value will be used. For example, 1.9 will be converted to 1 and -1.9 will be converted to -1.
When the missing value is categorical, the most frequently observed category among the k nearest neighbors will be used as the imputed value. In the case of a tie, one category will be randomly picked among the tied categories.

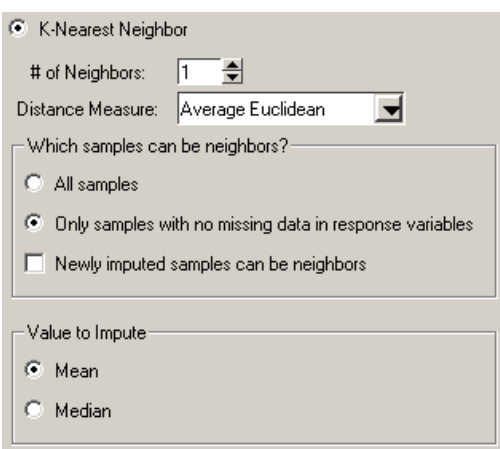


Figure 10. 3: Configuring the K-Nearest Neighbor imputation

Data transformations are commonly used as a remedy for outliers and for failures of normality and linearity. Although it is often useful to transform data into a normal distribution, there are also circumstances where this transformation can be detrimental. For example, transformation may make it more difficult to interpret the value of a variable since it is no longer represented in its natural scale.

Let $\{x_1,...,x_n\}$ denote the original array and let $\{T(x_1),...,T(x_n)\}$ be the transformation of the array elements. The array may be a row or column of a spreadsheet. Table 10. 1 shows data transformations that produce a more normal distribution.

Shape of Distribution	Transformation
Moderate negative Skewness	$T(x_i)=\sqrt{K-x_i}$
Moderate positive Skewness	$T(x_i)=\sqrt{x_i}$
Substantial negative Skewness	$T(x_i)=\log[K-x_i]$
Substantial positive Skewness	$T(x_i)=\log(x_i)$
Substantial positive Skewness with a min of zero	$T(x_i)=\log[x_i+C]$
Severe positive Skewness	$T(x_i)=1/x_i$
Severe positive Skewness with a min of zero	$T(x_i)=1/[x_i+C]$
Severe negative Skewness (J-shaped)	$T(x_i)=1/[K-x_i]$

Table 10. 1: Suggested Corrections for Non-Normal Distributions

C = a constant added to each value so that the smallest value is 1.
K = a constant from which each value is subtracted so that the smallest value is 1; usually equal to the largest value + 1.

- Select **Transform** > **Normalization & Scalin**g in the Partek main menu
- Select the columns or rows to be transformed by clicking the tab at the top of the dialog (Figure 10. 4)

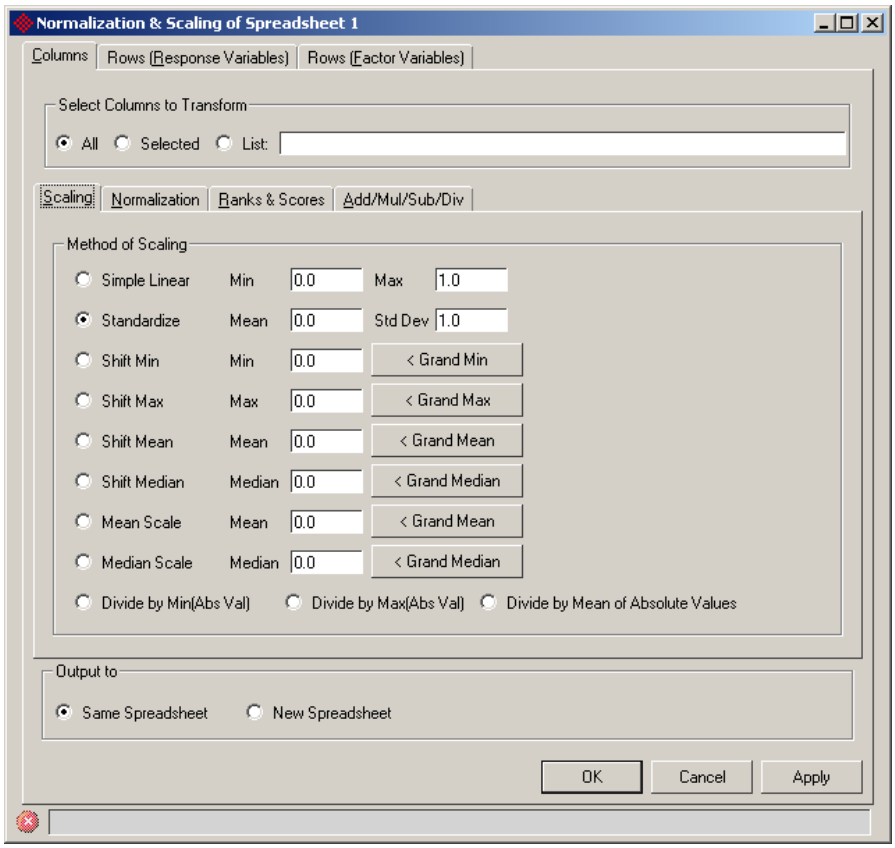


Figure 10. 4: Configuring the Scaling page of the Normalization & Scaling dialog

By default, the transformation is performed on the selected rows or columns. You can also provide a list of row or column numbers, e.g. "1-10 20 30", using "-" to specify a range of numbers and "space" to separate the numbers. Select **All** to transform all the numeric columns. When *Rows (Response Variables)* is selected, transformation will be done only on the response numeric columns of the specified rows; when the *Rows (Factor Variable)* page is selected, transformation will be done only on the factor numeric columns of the specified rows.

Linear Scaling

These transformations perform linear scaling on the data and affect only the location (mean, median) and spread (min, max, standard deviation) of the data. They do not change the shape of the distribution, thus they are not strictly "normalization" methods.
The following scaling methods are in Partek (Figure 10. 4):

- *Simple Linear*: This transformation linearly maps the array to the range [min, max] as specified.
- *Standardize*: Standardizing data shifts and scales the data in such a way that the data will have a mean of zero and a variance of one. This transformation is also called "converting to z-scores". It is computed by subtracting from each number the mean of the data and dividing it by the standard deviation. In Partek, this operation can be generalized to give any desired mean and standard deviation to the resulting distribution.
- *Shift Min*: Shift the distribution of the data to a specified minimum value of C. $T(x) = x + (C - \min(x))$. Click <*Grand Min* to adjust the minimum to the overall minimum of the entire spreadsheet.

- *Shift Max*: Shift the distribution of the data to a specified maximum value C.

$T(x_i) = x_i - \max(x_i) + C$. Click <*Grand Max* to adjust the maximum to the overall maximum of the entire spreadsheet.

- *Shift Mean*: Shift the distribution of the data to a specified mean value C.

$T(x_i) = x_i - \mathrm{mean}(x_i) + C$ click <*Grand Mean* to adjust the mean to the overall mean of the entire spreadsheet.

- *Shift Median*: Shift the distribution of the data to a specified median value C.

$T(x_i) = x_i - \mathrm{median}(x_i) + C$ click <*Grand Median* to adjust the median to the overall min of the entire spreadsheet.

- *Mean Scale*: Scale the distribution of the data to a specified mean value C.

$T(x_i) = C x_i / \mathrm{mean}(x_i)$

- *Median Scale*: Scale the distribution of the data to a specified median value C.

$T(x_i) = C x_i / \mathrm{median}(x_i)$

- *Divide by Min(Abs Value)*: It first gets all the absolute values, and then finds the minimum value. If the minimum of the absolute values is zero, an error will be generated. Otherwise, this transformation will scale data by dividing by the minimum absolute value.
- *Divide by Max(Abs Value)*: It first gets all the absolute values, and then finds the maximum value. If the maximum of the absolute values is zero, which implies all values are zero, an error will be generated. Otherwise, this transformation will scale data by dividing by the maximum absolute value.
- *Divide by Mean of Absolute Values*: It first gets all the absolute values, and then gets the means. If the mean of the absolute values is zero, which implies all values are zero, an error will be generated. Otherwise, this transformation will scale data by dividing by the mean of the absolute values.

Non-linear Transformations

Nonlinear transformations are usually used to improve normality, homoscedasticity, and linearity of the data. Unlike the linear transformations described above, these transformations change the shape of the distribution and are therefore often referred to as “data normalization” procedures. The dialog in Figure 10. 5 can be found by going to **Transform > Normalization & Scaling** in the Partek main menu.

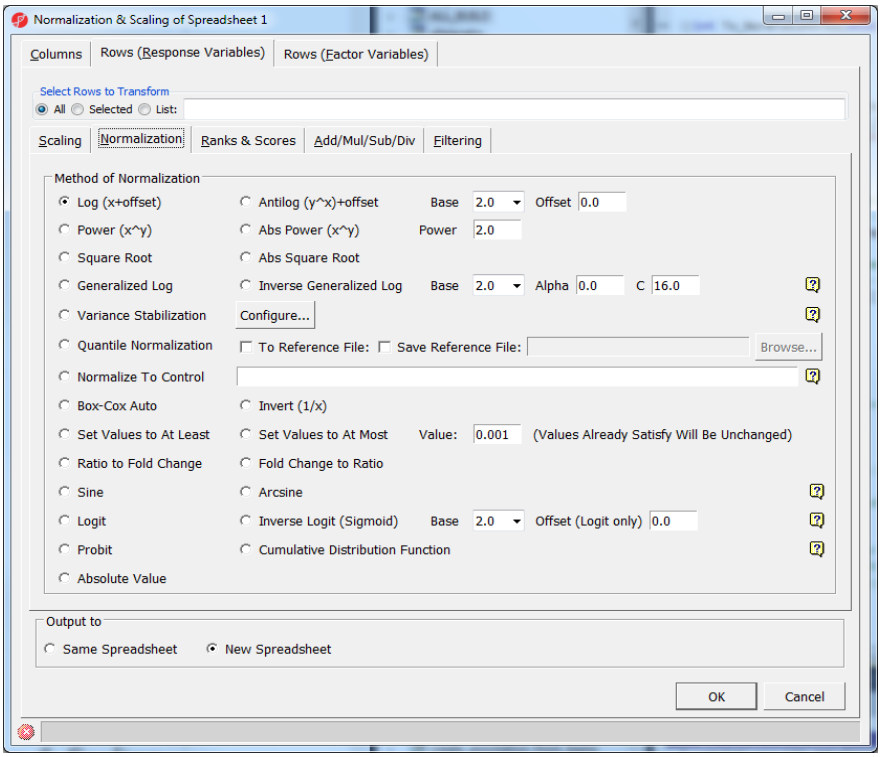


Figure 10. 5: Configuring the Normalization page of the Normalization & Scaling dialog

The following normalization methods are available in Partek.

- *Log (x + C)*: You must specify the base of the log and an offset to apply before taking the log. The offset is used when the data contains zero values or negative values, since the log is not defined for zero or negative values.
- *Antilog (x)+C*: You must specify the logarithm base and an offset to add after the antilog is applied. For example, if you first apply a log(x+1) transformation and you wish to "undo" it, apply an antilog with an offset of –1 to restore to original values (antilog(x)-1).
- *Power* (x^y): You must specify the power, y. If the data contains negative values and the power is not odd, an error is generated.
- *Square Root* ($\sqrt{x}$): Replaces each value with its square root. If the data contains negative values, an error is generated.
- *Abs Power*: This transformation is useful when you would like to apply a power transformation such as square root when the data contains negative values. For negative values, the sign is temporarily removed, power (such as 2 for square root) is applied, and the sign is put back on the result. For example, a transformation of absolute power 2 (absolute square root) 4 would be replaced by 2, and –4 would be replaced by –2.
- *Abs Square Root*: Same as *Abs Power* but the power does not have to be "2"..
- *Generalized Log*: It can be used to overcome the problem of taking Log upon 0 or negative values. It can also stabilize variance on some data whose variances increase with their mean values. For more information, see the Variance Stabilization section below. You must specify the base of the log, an alpha value, and a C value. Base must be positive. Alpha works like an offset. C must be positive.
- *Inverse Generalized*: If you first apply a Generalized Log and you wish to "undo" it, apply the Inverse Generalized Log with the same base, alpha, and C.

- *Quantile Normalization*: It is a rank based normalization method. It first takes m vectors as input: v1 = {x11, x12, x13, …, x1n}, v2 = {x21, x22, x23, …, x2n}, …, vm = {xm1, xm2, xm3, …, xmn}. Second, it calculates a vector V = {X1, X2, …, Xm} that is the average of sorted v1, v2, …, vm.

Namely, X1 is the average of the smallest values in v1, v2, …, vm; X2 is the average of the second smallest values in v1, v2, …, vm; …; Xm is the average of the largest values in v1, v2, …, vm; Then it replaces each value in a vector by using the value that has the same rank in V. For example, if x11 is ranked 5 of all values in v1, it will be changed to X5. For more details about quantile normalization, please refer to Bolstad et al 2003.

After quantile normalization, all vectors should share the same distribution shape except when there are tied values in a vector. For example, suppose, v1 = {0, 0, 0, 10, 20, 30}, v2 = {66, 55, 44, 33, 22, 11}, and the calculated reference vector v = {1, 2, 3, 4, 5, 6}. The three zeros in v1 are tied, so they will be assigned to the average of the first three smallest values in v, i.e. the mean of {1, 2, 3} = 2. After quantile normalization, v1 will become {2, 2, 2, 4, 5, 6} and v2 {6, 5, 4, 3, 2, 1}. In this example, they don't have the same distribution or histogram.

You can save the common distribution as a reference by checking *Save Reference* and specifying the *Reference File*. It then can be used to normalize new data by checking *Normalize to Reference* so the new data will also have this distribution.
Note: The new data and the reference do not necessarily have to be same size. Linear interpolation will be performed when it's needed.

- *Normalize To Control:* It uses the average geometric mean from the control set. Then scales the data to have the same geometric mean. For example, v1 = {x1control1, x1control2, …, x1controln}, v2 = {x2control1, x2control2, …, x2controln}, …, vm = {xmcontrol1, xmcontrol2, xmcontrol3, …, xmcontroln}. It first calculates the geometric mean gmean1, gmean2, …, gmeanm for v1, v2,..., vm. Then gets the average (arithmetic mean) target_gmean from gmean1, gmean2, …, gmeanm. Then transforms the data T(xi) = target_mean * xi/genometric_mean(x). The data must be positive to calculate the geometric mean. E.g., to use row 1, 3, and 5 as the control set, enter: 1 3 5. E.g., to use row from 5 to 9 as the control set, enter 5-9. Reference Paper: Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. By: J Vandesompele et. al., Genome Biology, 18 June 2002
- *Box-Cox Auto*: This applies the "Box-Cox transformation" (Box and Cox, 1964). It is a power transformation that automatically determines the transform to use which best normalizes the data: $f(x) = (x^{\lambda} - 1)/\lambda$. Lambda ($\lambda$) is chosen differently for each row or column in such a way to make the data in that row or column most normal.
- *Invert (1/x)*: This transformation replaced every value with its inverse. If there are zero values, an error is generated.
- *Set Values to At Least*: If x is less than the specified *Value* y, then set x to y; if x >= y the value will be unchanged. For example, if the *Value* y is 0, it will set all negative values to 0; values already >= 0 will be unchanged.

- *Set Values to At Most*: Similar to the transform above, if x > y, then set x to y; if x <= y the value will be unchanged.
- Variance Stabilization

Many popular statistical methods (e.g. ANOVA) have an assumption that the data must have the same variance across all value levels. Some data that violates the assumption can be cured by logarithm transformation. However, there are instances that the regular logarithm can't handle. For example in gene microarray data, log can not be performed upon 0 or negative expressions, which are common. Also, part of the transformed data at low levels still could not satisfy the assumption. Variance stabilization (Geller, Gregg, Hagerman, and Rocke, 2003) is a useful transformation in such cases. The *Variance Stabilization* dialog is shown in Figure 10. 6. The equation of the transformation is:

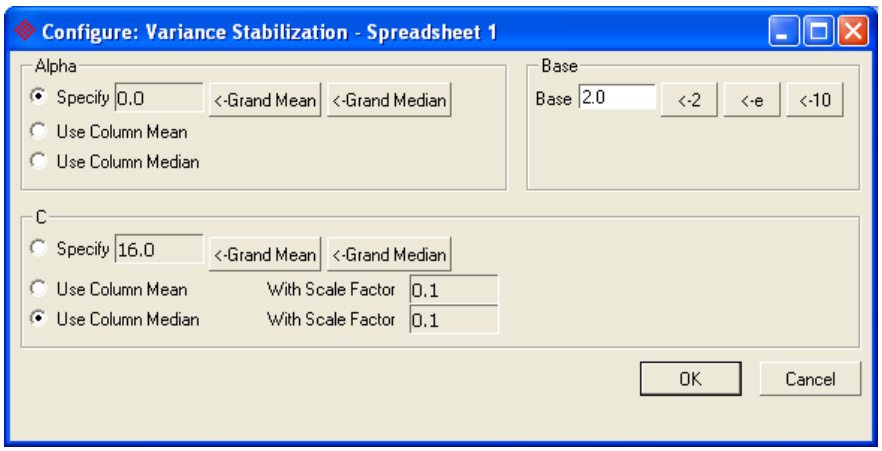


Figure 10. 6: Configuring the Variance Stabilization dialog

By default, the transformation is performed on the S*elected* rows or columns. You can also provide a *List* of row or column numbers, e.g. "1-10 20 30", using "-" to specify a range of numbers and space to separate the numbers. Select *All* to transform all the numeric columns, response variables, or factor variables. When the *Rows (Response Variables)* page is selected, transformation will be done only on the response numeric columns of the specified rows; when the *Rows (Factor Variable)* page is selected, transformation will be done only on the factor numeric columns of the specified rows.

As an example, perform *Variance Stabilization* on the row response variables. Transformations on columns and rows (Factor Variable) can be performed similarly except for the places that are explicitly pointed out.

As in the equation, alpha serves as an offset. You can *Specify* a constant that all the transformations will use the same alpha. Type in the edit box or click *Grand Mean* or click *Grand Median* to assign a constant to alpha. *Use Row Mean* to transform an element with the alpha that equals its row mean. Similarly, each row will have its own alpha when *Use Row Median* is selected.

You can *Specify* a constant C value, or *Use Row Mean/Median* to have different C values for different rows. *Scale Factor* is a handy way to change the sign of, proportionally increase, or decrease *Row Mean/Median.* Since C must > 0, the *Specified, Row Mean x Scale Factor* for each row, or *Row Median x Scale Factor* for each row MUST be positive.

Log *Base* can also be S*pecified*, use *2*, *e*, or *10*.

Ranks and Scores

Ranks and scores are used to transform data into a known (e.g. uniform or normal) distribution.

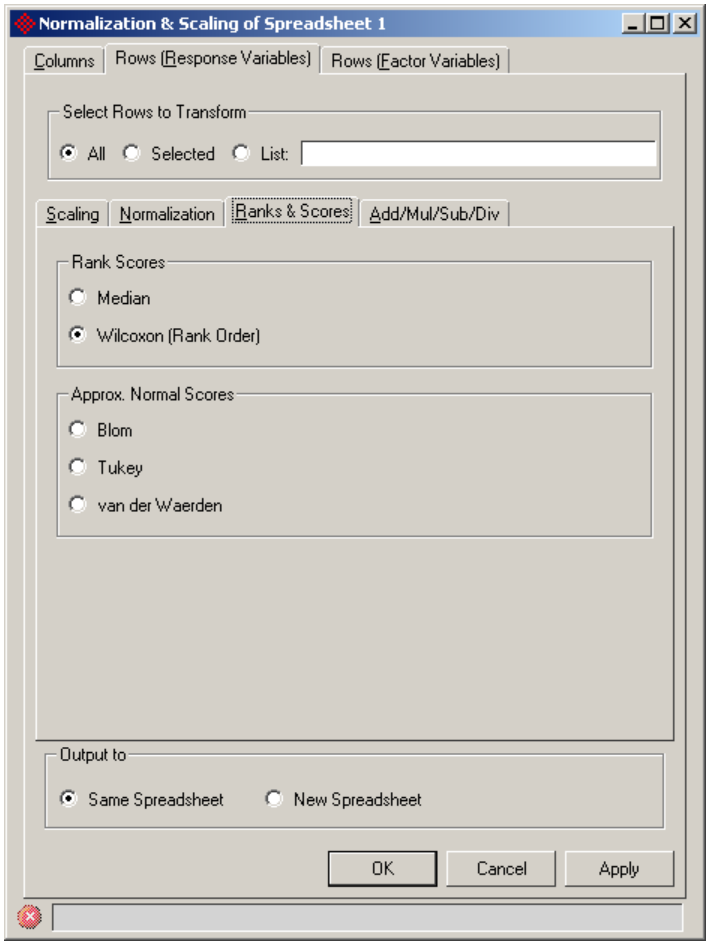


Figure 10. 7: Configuring the Ranks & Scores page of the Normalization & Scaling dialog

The following rank and order transformations are available from the graphical user interface of Partek. Other transformations can be applied by writing a Tcl script. Please contact Partek customer support if you need assistance developing these scripts.

- *Median*: This transformation replaces each value with a "1" or a "0" depending on whether it is greater than or less than the median of all values, respectively. If a value is exactly the median, it is replaced with a "0".
- *Wilcoxon (Rank Order)*: This transformation replaces each value with its rank in the list of sorted values. The smallest value is replaced by 1 and the largest value is replaced by the total number of non-missing values, N. If there are no tied values, this results in a perfectly uniform distribution. In the case of ties, all tied values receive the mean rank.
- *Blom*: This transformation replaces each value with the normal score of its rank in the list of sorted values. $T(x_i) = \phi^{-1}\left(\frac{r_i - 3/8}{n + 1/4}\right)$, where ϕ^{-1} is the inverse cumulative normal function and r_i is the rank of the data as calculated with the Wilcoxon rank transformation.

- *Tukey*: This transformation replaces each value with the normal score of its rank in the list of sorted values. $T(x_i)=\phi^{-1}\left(\frac{r_i-1/3}{n+1/3}\right)$, where ϕ^{-1} is the inverse cumulative normal function and r_i is the rank of the data as calculated with the Wilcoxon rank transformation.
- *van der Waerden*: This transformation replaces each value with the normal score of its rank in the list of sorted values. $T(x_i)=\phi^{-1}\left(\frac{r_i}{n+1}\right)$, where ϕ^{-1} is the inverse cumulative normal function and r_i is the rank of the data as calculated with the Wilcoxon rank transformation

Simple Transformations: Add, Multiply, Subtract, Divide

Use the *Add/Mul/Sub/Div* tab to perform simple addition, multiplication, subtraction, and division of constants, multiply, divide, add, and subtract value in a column, or add random number to a spreadsheet (Figure 10. 8).

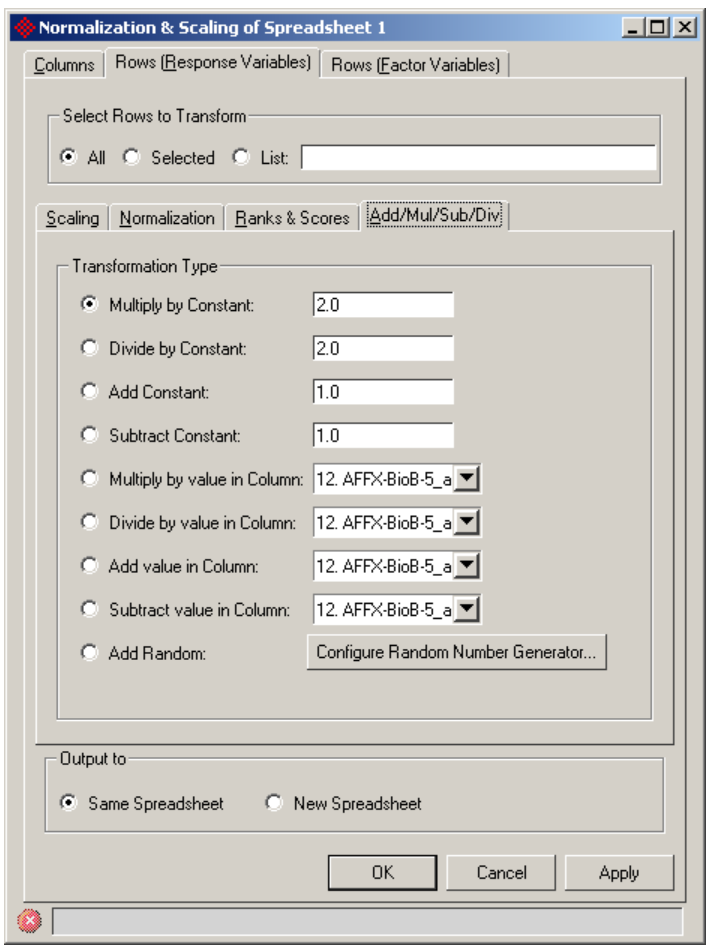


Figure 10. 8: Configuring the Add/Mul/Sub/Div page of the Normalization & Scaling dialog

The following simple transformations are available from the graphical user interface of Partek. Other transformations can be applied by writing a Tcl script. Please contact Partek customer support if you need assistance developing these scripts.

- *Multiply by Constant*: Multiply the rows or columns by a constant C. $T(x_i) = x_i * C$
- *Divide by Constant*: Divide the rows or columns by a value C. $T(x_i) = x_i / C$
- *Add Constant*: Add the rows or columns a value C. $T(x_i) = x_i + C$
- *Subtract Constant*: Subtract the rows or columns by a value C. $T(x_i) = x_i - C$
- *Multiply by value in Column*: Multiply the rows or columns by the corresponding value in a column. Column transform: $T(x_i) = x_i * x_c$. Row transform: $T(x_i) = x_i * x_c$
- *Divide by value in Column*: Divide the rows or columns by the corresponding value in a column. Column transform: $T(x_i) = x_i / x_c$. Row transform $T(x_i) = x_i / x_c$
- *Add value in Column*: Add the rows or columns by the corresponding value in a column. Column transform: $T(x_i) = x_i + x_c$. Row transform $T(x_i) = x_i + x_c$
- *Subtract value in Column*: Subtract the rows or columns by the corresponding value in a column. Column transform: $T(x_i) = x_i - x_c$. Row transform $T(x_i) = x_i - x_c$
- *Add Random*: Add the rows or columns to a random number. $T(x_i) = x_i + R$ where *R* is a random number from a distribution that can be configured by clicking the *Configure Random Number Generator* button

Normalize to Baseline

The *Normalize to Baseline* dialog (Figure 10. 9) allows you to normalize your data. The first option in the dialog box, *Load Baseline*, uses a baseline file (as generated by **Tools > Create Baseline**). If you have a categorical column that identifies the baseline samples then specify it by *Identifier Column* and choose the *Baseline Category*. Each column for each sample will be divided by the average of all samples specified as baseline.

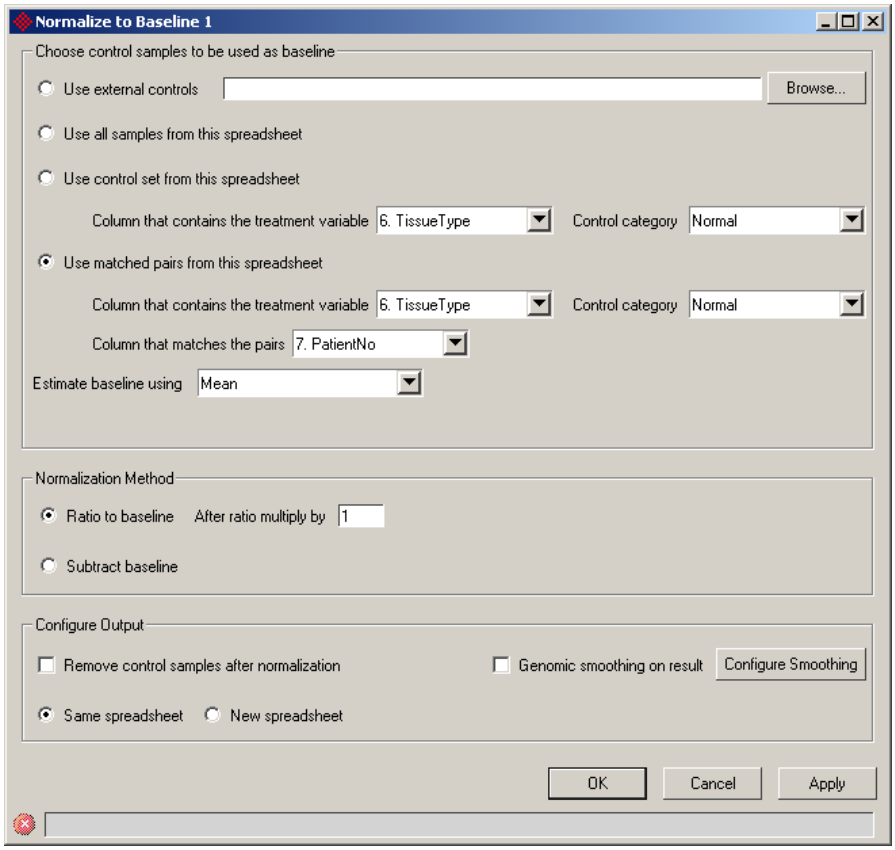


Figure 10. 9: Normalize to Baseline dialog

If you are normalizing using an *Identifier column*, you can check **Delete normals** to remove the baseline after normalization.

If the spreadsheet is associated with genomic information, you have the option to perform genomic smoothing immediately after normalization. The **Configure Smoothing** button invokes a dialog like the one from **Transform > Genomic Smoothing** (see below).

Smoothing

Smoothing is performed to smooth out (usually time-series) data. The smoothing operations all convolve a smoothing filter of a specified size and weight along a vector of numbers.

Go to **Transform > Smoothing** in the Partek main menu; the dialog in Figure 10. 10 will appear. Let *n* be the filter length. If it is an odd number, the value of a cell in the vector is smoothed by the values of the cells to the left and cells to the right of that cell (if possible). If *n* is even, the values of cells to the left and cells to the right of that cell are used for smoothing. Let *n* be the set of cells of the vector that fall in the filter. Note that the boundary cells (cells in the beginning and at the end of the vector), which may not have enough cells to the left or right, are smoothed by whatever cells that are available.

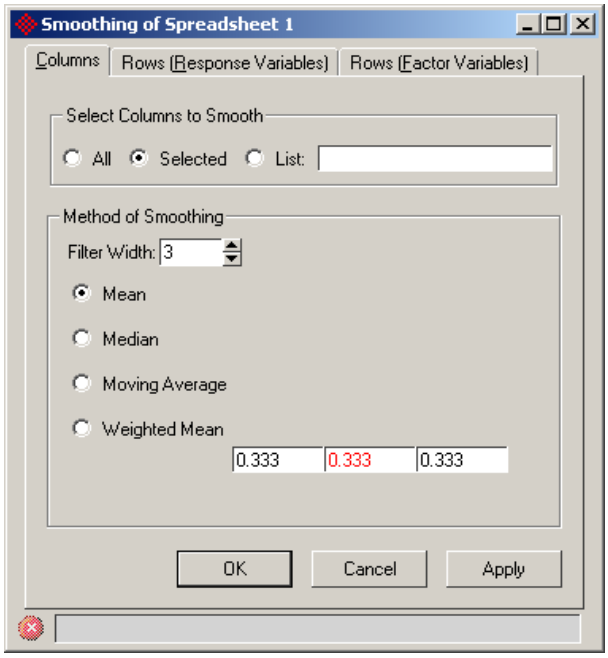


Figure 10. 10: Configuring the Smoothing dialog

For example, if the *Filter Width* is set to **5** the first cell in the vector will be replaced with the average of the first 3 (array) elements. The second element will be replaced with the average of the first 4 elements. The third (and subsequent non-boundary) element will be replaced with the mean of itself, the two preceding and two following elements. The same idea is applied at the end of the vector.

- *Mean Filter*: The value of a cell is replaced by the mean of the values of the cells that fall in the filter.
- *Median Filter*: The value of a cell is replaced by the median of the values of the cells that fall in the filter.
- *Weighted Mean Filter*: This is the same as mean filter, except that the value of each cell is multiplied by the weight of that cell.
- *Moving Average*: For this type of smoothing, the filter configuration is different. Let *n* be the filter size. Each cell of the array is replaced by the average of the *n* cells including and preceding the cell itself. If there are not *n* cells to the left of a cell, the value of the cell is replaced by the average of cells available.

Genomic Smoothing

The *Genomic Smoothing* dialog (Figure 10. 11) performs Gaussian smoothing with the location of a column determined by the genomic location of the probe set. Genomic smoothing is most useful for the visual detection of large changes (such as trisomy).

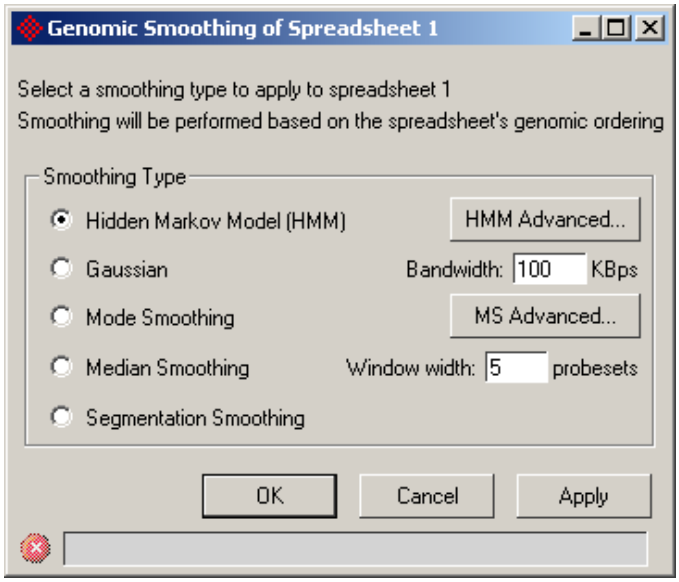


Figure 10. 11: Genomic Smoothing dialog

Smoothing is an optional step and should be performed after analysis. The window width should be smaller than the size of the regions you detect.

If you are looking for very small regions of change (such as microdeletions) then you should not perform smoothing.

Hidden Markov Model (HMM) smoothing finds the most likely state at each genomic locus by assigning a hidden state at each locus based on the observed data and the neighboring states. You may specify a list of known states that are all potential values for the hidden states. The probability of no transition specifies the probability of retaining the same hidden state as the neighboring locus. This is a constant value independent of genomic distance between neighboring loci.

Genomic decay will make the probability follow an exponential decay from the max probability to the initial probability. This allows probes with larger gaps in the data to be treated more independently than closely spaced probes. Decay can be disabled by setting the parameter's value to 0.The decay function is described more specifically as:

alpha = e^(d / decay)
P(S(t) == S(t+1)) = alpha * MaxProbability + (1 - alpha) *
InitialProbability
P(S(t) != S(t+1)) = 1 - P(S(t) == S(t+1)

where d is the distance in basepairs between two probes, decay and MaxProbability are the specified parameters, and S(t) is the hidden state at observation t. The InitialProbability is 1 / # states; i.e. all states are assumed to be equally likely.

The sigma parameter describes the width of the normal distribution from which observations are drawn for each hidden state.

The *Gaussian Bandwidth* is the width of the Gaussian kernel. An increase in value results in more contribution from distant loci.

Mean Shift smoothing is a mode seeking procedure using a multivariate, nonparametric density estimation technique. It is possible to specify the signal and genomic distance domain bandwidths independently.

Create a Transposed Spreadsheet

Create Transposed Spreadsheet will automatically transpose all the selected rows and columns in the spreadsheet. Select **Transpose > Create Transposed Spreadsheet** from the Partek main menu.

Random Number Generator

There are six different ways to generate random numbers in Partek—*Uniform*, *Normal*, *Exponential*, *Gamma*, *Binomial* and *Poisson*. To reproduce the resulting sequence of random numbers, you will need to specify the *Seed* (Figure 10. 12) and some criteria like mean and standard deviation for normal distribution data.

Note: The *Random Number Generator* is available on the *Add/Mul/Sub/Div* tab from **Transform > Normalization & Scaling**.

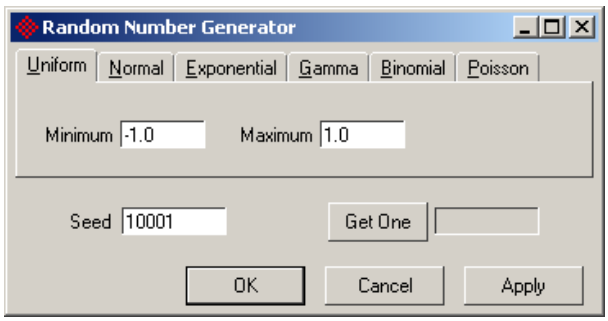


Figure 10. 12: Configuring the Random Number Generator dialog

Click the **Get One** button; the number will be shown in the entry next to the button.

- Uniform distribution

The uniform distribution has a constant probability of occurrence in an interval (min, max) and zero probability elsewhere. It is also called rectangular distribution.

- Normal distribution
- Exponential distribution
- Gamma distribution
- Binomial distribution

The binomial distribution occurs when observing a stationary Bernoulli process. The binomial random variable can take on only integer values ranging from 0 to N inclusive, where N is the number of observations (trials). The probability of observing any particular value, r, is given by

$$p(x = r; N, p) = \binom{N}{r} p^r (1-p)^{N-r}$$

where p is the probability of success at each observation.

- Poisson distribution

The Poisson distribution is similar to the binomial distribution. A random variable drawn from Poisson distribution also takes on only integer values. The Poisson distribution might be best understood by regarding it as a special case of the binomial distribution where N is very large and p is very small – thus making the binomial probability difficult to calculate. The probability function for a Poisson random variable is:

$$P(X=x)=\frac{e^{-m}m^{x}}{x!}$$

where m=Np is the intensity of the distribution.

Gene Summary

Certain expression values, such as exon data, can be summarized to more general values, such as transcripts. You must have exon data or a meta-probeset file associated with this data file for this option to appear.

- To associate a spreadsheet with a meta-probeset file select **File > Properties**
- Select the **Advanced** button at the bottom left corner of the *Properties* dialog (Figure 10. 21)

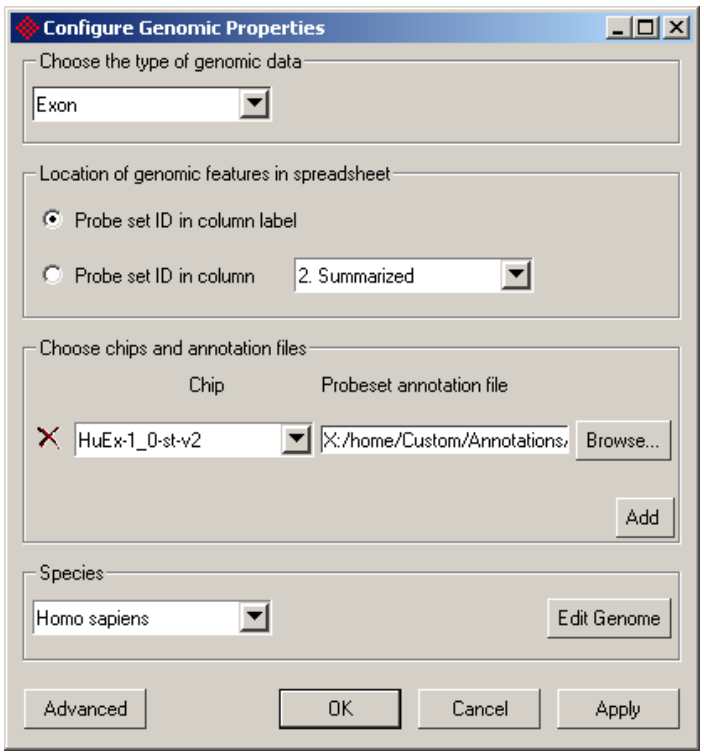


Figure 10. 13: File properties

- Select **Meta Probeset** from the *Add Property* drop down list and click **Add** (Figure 10. 14)

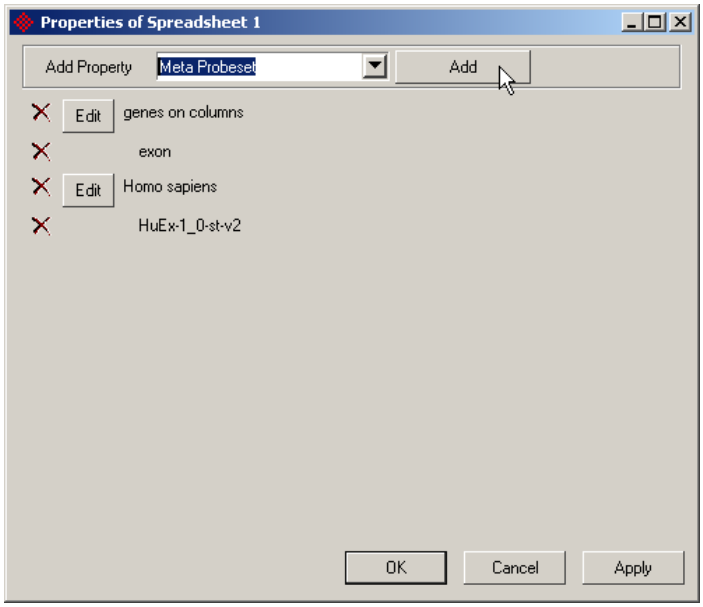


Figure 10. 14: Selecting the Metaprobeset

- Choose a pre-defined description or enter your own. You will be prompted for the location of the meta-probeset file when you invoke a dialog that requires it (Figure 10. 15)

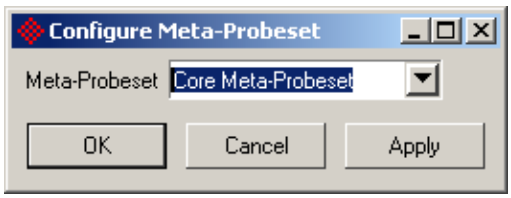


Figure 10. 15: Configuring the Meta-Probeset property

The meta-probeset file is a tab-delimited file provided by Affymetrix which maps from transcript_cluster_ids to exon probeset_ids.

The gene summary will have one column for each row in the meta-probeset file. The first column of the meta-probeset file is the new probeset id. This column will be used as the column label.

The second column is the transcript_cluster_id. In the Affymetrix supplied meta-probeset files the probeset_id matches the transcript_cluster_id.

The third column is a space-separated list of probeset ids in the original spreadsheet that belong to the transcript cluster specified in the first column.

Example:
#comments start with #
#%create_date=Thu Jan 5 15:32:53 PDT 2006
probeset_id transcript_cluster_id probeset_list probe_count
3948543 3948543 3948549 3948555 3948556 3948570 3948572 3948577 3948584 28

Configuring the Gene Summary dialog

- Open the *Gene Summary* dialog by selecting **Tools > Gene Summary.** The dialog shown in Figure 10. 16 will appear

- Choose the summarization method and the output file
- Click **OK** or **Apply**

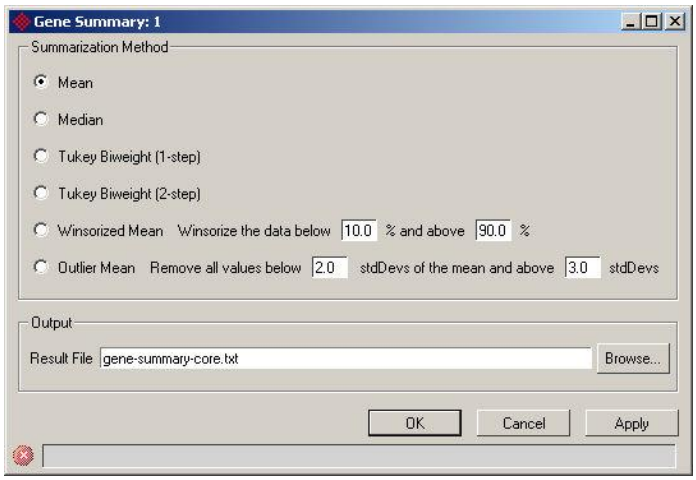


Figure 10. 16 Configuring the Gene Summary Dialog

The result spreadsheet will contain a column for each row in the meta-probeset file with the column label equal to the value in the first column.

References

Box, G. E. P. and Cox, D. R. (1964) An analysis of transformations. *JRSS B* **26** 211–246.

Cho, R. J., et. al. "A Genome-Wide Transcriptional Analysis of the Mitotic Cell Cycle". Molecular Cell **2**, 65-73 (1998).

Geller, S.C. Gregg, J.P. Hagerman, P. and Rocke, D.M. Transformation and Normalization of Oligonucleotide Microarray Data Bioinformatics, 2003, 19, 1817-1823

Lehmann, E.L. (1975). Nonparametrics: Statistical Methods Based on Ranks, Holden-Day, San Francisco.

Bolstad, B.M., Irizarry R. A., Astrand, M., and Speed, T.P. (2003) A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Bias and Variance. Bioinformatics 19(2):185-193

Inferential Statistics

Introduction

The inferential statistics tests in Partek are explained in this chapter. The tests are grouped as parametric and non-parametric. Parametric tests, such as the two sample t-test, paired sample t-test, one sample t-test, one sample z-test, and the ANOVA, take a measurement from a sample to represent a population. Non-parametric tests such as the Mann-Whitney, Kruskal-Wallis, Kolmogrov-Smirnov, Friedman, and Quade do not follow the same rules as do parametric tests; they assume that a normal distribution within a population does not exist.

Parametric Tests

Two-Sample *t*-test

Introduction

The two-sample t-test is used to test for a difference in means between two groups when there are different subjects in each group. It can also be used to test whether the means of two groups are different by a specific amount. The two groups are assumed to be normally distributed and have equal variance (for the equal variance *t*-test). If the variances of the two groups are different, the unequal variance *t*-test should be used. However, when the variances are equal, the equal variance t-test has more power. The equal variance *t*-test is equivalent to a one-way analysis of variance (ANOVA) when comparing only two groups.

Implementation Details

Partek uses Satterwaithe's approximation when computing degrees of freedom for the unequal variance *t*-test.

Configuring the Two-Sample *t*-test Dialog

Open the two-sample *t*-test dialog by selecting **Stat > Parametric Tests > Two-Sample t-test…** from the Partek main window (Figure 11. 2). You will use this dialog to specify the hypothesized difference, the grouping variable (factor), the response variable(s) to be tested, multiple test corrections and whether to use the equal or unequal variance version of the test. Figure 11. 1is configured to compare an equal variance *t*-test between ALL and AML on all numeric variables.

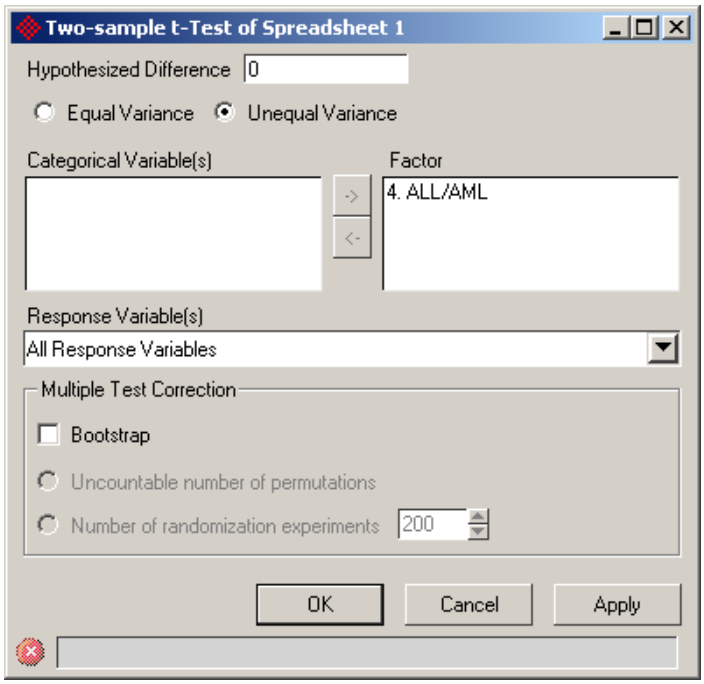


Figure 11. 1: Configuring the Two-Sample t-test dialog for multiple tests

Selecting Grouping Variables

The grouping variable (or factor) must be selected from the *Categorical Variable(s)* list, which contains variables that have only two categories (levels) in the spreadsheet. There can be only one grouping variable in a t-test computation. When an item in the *Categorical Variable(s)* list box is selected, the -> button next to the *Factor* list box will be enabled. Click on it to move the selected item to the *Factor* list box. This can also be done by double clicking the item. To remove a factor, select it in the *Factor* list box and the <- button next to it will be enabled. Click on the <- button and the item selected in the *Grouping Variable* list box will be moved back to the *Categorical Variable(s)* list box.

Selecting Response Variables

By default, *All Response Variables* will be shown as the *Response Variable(s)* to test all of response numeric variables at one time. If *All Variables* is chosen, all the factor and response numeric variables will be tested. To choose a specific response variable to test, select the variable name from the drop-down list.

When an analysis is performed on all numerical variables, the results will be summarized in a new spreadsheet that is a child of the original. In the results spreadsheet, each variable tested in a row is summarized by the column number and name of the variable, and followed by summary statistics including the p-values, means, and standard deviations. The rows are automatically sorted by the column of p-values. To sort by a different column, right click on the column heading and select **Sort Ascending/Descending** in the pop-up menu. Detailed reports about individual test variables can be viewed by right-clicking on the row label

corresponding to the variable and selecting the **HTML Report** option on the pop-up menu (Figure 11. 2).

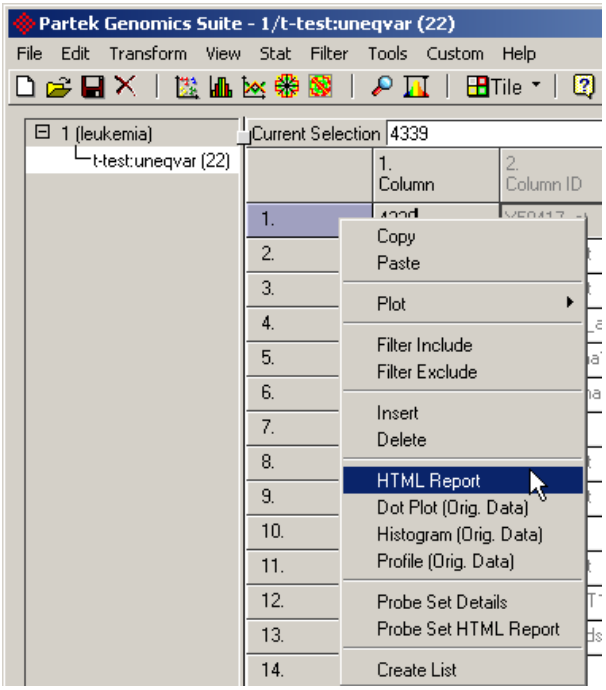


Figure 11. 2: Viewing the Result Spreadsheet of the Two-Sample t-test, multiple tests

To test a single response variable, select the variable from the drop-down list, the result will be displayed in a HTML report (Figure 11. 3).

Equal Variance Two Sample t-test of "leukemia"
on AFFX-BioB-5_at

Grouping Variable: ALL/AML
Response Variable: AFFX-BioB-5_at
Hypothesized Difference: 0

Factor Level Information

Factor	Levels	Level Values
ALL/AML	2	ALL, AML

Descriptive Statistics

ALL/AML	N	Mean	Std. Dev.	Std. Err.	Minimum	Maximum
ALL	47	-122.723	105.361	15.3685	-476	86
AML	25	-99.28	81.3011	16.2602	-318	15
Total	72	-114.583	97.781*	11.5236*	-476	86

* pooled standard deviation

Figure 11. 3: Viewing the Two-Sample t-test report for a single test

Bootstrap

When multiple response variables are tested, Partek can compute corrected p-values using a variety of methods. To compute bootstrap, check the box in front of

Bootstrap in the *Multiple Test Correction* panel of Figure 11. 1. For more information about bootstrap, please see the Multiple Test Correction for P-Values section below

Running the Computations

OK will perform the configured *t*-test computation and dismiss the dialog.

Apply will perform the configured *t*-test computation, but the two-sample *t*-test dialog will remain to allow for another computation.

Cancel will close the dialog without doing any computation.

The Paired Sample t-Test

Introduction

The Paired sample t-test is used to test for a difference in means between two groups when the two groups contain different measures on the same subjects. Examples of when to use the paired t-Test include measuring the effect a drug has on a group of subjects at two different time points or comparing two tissue types taken from the same subject. It can also be used to test whether the means of two groups are different by a specific amount. The two groups are assumed to be normally distributed and have equal variance.

Implementation Details

The data has to be balanced, meaning each subject must have two measurements. If for any subject, there are not exactly two samples, all samples from the subject will be omitted from the calculation.

Configuring the Paired Sample t-test Dialog

Open the paired sample *t*-test dialog by selecting **Stat > Parametric Tests > Paired Sample t-test…** from the Partek main window (Figure 11. 4). This dialog is used to specify the hypothesized difference, the subject variable, the grouping variable (factor), the response variable(s) to be tested, and multiple test corrections. Figure 11. 4 is configured to use the paired *t*-test to compare two different tissue types taken from each animal on all numeric variables.

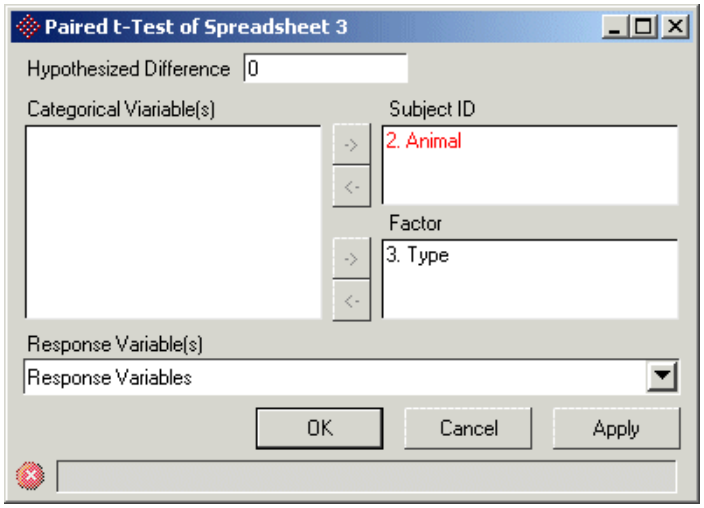


Figure 11. 4: Configuring the Paired Sample t-test dialog, multiple tests

Selecting Grouping Variables

The *Subject ID* and *Factor* must be selected from the *Categorical Variable(s)* list, which contains variables that have more than two categories (levels) in the spreadsheet. There can be only one subject variable and one factor variable in a paired t-test computation. When an item in the *Categorical Variable(s)* list box is selected, the -> button next to the *Grouping Variable* list box will be enabled. Click on it to move the selected item to the *Subject ID* or *Factor* list box. To remove a subject or a factor, select it and the <- button next to it will be enabled. Click on the <- button and the item selected will be moved back to the *Categorical Variable(s)* list box. The *Factor* variable must have two subgroups (levels).

Selecting Response Variables

By default, *All Response Variables* will be shown as the *Response Variable(s)* to test all of response numeric variables at one time. If *All Variables* is chosen, all the factor and response numeric variables will be tested. To choose a specific response variable to test, select the variable name from the drop-down list.

When an analysis is performed on all numerical variables, the results will be summarized in a new spreadsheet, which is a child of the original. In the results spreadsheet, each variable tested in a row is summarized by the column number and name of the variable, and followed by summary statistics including the p-values, means, and standard deviations. The rows are automatically sorted by the column of p-values. To sort by a different column, right click on the column header and select **Sort Ascending/Descending** in the pop-up menu. Detailed reports about individual test variables can be viewed by right-clicking on the row label corresponding to the variable and selecting the **HTML Report** option on the pop-up menu (Figure 11. 5).

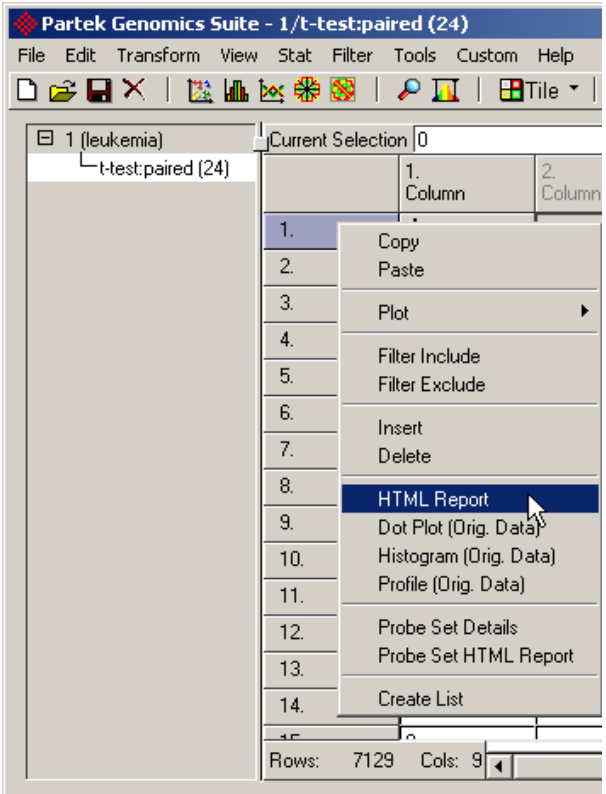


Figure 11. 5: Selecting the HTML Report from the result spreadsheet for multiple tests, paired sample t-test

To test a single response variable, select the variable from the drop-down list; the result will be displayed in a HTML report (Figure 11. 6).

Paired sample t-test of "Two-Tissues" on A01157cds_s_at

Factor Level Information

Factor	Levels	Level Values
Type	2	substantia, ventral

Descriptive Statistics

Type	N	Mean	Std. Dev.	Std. Err.	Minimum	Maximum
substantia	7	5.77788	0.119607	0.0452074	5.59735	5.87863
ventral	7	6.02222	0.0882777	0.0333658	5.94261	6.14536
Total	14	5.90005	0.16209	0.0433204	5.59735	6.14536

t-test

t Statistic	DF	p-value (2-tailed)
-7.0369	6	0.000411567

Figure 11. 6: Viewing the HTML report for a single test

Running the Computations

OK will perform the configured *t*-test computation and dismiss the dialog.

Apply will perform the configured *t*-test computation, but the paired sample *t*-test dialog will remain to allow for another computation.

Cancel will close the dialog without doing any computation.

One-Sample *t*-test

Introduction

The one-sample *t*-test is used to test for a difference in means between a sample group and a hypothesized population. Like many other parametric tests, the assumptions are that samples are independent and the values are normally distributed. The one sample *t*-test is used instead of the one-sample *z*-test when the population standard deviation is unknown.

Configuring the One-Sample *t*-test Dialog

Open the one-sample *t*-test dialog by selecting **Stat > Parametric Tests > One-Sample t-test…** from the Partek main window (Figure 11. 7). You will use this dialog to specify the hypothesized mean, the response variable(s) to be tested, and multiple test corrections. Figure 11. 7 is configured to compare a sample mean to 0 on all the numeric variables.

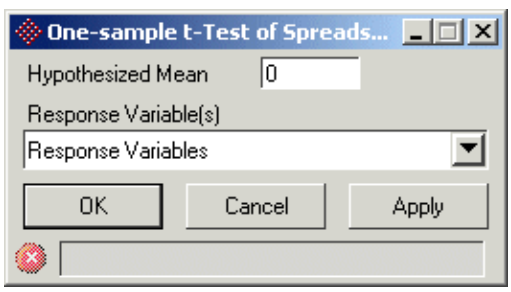


Figure 11. 7: Configuring the one-Sample t-test dialog for multiple tests

Selecting Response Variables

By default, if there is more than one numeric variable in the spreadsheet, *Response Variables* will be shown as the *Response Variable(s)* to test all of the variables at one time. To choose a specific response variable to test, select the variable name from the drop-down list; however, if there is only one numeric variable in the spreadsheet, by default the variable name will be selected as the *Response Variable*.

When an analysis is performed on all numerical variables, the results will be summarized in a new spreadsheet that is a child of the original. In the results spreadsheet, each variable tested in a row is summarized by the column number and name of the variable, and followed by summary statistics including the p-values, means, and standard deviations. The rows are automatically sorted by the first column of p-values. To sort by a different p-value, right click on the column heading and select **Sort Ascending** in the pop-up menu. Detailed reports about

individual test variables can be viewed by right-clicking on the row label corresponding to the variable and selecting the **HTML Report** option on the pop-up menu (Figure 11. 8).

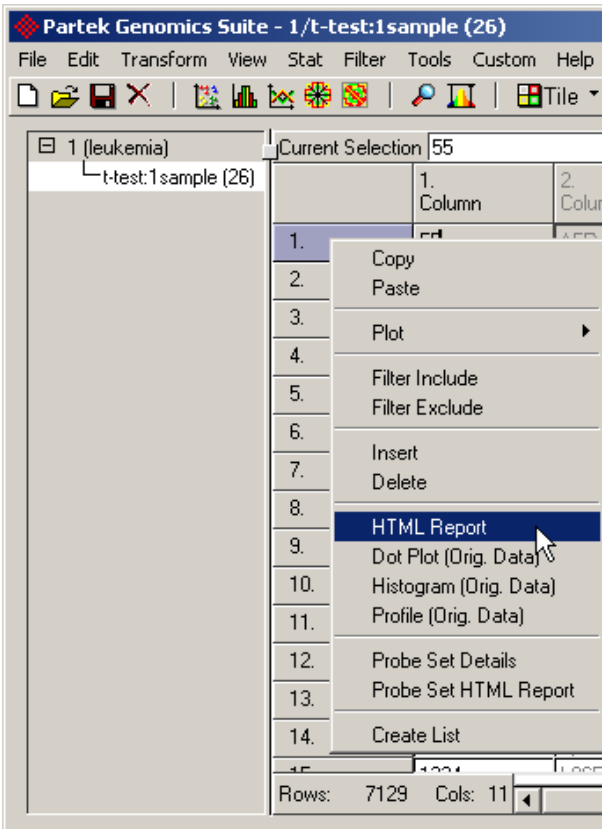


Figure 11. 8: Selecting an HTML report from the result spreadsheet for a multiple test of a One-Sample t-test

To test a single response variable, select the variable from the drop-down list; the result will be displayed in a HTML report (Figure 11. 9).

One Sample t-test of "leukemia" on AFFX-BioB-5_at

Response Variable: AFFX-BioB-5_at
Hypothesized Mean: 0

Descriptive Statistics

N	Mean	Std. Dev.	Std. Err.	Minimum	Maximum
72	-114.583	97.7383	11.5186	-476	86

t-test

t Statistic	DF	p-value (2-tailed)
-9.9477	71	4.27279e-015

Figure 11. 9: Viewing the report of the One Sample t-test, single test

Running the Computations

OK will perform the configured *t*-test computation and dismiss the dialog.

Apply will perform the configured *t*-test computation, but the one-sample t-test dialog will remain to allow for another computation.

Cancel will close the dialog without doing any computation.

One-Sample *z*-test

Introduction

The one-sample *z*-test is used to test for a difference in means between a sample group and a hypothesized population with a specified standard deviation. Like many other parametric tests, the most important assumption is that the data is normally distributed. The *z*-test is used instead of the one-sample *t*-test when the population standard deviation is known.

Configuring the One-Sample *z*-test Dialog

Open the one-sample *z*-test dialog by selecting **Stat > Parametric Tests > One-Sample z-test…** from the Partek main window (Figure 11. 11). You will use this dialog to specify the hypothesized mean, population standard deviation, the response variable(s) to be tested, and multiple test corrections. Figure 11. 10 is configured to compare a sample mean to 0 with the standard deviation as 1 on all the numeric variables.

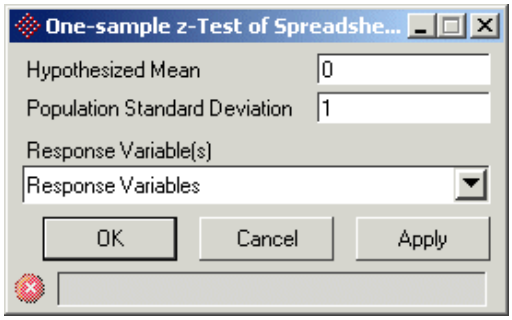


Figure 11. 10: Configuring the One-Sample z-test dialog, multiple tests

Selecting Response Variables

By default, if there is more than one numeric variable in the spreadsheet, *Response Variables* will be shown as the *Response Variable(s)* to test all of the variables at one time. To choose a specific response variable to test, select the variable name from the drop-down list; however, if there is only one numeric variable in the spreadsheet, the variable name will be selected as the *Response Variable* by default.

When an analysis is performed on all numerical variables, the results will be summarized in a new spreadsheet that is a child of the original. In the results

spreadsheet, each variable tested in a row is summarized by the column number and name of the variable, and followed by summary statistics including the p-values, means, and standard deviations. The rows are automatically sorted by the first column of p-values. To sort by a different p-value, right click on the column heading and select **Sort Ascending** in the pop-up menu. Detailed reports about individual test variables can be viewed by right-clicking on the row label corresponding to the variable and selecting the **HTML Report** option on the pop-up menu (Figure 11. 11).

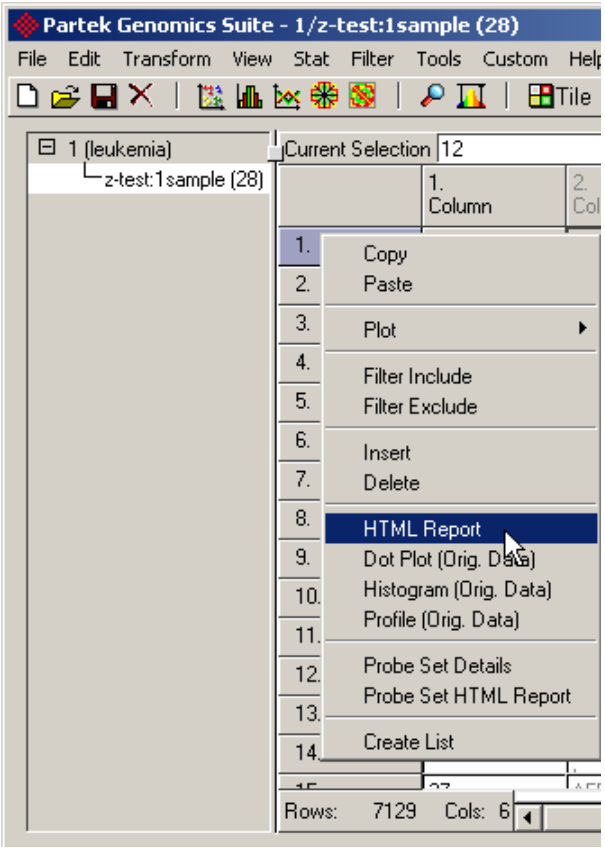


Figure 11. 11: Selecting an HTML report from the results spreadsheet for a multiple test of a One-Sample z-test

To test a single response variable, select the variable from the drop-down list; the result will be displayed in a HTML report (Figure 11. 12).

One Sample z-test of "leukemia" on AFFX-BioB-5_at

Response Variable: AFFX-BioB-5_at
Hypothesized mean: 0
Population Standard Deviation: 1

Descriptive Statistics

N	Mean	Std. Dev.	Std. Err.	Minimum	Maximum
72	-114.583	97.7383	11.5186	-476	86

z-test

z Statistic	p-value (2-tailed)
-972.272	2.57825e-203

Figure 11. 12: Viewing the report for the One-Sample z-test, single test

Running the Computations

OK will perform the configured *z*-test computation and dismiss the dialog.

Apply will perform the configured *z*-test computation, but the one-sample *z*-test dialog will remain to allow for another computation.

Cancel will close the dialog without doing any computation.

Analysis of Variance

Introduction

Analysis of variance (ANOVA) is a family of statistical models used to test for the difference in means of a response variable between different groups. Because ANOVA is a *parametric* test, it makes certain assumptions about the distribution of the response variable. The most important assumptions are that the data is normally distributed and that the variance is approximately equal between the groups (homogeneity of variance). Although ANOVA is most powerful when these assumptions are met, in many cases ANOVA is very robust to violations of these assumptions.

Implementation Details

Sir Ronald Fisher first developed analysis of variance in 1925. Many intermediate statistical textbooks serve as an introduction to analysis of variance (e.g. Steel and Torrie (1980) or Snedecor and Cochran (1980)). Scheffé (1959) is a classic reference for analysis of variance.

Partek's ANOVA can handle:

- a balanced and an unbalanced design
- random and fixed effects (mixed-model ANOVA), nested factors

- any number of categorical effects (multi-way ANOVA)
- numeric covariates (multi-way Analysis of Covariance, or ANCOVA)

Examples of each are provided below.

Example of a Balanced Experimental Design

A design is balanced when the number of samples is the same for each factor level. Consider the two factors, Treatment and Time. This is referred to as a 2X6 experiment design because Treatment has 2 levels and Time has 6 levels.

Factor	Levels
Treatment	Control, Treated
Time	T1, T2, T3, T4, T5, T6

Time vs. Treatment

Time\Treatment	Control	Treated	Total
T1	3	3	6
T2	3	3	6
T3	3	3	6
T4	3	3	6
T5	3	3	6
T6	3	3	6
Total	18	18	36

Figure 11. 13: An example of a balanced experimental design

Figure 11. 13 shows a balanced experimental design; in this case, a two-way crossed ANOVA. Every level of the factor Time occurs with every level of the factor Treatment. This is a balanced design because all of the levels of the two factors have the same number of samples (3).

An Example of Unbalanced Experimental Design

A design is unbalanced when the number of samples is not the same for each factor level. Below, in the Time and Treatment example, when a subject died at T6 (Time 6), the experiment became unbalanced (Figure 11. 14).

Time vs. Treatment

Time\Treatment	Control	Treated	Total
T1	3	3	6
T2	3	3	6
T3	3	3	6
T4	3	3	6
T5	3	3	6
T6	3	2	5
Total	18	17	35

Figure 11. 14: An example of an unbalanced experimental design

Missing Values & Missing Treatment Combinations

If the levels of all the factors are completely crossed (Figure 11. 13, Figure 11. 14), Type III sums of squares is used. However, if missing treatment combinations occur in any interaction, Type IV sums of squares is used.

A missing treatment combination occurs when one of the cells in the multi-way ANOVA table has no entries. Consider all three treated samples at time T6 are not available; the Treated X T6 combination is missing (Figure 11. 15).

Time vs. Treatment

Time\Treatment	Control	Treated	Total
T1	3	3	6
T2	3	3	6
T3	3	3	6
T4	3	3	6
T5	3	3	6
T6	3	0	3
Total	18	15	33

Figure 11. 15: An example of a missing treatment combination

If an interaction corresponding to a treatment combination has no replication, Partek will automatically remove that interaction from the model. Therefore, when testing multiple response variables, the p-values of the removed interactions will be represented by question marks ("?") to indicate that the value could not be computed, and a message that the interaction is removed for that variable will be displayed at the bottom of the HTML report.

Mixed Model ANOVA

To obtain estimates of variance components for mixed models, Partek has the method of moments estimation (Eisenhart, 1947), restricted maximum likelihood estimation (REML) (W.A. Thompson, 1962), and minimum variance quadratic unbiased estimation (MIVQUE) (Rao, 1971).

The method of moments is used to equate analysis of variance mean sum of squares to their expected values ($s=C\sigma^2$). S is a vector of the mean sum of squares, C is a matrix, and σ^2 is a vector of variance components. The estimates of σ^2 are $C^{-1}s$. However, the method of moments method can produce negative estimates.

The MIVQUE method gives minimum variance quadratic unbiased estimates that are invariant with respect to the fixed effects of the model. Given the initial prior values $r_i^2=0$ where i=1,…, k and $rk_{+1}^2=1$, k is the number of random variables. The MIVQUE of σ^2 are obtained as a solution of the linear system of equations: $S\ \sigma^2=q$ where $S=\{s_{ij}\}$ is a (k+1)*(k+1) symmetric matrix, $q=\{q_i\}$ and $\sigma^2=\{\sigma^2_i\}$ are (k+1) vectors. $s_{ij}=SSQ\ (X_i'MX_i)$ is computed, where $M=I-X_0(X_0'\ X_0)^-\ X_0'$ and X_0' is part of the design matrix for the fixed effects, X_i is part of the design matrix for the random effects and SSQ is an operator that takes the sum of squares of the

elements. The estimators obtained by MIVQUE are functions of priori values used. The minimality property applies only at these a priori values.

The restricted maximum likelihood method optimizes the parameter estimates for the effects in the model. REML only maximizes the likelihood function on random effects. The procedure uses a Newton-Raphson algorithm, iterating until convergence is reached for the log-likelihood function of the portion of the likelihood that does not contain the fixed effects. The objective function for REML is $Ln(|V|)+r'V^{-1}r + Ln(|X_0'V^{-1}X_0|)$, where $r=y-X_0(X_0'V^{-1}X_0)^{-}X_0'V^{-1}y$, $V= \sigma^2_0 I+ \Sigma_i\sigma^2_i X_i X_i'$, where σ^2_0 is the residual variance, i=1,…,n and n is the number of random effects in the model. σ^2_i represents the variance components; X_i is part of the design matrix for one of the random effects.

Configuring the ANOVA Dialog

Open the ANOVA dialog by selecting **Stat > ANOVA…** from the Partek main menu. The ANOVA dialog is shown in Figure 11. 16.

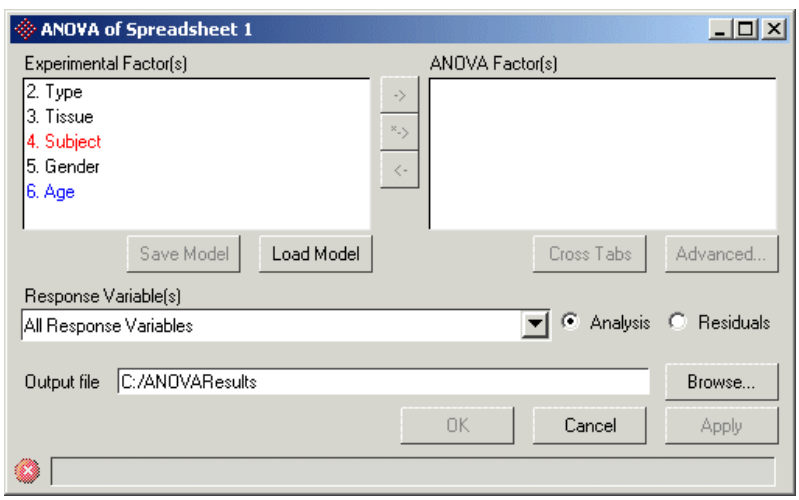


Figure 11. 16: Configuring the ANOVA dialog

This dialog is used to specify the factor(s), the response variable(s), and any advanced settings, such as interactions, post-hoc analysis. In the *Experimental Factor(s)* box, if the categorical variable is specified as a random effect, it will be in red. The candidate variables also include numeric variables. If their attributes are factor, they are labeled in blue; they are candidates of covariate in the ANCOVA model. Figure 11. 17 is configured to perform a 3-way ANOVA to test the difference among the categories in Type, Tissue, and Subject factors on all of the numeric variables. This model includes both fixed and random effects. Three methods of variance component estimation are shown with the default as *Method of Moments*.

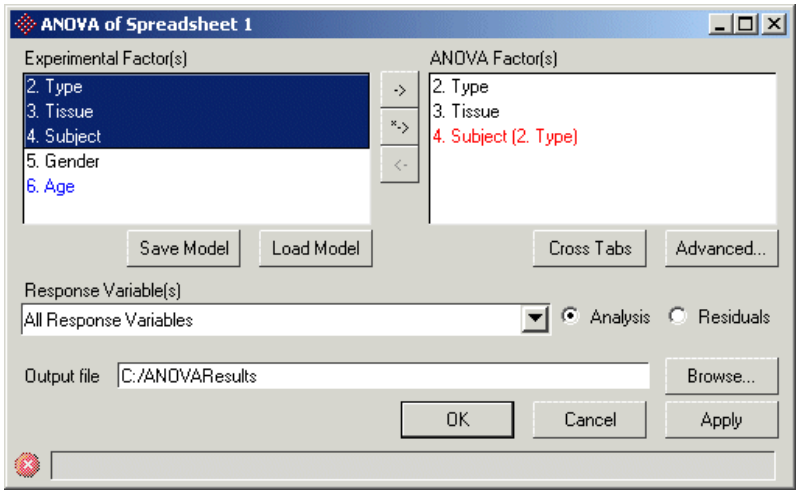


Figure 11. 17: Three-way Mixed Model ANOVA configuration

Selecting Factor(s)

In order to run ANOVA, there must be at least one categorical variable in the spreadsheet. The *Factor(s)* (or grouping variables) must be selected from the *Experimental Factor(s)* list, which contains all categorical variables and numeric variables whose attributes are factor of the spreadsheet. There are no software imposed limits on the number of *Factors* in Partek's ANOVA. A 1-way ANOVA is used for one *Factor*, a 2-way ANOVA is used for two *Factors*, a 3-way for three *Factors*, and so on.

When there is at least one item in the *Experimental Factor(s)* list box selected, the -> button next to the *ANOVA Factor(s)* list box will be enabled. Click on it to move the selected item(s) into the *Factor(s)* list box. The order of the items in the *ANOVA Factor(s)* list box is based on the column number. To remove a factor, select it in the *ANOVA Factor(s)* list box, the <- button next to it will be enabled. Click on the <- button and the item(s) selected in the *Factor*(s) list box will be moved. Double clicking on one item at a time will also move it to the other list box.

Note: if a factor is removed from the *Factor(s)* panel, *Contrast* and *Results* information related to that factor will be deleted.

Selecting Response Variables

Figure 11. 18: Selecting a response variable from the drop-down list

A list of all numeric variables in the spreadsheet is shown in the *Response Variable(s)* drop down list. By default, *All* will be shown as the *Response Variable(s)* if there is more than one numeric variable in the spreadsheet. To choose a specific response variable to test, select the variable name from the drop-down list.

By default, the *Analysis* button is selected to produce ANOVA results; if *Residuals* is selected, however, the residual of the configured ANOVA model will be calculated on the selected response variable. The result will be in a child spreadsheet that has the same format as the original spreadsheet, by default, the result file is saved in the same folder as the original file; the file is called *ANOVAResults*. Use the *Browse* button store the file in a different folder. (Figure 11. 19).

When an analysis is performed on all numerical variables, the results will be summarized in a new spreadsheet that is a child of the original (Figure 11. 19). Each row of the *Results Spreadsheet* corresponds to one of the numeric columns in the parent spreadsheet.

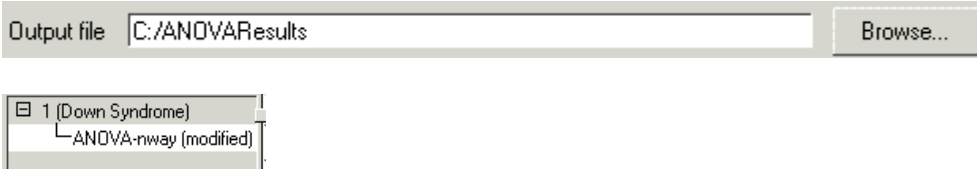


Figure 11. 19: Viewing the spreadsheet hierarchy, the ANOVA result spreadsheet

In the *Results Spreadsheet*, each variable is tested in a row and is summarized by the column number and name of the variable. It is followed by summary statistics including the p-values for each factor. When more than one factor is used, the first factor in the *Grouping Variable(s)* box corresponds to the first column of p-values (Figure 11. 17). The first column of p-values automatically sorts the rows; however, to sort by a different p-value, simply right click on the column heading and select **Sort Ascending** in the pop-up menu. Detailed reports about individual test variables can be viewed by right-clicking on the row label corresponding to the variable and selecting the **HTML Report** option on the pop-up menu (Figure 11. 20).

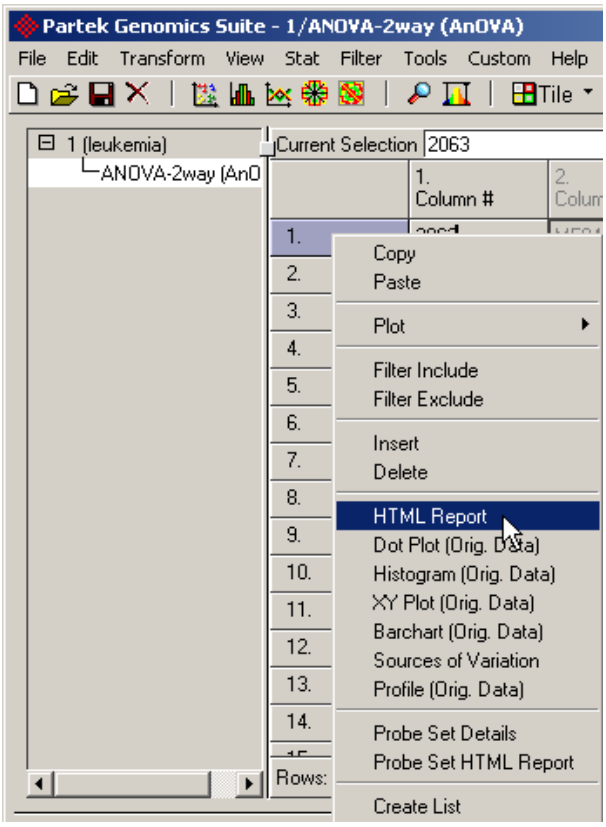


Figure 11. 20: Selecting an HTML report from the three-way ANOVA result spreadsheet

To test a single response variable, select the variable from the drop-down list; the result will be displayed in a HTML report (Figure 11. 21).

ANOVA-nway report of variable 200677_at

Grouping Variable(s): "Type" "Tissue" "Subject"
Response Variable: "200677_at"

Grouping Variable Information

Grouping Variable(s)	Levels	Level Values
Type	2	Normal, TS21
Tissue	4	astrocyte, cerebellum, cerebrum, heart
Subject	10	1218, 1389, 1390, 1411, 1478, 1479, 1521, 1565, 748, 847

Descriptive

Type	N	Mean	Std. Dev.	Std. Err.	Minimum	Maximum
Normal	14	970.571429	624.669831	166.950035	462.400000	2396.500000
TS21	11	1716.572727	1160.521463	349.910387	614.100000	3972.800000

Tissue	N	Mean	Std. Dev.	Std. Err.	Minimum	Maximum
astrocyte	4	3125.750000	892.244124	446.122062	2314.500000	3972.800000
cerebellum	6	1003.933333	200.165458	81.717206	799.400000	1271.100000

Figure 11. 21: Viewing an example report for the ANOVA, single test

Crosstabulations

Click on the **Crosstabulations** button in the *ANOVA* dialog box to show the combined frequencies of the samples for the two-factor pairs in the *Grouping Variables* spreadsheet (Figure 11. 22).

Crosstabulations

Factors: Type; Tissue; Subject;
Main Effects: Type; Tissue;
Interactions: Subject*Type; Tissue*Type; Subject*Tissue*Type;

Tissue vs. Type

Tissue\Type	Normal	TS21	Total
astrocyte	2	2	4
cerebellum	3	3	6
cerebrum	7	4	11
heart	2	2	4
Total	14	11	25

Subject vs. Type *

Subject\Type	Normal	TS21	Total
1218	0	3	3
1389	0	2	2
1390	4	0	4

Figure 11. 22: Viewing the crosstabulations report

Settings for the ANOVA Model

Partek's ANOVA is very robust, allowing specification of main effects, interactions, and covariates.

Nested Factors

In a two-way or a multi-way ANOVA, if each level of one factor occurs with each level of another factor, the factors are said to be "crossed". If, however, the levels of one factor only occur within a single level of another factor, then one factor is said to be "nested in" the other factor.
In the example below, there are two factors: Type and Subject ID. Type has 2 levels and Subject has 10 levels.

Factor	Levels
Type	Normal, TS21
Subject	1218, 1389, 1390, 1411, 1478, 1479, 1521, 1565, 748, 847

Notice in the Crosstabulations table (Figure 11. 23) that each level of Subject occurred within one level of Type. Therefore, Subject is nested within Type.

Subject vs. Type *

Subject\Type	Normal	TS21	Total
1218	0	3	3
1389	0	2	2
1390	4	0	4
1411	4	0	4
1478	0	4	4
1479	1	0	1
1521	4	0	4
1565	1	0	1
748	0	1	1
847	0	1	1
Total	14	11	25

* Subject is nested in Type.

Figure 11. 23: An example of a nested factor

The notation of Subject is nested in Type is Subject(Type)

Interactions

An interaction is the variation among the differences between means for the different levels of a factor (grouping variable) over different levels of the other factor. When there is more than one factor in the *Experimental actor(s)* list box selected, the *-> button will be enabled to allow configuration of the interactions in the dialog (Figure 11. 24).

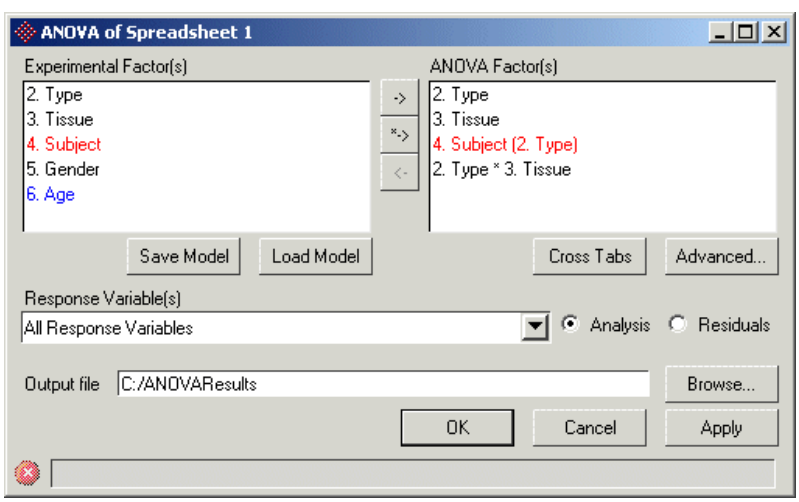


Figure 11. 24: Configuring the interaction dialog

When a main effect or interaction is removed from the *ANOVA Model* panel, any previously configured *Contrasts* and/or *Results* information that was related to the main effect or interaction will be removed. However, if another interaction is added, the previous configured *Contrast* and *Results* configuration will remain.

Random vs. Fixed Effects

Most factors in an analysis of variance are fixed factors, i.e. the levels of that factor represent all the levels of interest. Examples of fixed factors include gender, race, strain, etc. However, in experiments that are more sophisticated a factor can be a random effect, meaning the levels of the factor only represent a random sample of all of the levels of interest. Examples of random effects include subject and batch. Consider the example where one factor is *type* (with levels *normal*, *diseased*), and another factor is *subject* (the subjects selected for the experiment). In this example, *type* is a fixed factor since the levels *normal* and *diseased* represent all conditions of interest. *Subject*, on the other hand, is a random effect since the subjects are only a random sample of all the levels of that factor.

Specifying a Categorical Variable as a Random Effect

To specify a categorical variable as a random effect, in the Partek spreadsheet, right click on the column in the spreadsheet and select *Properties* (Figure 11. 25), and check the *Random Effect* box. In the *Main Effect* page of the *ANOVA Model Configuration* dialog, the random effect will be shown in red (Figure 11. 24). By saving the spreadsheet, Partek will automatically know which factors are random and which are fixed when they are used in an ANOVA model.

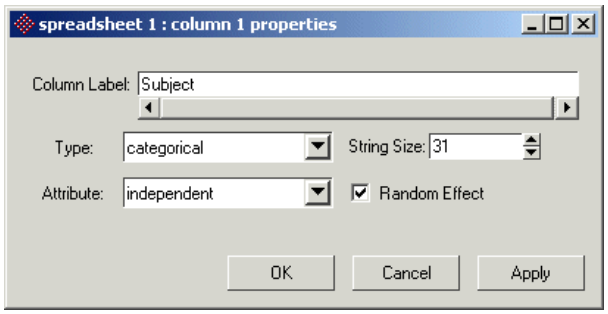


Figure 11. 25: Configuring the Column Properties dialog of spreadsheet 1

Analysis of Covariance (ANCOVA)

A covariate is an extraneous variable such as gender, age, or race that might be a significant source of variability but was not controlled during the experiment. For example, if testing for a difference between a diseased group of subjects and a normal group of subjects, 70% of the diseased subjects are male and only 40% of the normal subjects are male; gender is partially confounded with the disease state. In other words, some of the difference between the normal and diseased groups could be due to gender. Gender is a confounding nuisance variable, or a variable that could statistically distort the results, and should be accounted for by using analysis of covariance (ANCOVA). If the confounding variable is a categorical variable, include it in the model as another factor. For instance, in the above example, the one-way ANCOVA on disease type with gender as a covariate is the same as a 2-way ANOVA on both disease type and gender.

To specify a numeric variable as a covariate, right click on the variable column header on the spreadsheet; click **Properties** to change the attribute of the column to

factor (Figure 11. 26). In the ANOVA dialog, the numeric variable will be automatically included in the *Candidate Variable(s)* list and labeled in blue. Include this variable in *Grouping Variable(s)*. In Partek, there is no limitation on the number of covariates, and the covariates can interact with other factors. In addition, any number of covariates can be selected. The partial correlation of a covariate can be found on the result spreadsheet. The partial correlation can be calculated as the square root of R^2-r^2. The sign is from the coefficient of the covariate in the model.

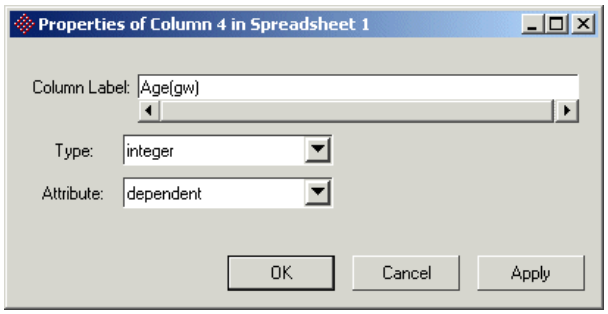


Figure 11. 26: Configuring a variable's attribute as a factor

Configuring the ANOVA Results

By default, only p-values and F ratios of the factors/interactions are displayed in the ANOVA result spreadsheet. To present more information about ANOVA, click on the **Advanced…** button.

Contrasts

The *Contrast* page allows performing a linear contrast between two specific groups within the context of ANOVA (Figure 11. 27).

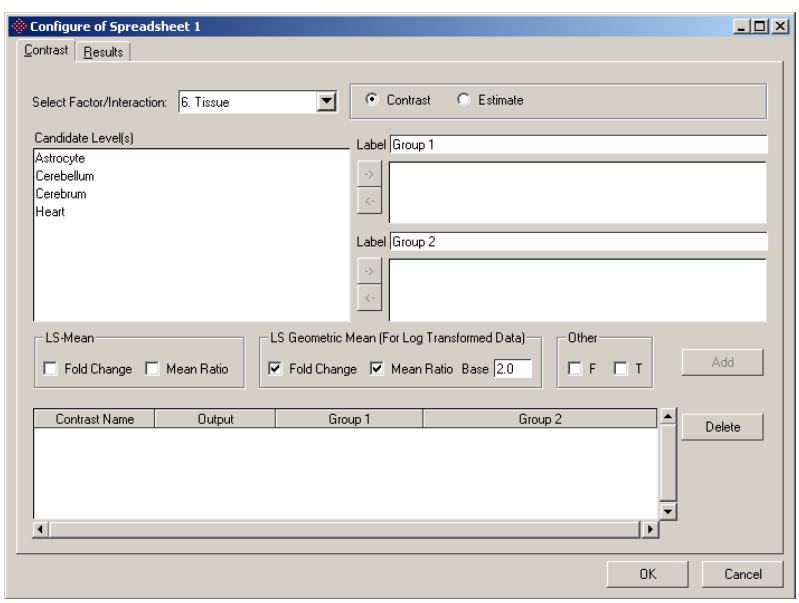


Figure 11. 27: Configuring the contrast of the mean of different tissue

Specify the factor or interaction to perform the contrast (Figure 11. 28).

Figure 11. 28: Selecting the Factor or Interaction to perform the contrast

If **Contrast** is selected (Figure 11. 29), you must have two groups to compare. The coefficients of the levels in the two compared groups will add up to 0; if **Estimate** is selected, it can be calculated on one or two groups. The coefficients of all the levels in the group(s) might not add up to 0.

Figure 11. 29: Selecting Contrast or Estimate

The *Candidate Level(s)* list all the levels (subgroups) of the selected factor or interaction (Figure 11. 30).

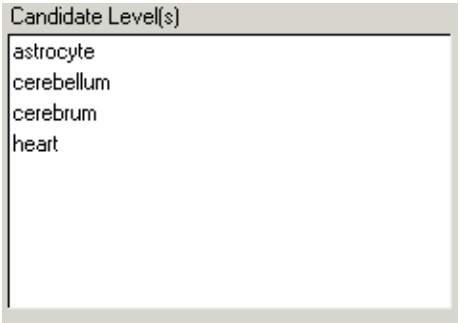


Figure 11. 30: Viewing the selected Candidate Level(s)

Select the levels and click the **->** button to move them to *Group1* or *Group2*. The label of the groups can be changed (Figure 11. 31).

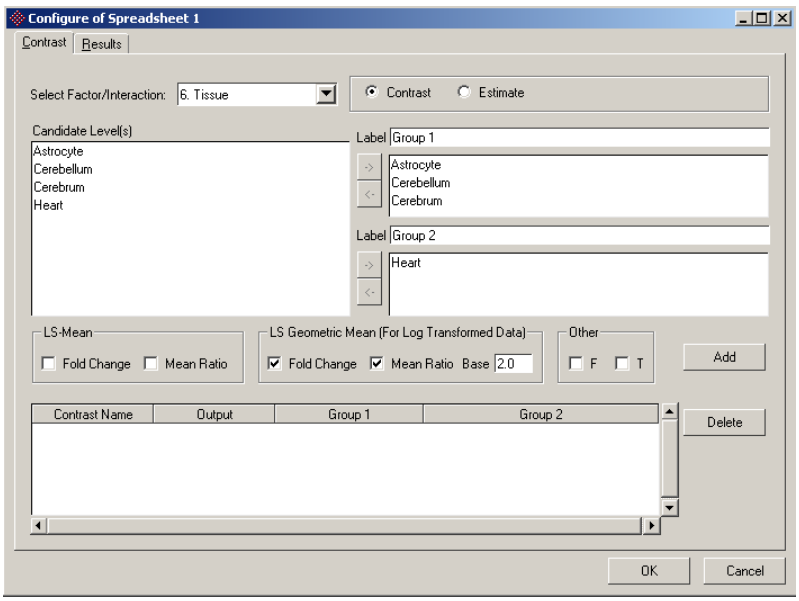


Figure 11. 31: Specifying the groups to compare

Figure 11. 31 shows all the brain tissue grouped together to compare to the heart tissue. The labels of the groups were changed into *Brain* and *Heart*, respectively.

- Click **Add** to add the specified contrast to the bottom panel

If **Contrast** is selected, the coefficients of levels in Group1 (Drug) are added up to 1, and the coefficients of levels in Group2 (Control) are –1 (Figure 11. 32).

Contrast Name	Output	Group 1	Group 2
Brain vs. Heart	p-value FoldChange MeanRatio	1/3 astrocyte+1/3 cerebellum+1/3 cerebrum	-1 heart

Figure 11. 32: Viewing the contrast between the Brain and Heart

If **Estimate** is selected, the coefficient of each level in Group1 (Brain) is 1, and the coefficient of each level in Group2 (Heart) is –1 (Figure 11. 33).

Contrast Name	Output	Group 1	Group 2
Brain vs. Heart	p-value Estimate	1 astrocyte+1 cerebellum+1 cerebrum	-1 heart

Figure 11. 33: Viewing the estimate between the Brain and Heart

Since all the levels in each group have the same weight, specify double level weight in a group by dragging it to the group twice.

To remove the specified contrast in the bottom panel, click **Delete**. To select more than one item, click the left-mouse button and drag or press **<Ctrl>** and left click, then click **Delete**.

By default, the p-values of each Contrast and/or Estimate are calculated and put out as a column on the *ANOVA Result* spreadsheet. In *Contrast*, you can configure *mean ratio* and *fold change* of the two groups; in *estimate*, you can configure *estimate value*.

The computations of p-values are based on LS-means (Least-squares means), which are the means adjusted by other factors.

Contrast Equations

When contrasting the average of treatments A and B versus treatment C, the contrast equation is

$$\frac{1}{2}A+\frac{1}{2}B-1C=0$$

The ratio is calculated using the least square mean (LS Mean) of each term, thus the ratio for the contrast is given by:

$$\frac{\frac{1}{2}LSMean(A)+\frac{1}{2}LSMean(B)}{LSMean(C)}$$

A second example contrasts the average of treatments A, B, and C versus the average of treatments C and D. The contrast equation is

$$\frac{1}{3}A+\frac{1}{3}B+\frac{1}{3}C-\frac{1}{2}D-\frac{1}{2}E=0$$

The ratio for the contrast is given by:

$$\frac{\frac{1}{3}LSMean(A)+\frac{1}{3}LSMean(B)+\frac{1}{3}LSMean(C)}{\frac{1}{2}LSMean(D)+\frac{1}{2}LSMean(D)}$$

Fold changes are calculated in a similar fashion using LSMeans.

LS Mean and Geometric Mean

The LS Mean (Least Squares Mean) is calculated as the linear combination (sum) of the estimated means from a linear model (e.g. ANOVA, regression, etc). The LS mean is based on the factors specified in the model, thus, the LS mean is "model dependent" whereas arithmetic mean is "model independent". When the data results from a balanced experiment (same number of treatment combinations in each group), the arithmetic mean and LS mean are identical. In unbalanced data, the arithmetic mean and LS mean are different. In an unbalanced experiment, the LS means are preferred because they reflect the model being fit to the data.

Consider a simple unbalanced two-factor experiment containing a control group and a treated group, with unequal number of male and female animals in each group (Figure 11. 34). The control group contains 4 females and 2 males, and the treated group contains 2 females and 5 males.

Treatment vs. Gender

Treatment\Gender	Female	Male	Total
Control	4	2	6
Treated	2	5	7
Total	6	7	13

Figure 11. 34: Unbalanced two-factor experiment crosstabulation

Using the arithmetic mean to estimate the means of the control and treated groups ignores the imbalance of male and female in the two groups and may be biased. For example, if you are estimating the effects of a gene's expression which lies on Y chromosome (females don't have a Y chromosome, thus they will have lower expression than males on this gene), the arithmetic mean would overestimate the

mean in treated group since the treated group contains more males, and underestimate the control group since the control group contains more females (Figure 11. 35).

Arithmetic Mean

Treatment	N	Mean
Control	6	54.366540
Treated	7	255.856519

Least Squares Mean

Treatment	LSMean
Control	91.120693
Treated	208.601179

Figure 11. 35: Comparison of the arithmetic mean and LS mean in the control and treated group of gene's expression that lies on Y chromosome

The LS mean uses the estimates for both factors in the design, treatment, and gender, and adjusts the means for the treated and control groups to account for the imbalance in gender between the groups. The LS mean would produce a more accurate, unbiased estimate of the mean of the treated and control groups in this example (Figure 11. 35).

Data is often log transformed prior to doing statistical analysis in order to transform a multiplicative effect into an additive effect. However, scientists sometimes want to interpret effects as ratios, in which case log transformed data is inappropriate, since it has been converted from a multiplicative effect to an additive effect. Simply anti-logging the mean of logged data does not produce the mean of the un-logged data; however, it does produce the geometric mean of the un-logged data. Anti-logging a least squares mean produces a value that we call a "least squares geometric mean".

When a ratio is calculated based on LS means, the ratio of Group1 vs. Group2 is:

$$\frac{LSMean(Group1)}{LSMean(Group2)}$$

When a ratio is calculated based on least squares geometric means, the estimate of the LS means on logged data is first calculated for each group, and the difference of the LS means is then anti-logged using the same base:

$$a^{LSMean(Group1)-LSMean(Group2)}$$

"***a***" is the base the data is log transformed on. This is equivalent to calculating the ratio of the least squares geometric means for the two groups:

$$\frac{a^{LSMean(Group1)}}{a^{LSMean(Group2)}}$$

This ratio is more appropriate than the simple ratio of LS means in the case that analyses have been performed on logged data.

Results

The *Results* page allows customizing of the items to show on the result spreadsheet in addition to p-values and the F ratio of the factors/interaction in ANOVA (Figure 11. 36).

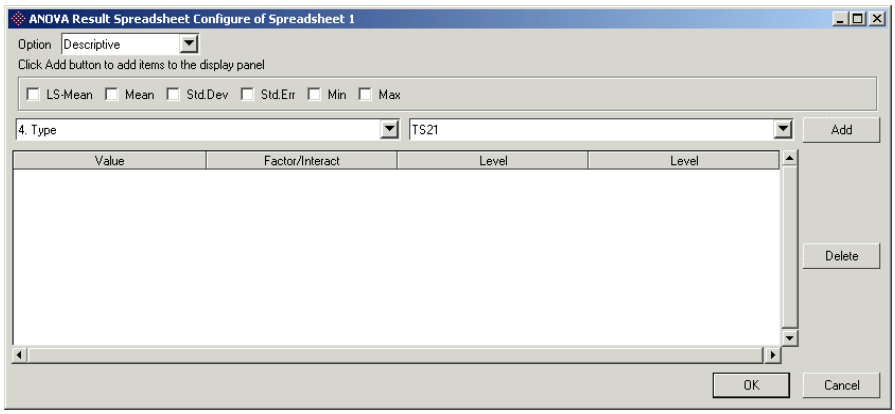


Figure 11. 36: Configuring the ANOVA result spreadsheet dialog

The *Option* drop-down list contains *Descriptive, Pairwise Comparisons,* and *Source Information.*

Descriptive can add *mean, lsmean, std.dev, std.err, min,* and *max* values of one level at a time from a factor or interaction of factors. Means that have been corrected for imbalances in other variables are called *adjusted means* or *least squares means* or *lsmeans*. If the lsmean is un-estimable due to unbalanced data, "?" will occur instead of data.

Pairwise Comparisons can add *difference of lsmeans, p-value, std. err of difference, mean ratio* and *fold change* values between two levels from the same factor or interaction of factors. When ANOVA is used to test for a difference between three or more groups, a post hoc analysis can be performed to see which groups were different. For example, if there are three groups, A, B, and C, and ANOVA shows a difference in means, it is difficult to know if the difference was between A and B, A and C, and/or B and C. The pairwise comparison presentation of testing one response variable is different from testing all the response variables at a time. This allows for easier judgment of group difference.

Source Information can add *Mean Squares, Sum of Squares* and *Degree of Freedom* values of a factor or interaction of factors.

Model Information can add p-value of *Model, Adjusted R-square, R-Square, F of Model, DOF of Model, DOF of Error,* and *MS of Error,* which tells how good the model fits the data.

When *Descriptive* is selected, select the values that will be displayed on the result spreadsheet by checking the checkboxes. Two drop-down lists will be shown to

select a factor/interaction from the grouping variables selected, and a level of that factor/interaction; when *Pairwise Comparison* is selected, customize a value or a factor/interaction and two levels from that factor/interaction. Click **Add** to include the selected value (or values, if *All* is selected) of the configured level (or pair of levels) to the display panel. Figure 11. 37 shows the *p-value*, *Mean Ratio*, *Fold Change* and *Mean Difference* of TS21/control of factor *Type* on the results spreadsheet.

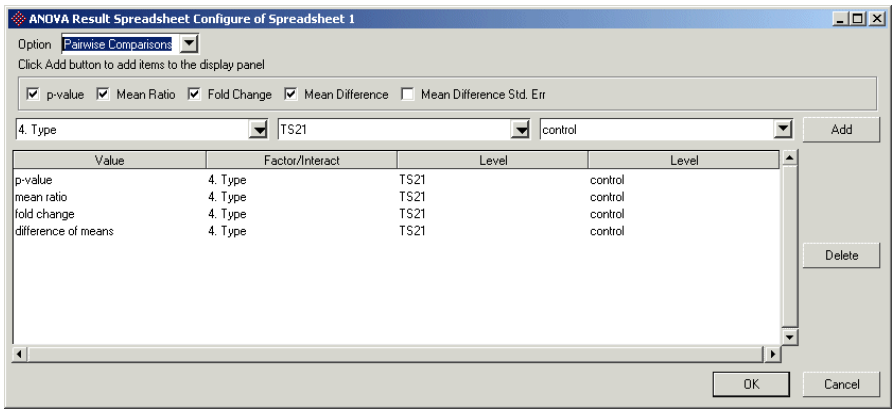


Figure 11. 37: Configuring to add TS21/control p-value, Mean Ratio, Fold Change, and Mean Difference on the result spreadsheet

The computations in the Pairwise Comparisons are based on least square means of each level.

When the *Source Information* option is chosen, a value and a factor/interaction need to be provided (Figure 11. 38).

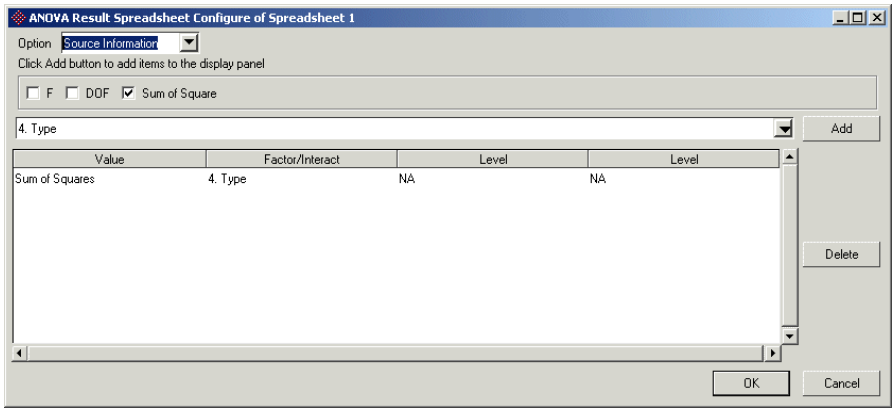


Figure 11. 38: Configuring to add the tissue sum of squares on the result spreadsheet

To remove the items selected on the display panel, select the item and click the **Delete** button. Press <Ctrl> or <Shift> while left clicking to choose multiple selections. Click **OK** to set the selection in the display panel and dismiss the dialog.

Method

If there is random effect factor(s) in the *ANOVA Factor(s)* panel, there will be *Method* page in the *Advanced* dialog (Figure 11. 39). There are three methods to estimate the amount of variance attributed to random effects. The methods are *Method of Moments*, *REML*, and *MINQUE*. The default method is Method of Moments. To change the method, select a different method and click **OK**.

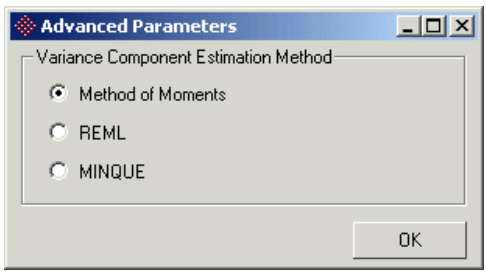


Figure 11. 39: Advanced Parameters Dialog

Running the Computations

Clicking **OK** or the **<Enter>** key will perform the configured ANOVA computation and dismiss the dialog. Clicking **Apply** will perform the configured ANOVA computation, but the *ANOVA* dialog will remain to allow for another computation.

Clicking **Cancel** or the **<Esc>** key will close the dialog without doing any computation.

The configured ANOVA model with all the results and contrast parameters can be saved as a .pam file. Click on **Save Model** and name the file (Figure 11. 40).

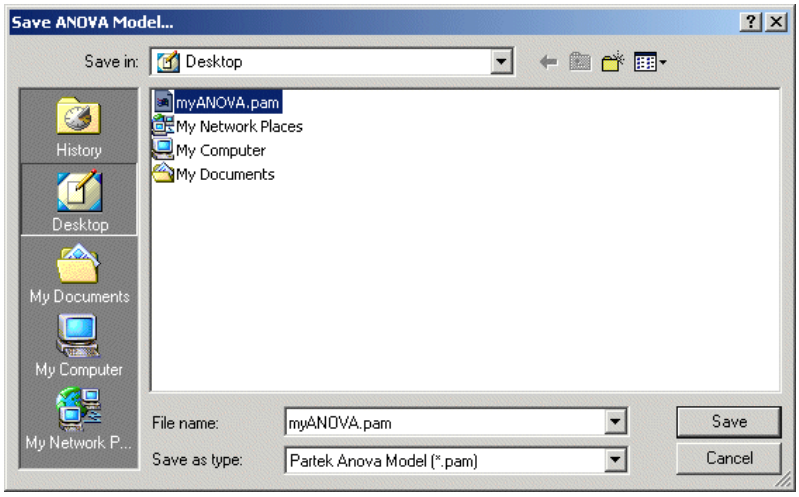


Figure 11. 40: Save ANOVA Model Dialog

If you want to perform the same ANOVA model on the same data set, you can click **Load Model** to load the saved model.

Note: Save Model only applies to spreadsheets that have the same sample information columns as the original spreadsheet from which the model was saved. All categorical variables and numeric factor variables should have the same column number and header on both spreadsheets.

Removing Batch Effects in Partek

Introduction

The Partek Batch Remover™ is used to remove the effects of nuisance batches or undesirable numeric or categorical factors from experimental data. In many cases, the effects, such as sample preparation batch and reagent lot, are large relative to the treatment effects being studied; therefore, the variability due to the batch effects may obscure the main effects. However, if the processing batches are included in the experiment design or are relatively well balanced with the treatments, their effects can be identified and removed from the data, and the true treatment effects can be revealed.

Assume the processing date has a significant effect on the expression values and you wish to remove this effect. A mixed model ANOVA is used to estimate the batch effects and the data is adjusted to what it would be if all batches were equal.

Therefore, the batch remover should be used as follows:

- Decide on the model that best fits your data, including treatment effects, and numeric or categorical technical effects
- Run the ANOVA
- Run the Batch Remover™ with the identical ANOVA model and identify those effects you wish to remove. You will have a choice to overwrite the data in the existing spreadsheet or to create a new spreadsheet with the batch-adjusted data
- On the batch-adjusted data, run the ANOVA again with the **exact** same model to make sure it worked perfectly. If it worked "perfectly" all p-values for the covariates, factors that have been removed should be 1.0. The p-values for the other factors should remain **exactly** unchanged

The batch remover works for balanced and unbalanced designs, and can remove any combination of numeric and categorical factors. It can even handle many incomplete designs (missing treatment combinations), but we recommend you check with step #4 above because it cannot work perfectly for some extremely sparse incomplete designs. If you find a design for which it does not work perfectly, please contact Partek customer support for assistance. For assistance in deciding which ANOVA model to use, see the **Analysis of Variance** section above

After a batch effect has been removed, that factor still must be included in ANOVA (or other statistical test) to account for the degrees of freedom used in the batch removal process. Another way to think about this is that ANOVA does

not need a batch remover - ANOVA removes batch effects by simply including the effects in the calculations. The usefulness of the batch remover is mostly for visualization purposes such as PCA, multidimensional scaling, clustering, etc. which do not have the ability to estimate or remove batch effects. The purpose of the batch remover is so that the visualizations will show treatment effects rather than technical effects. Visualizing batch adjusted data allows you to see the data the way that the multi-factor ANOVA model sees the data since the ANOVA model adjusts estimates for the batch effects in an identical way as the batch remover adjusts the data.

Since ratios and fold-changes computed in Partek's ANOVA are based on the least square means, they also don't need a batch removal, and will be identical whether the batch was removed or not.

Implementation Details

The batch remover works for balanced and unbalanced designs, and can remove any combination of numeric and categorical factors. It can even handle many incomplete designs (missing treatment combinations). You need to first decide the ANOVA model that will be used. After a batch effect has been removed, it still needs to be included in the ANOVA (or other statistical test) to account for the degrees of freedom used in the batch removal process. For assistance in deciding which ANOVA model to use, see the **Analysis of Variance** section above.

Configuring the Remove Batch Effects Dialog

Open the dialog by selecting **Stat > Remove Batch Effect…** from the Partek main menu (Figure 11. 41).

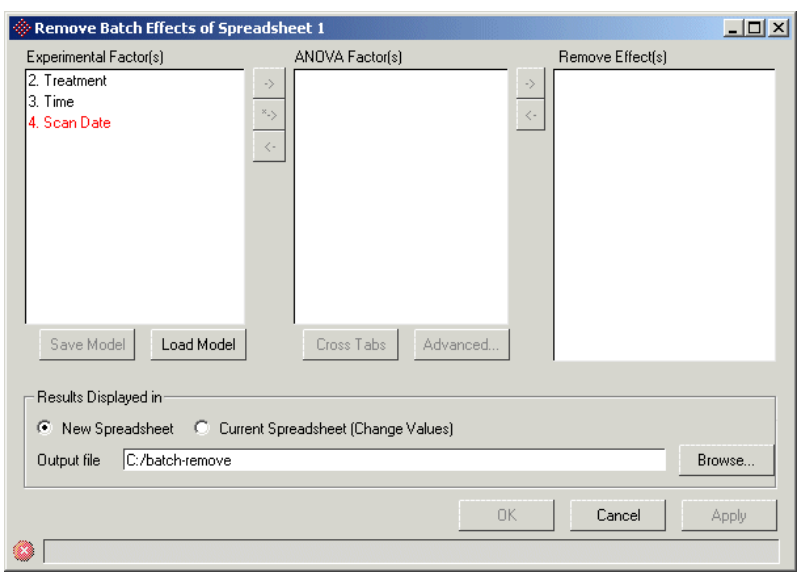


Figure 11. 41: Remove Batch Effects dialog

Select variables from *Experimental Factor(s)* and move them to *ANOVA Factor(s)* to configure the ANOVA model. For the details on how to configure the ANOVA dialog, see the **Analysis of Variance** section above. When an item in the *ANOVA Factor(s)* list box is selected, the -> button next to the *Remove Effect(s)* list box will be enabled. Click on the button to move the selected item into the *Remove Effect(s)* list box (Figure 11.). Click to select an item in the *Remove Effect(s)* list box and then click the enabled <- button to remove. Double clicking on an item will also move it to the other list box.

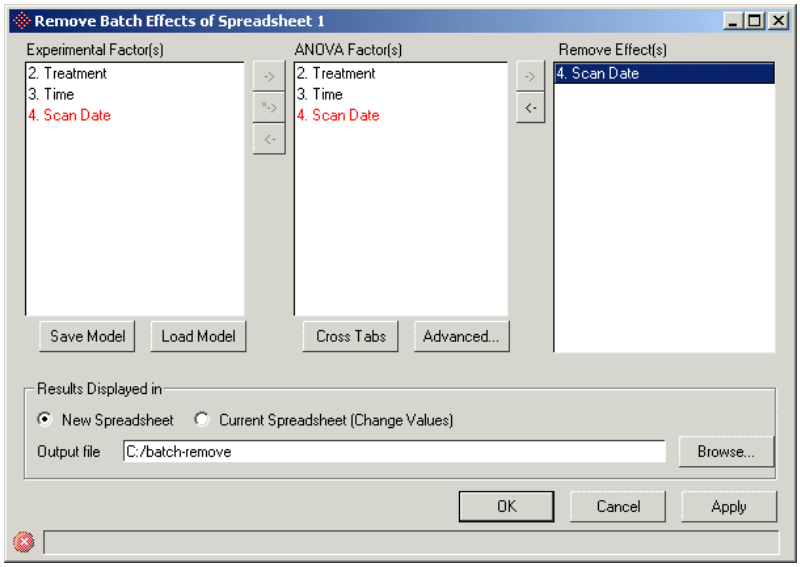


Figure 11. 42: Configuring the batch remover ; removing the scan date effect

Click **OK** or **Apply** to remove the batch effect on all numeric response columns of the spreadsheet. By default, the results will be displayed in a new child spreadsheet, you can specify the name of the result use the **Browse…** button; however, you can choose to change the values in the current spreadsheet instead (Figure 11. 43).

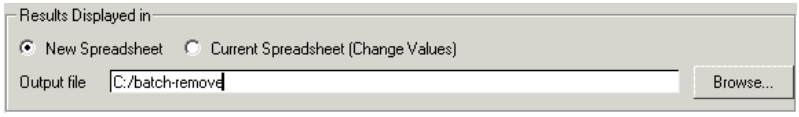


Figure 11. 43: Display the results in the current spreadsheet

Alternative Splice ANOVA

Introduction

One of the goals of exon array data analysis is to detect an alternative splicing event. There are two types of alter-splicing event for a gene, tissue independent and tissue dependent. In tissue independent alt-splicing, the different exons express differently regardless of tissues, treatments, diseases, etc. In tissue dependent alt-splicing, the exon's expression depends on different tissues, treatments, diseases,

etc. Figure 11. 48 shows an example of a gene with two exons –exon A and exon B. The height of the bar represents the expression level of the exon in disease tissue (D) and normal tissue (N). The first picture on the left shows no alternative-splicing, both exon A and B have the same expression level in both tissue types; the second picture shows that exon A has a higher expression than exon B in both tissue type, this gene demonstrates that alt-splicing is tissue independent; the two pictures on the right shows that the expression level of exons A and B depend on the tissue type.

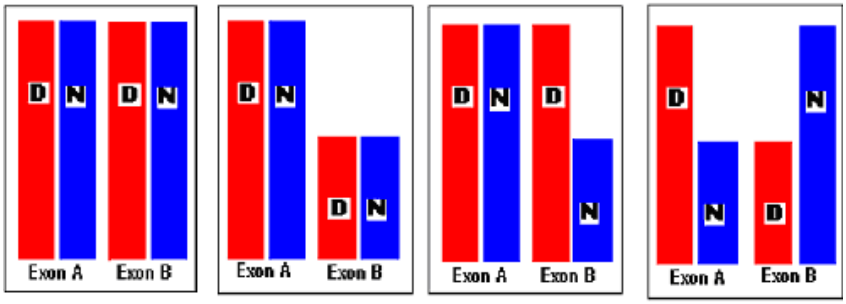


Figure 11. 44: Examples of different types of alternative splicing, the graphics show a gene that has two exons –Exon A, Exon B. The height of the bar represents the expression level, the red bar (D) represents Disease tissue, and the blue bar (N) represents Normal tissue. The first picture on the left shows now alternative splicing even; the second one from the left shows alternative splicing regardless of tissue; the 3rd picture from the left illustrates exon A has high expression in both tissue, but exon B has high expression only in disease tissue; the 4th picture shows exon A has high expression in disease tissue, but exon B has high expression in normal tissue, both the 3rd and 4th pictures demonstrates alternative splicing depends on tissue

Implementation Details

The method to detect alternative splicing is ANOVA, for detailed information about ANOVA, see the **Analysis of Variance** section above. The result is summarized at gene level based on the mean expression of all the exons in a gene. Besides all the factors specified in the ANOVA model by the user, extra terms added to the model by Partek automatically:

- Since not all exons in a gene express at the same level, exon ID is added to the model to account for exon-to-exon differences
- Interaction of exonID with the factor to detect alt-splicing is added to estimate an exon has different expression in different levels of the factor
- Since multiple measurements (on the multiple exons) come from the same sample, sample ID is added to the model, otherwise the ANOVA assumption of sample independence is violated

Suppose there is a paired designed experiment to find exons differentially expressed in two tissues, two different tissues are taken from each subject, paired sample t-

test, or 2-way ANOVA will be used to analyze the data. The alt-splicing ANOVA dialog allows the user to specify the ANOVA model, which include the two factors: tissue and patient ID, the analysis is performed at exon level, but the result is displayed at gene level. The equation of the model that the user specified is:

$$y = \mu + T + P + \varepsilon$$

y: expression of a gene
μ: average expression of the gene
T: tissue-to-tissue effect
P: patient-to-patient effect (this is a random effect)
ε: error term

When the alt-splicing factor is set as tissue, which means to detect exons that express differently depending on tissues, and when the **OK** or **Apply** button is pressed, the ANOVA model becomes as followings:

$$y = \mu + T + P + E + T * E + S(T * P) + \varepsilon$$

y: expression of a gene
μ: average expression of the gene
T: tissue-to-tissue effect
P: patient-to-patient effect (this is a random effect)
E: exon-to-exon effect (alt-splicing independent to tissue type)
T*E: an exon expresses differently in different tissue (alt-splicing dependent to tissue type)
S (T*P): sample-to-sample effect (this is a random effect, and nested in tissue and patient)
ε: is the error term

Alternative Splicing Score

If there are only two samples in the spreadsheet then Partek cannot calculate a type by probe set interaction. In this case, the result spreadsheet will contain a column labeled *Alt-Splice score*. First, for each probe set on the transcript calculate the difference between the two samples. The alt-splice score is the minimum p-value from the z-test of each probe set's difference against the rest. A low alt-splice score indicates that at least one probe set behaves differently from the rest.

Configuring the Alternative Splicing ANOVA Dialog

The dialog for alt splicing can be found by selecting **Stat > Alternative Splicing ANOVA…**. If this menu is not visible, make sure the active spreadsheet has the associated "exon" property to notify Partek that the spreadsheet has data appropriate for alternative splicing. Properties can be added to a spreadsheet by selecting **File > Properties...**.

The spreadsheet must have external link to exon level annotation file, transcript annotation file and annotation file that maps exon ID to transcript ID. If the links are not specified, the following dialog will be open (Figure 11. 45).

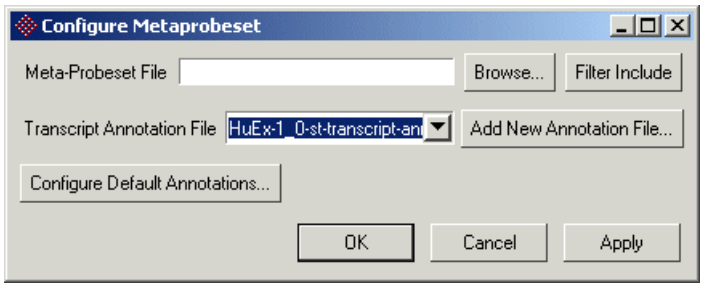


Figure 11. 45: Configure annotation files that link to the spreadsheet

Use the **Browse** button to specify the *Meta-Probeset File*, which maps the probe sets in the spreadsheet to gene; choose transcript annotation file from the drop-down list, this file contains information about the genes defined in the *Meta-Probeset File*. If it doesn't exist in the drop-down list, click **Add New Annotation File** to add it, click **OK**.

In the alternative splice ANOVA dialog, select variables from *Experimental Factor(s)* and move them to *ANOVA Factor(s)* to configure the ANOVA model. For the details on how to configure the ANOVA dialog, see the **Analysis of Variance** section above. When an item in the *ANOVA Factor(s)* list box is selected, the -> button next to the *Alternative Splice Factor(s)* list box will be enabled. Click on the button to move the selected item into the *Alternative Splice Factor(s)* panel (Figure 11. 46). Click to select an item in the *Alternative Splice Factor(s)* panel and then click the enabled <- button to remove. Double clicking on an item will also move it to the other list box.

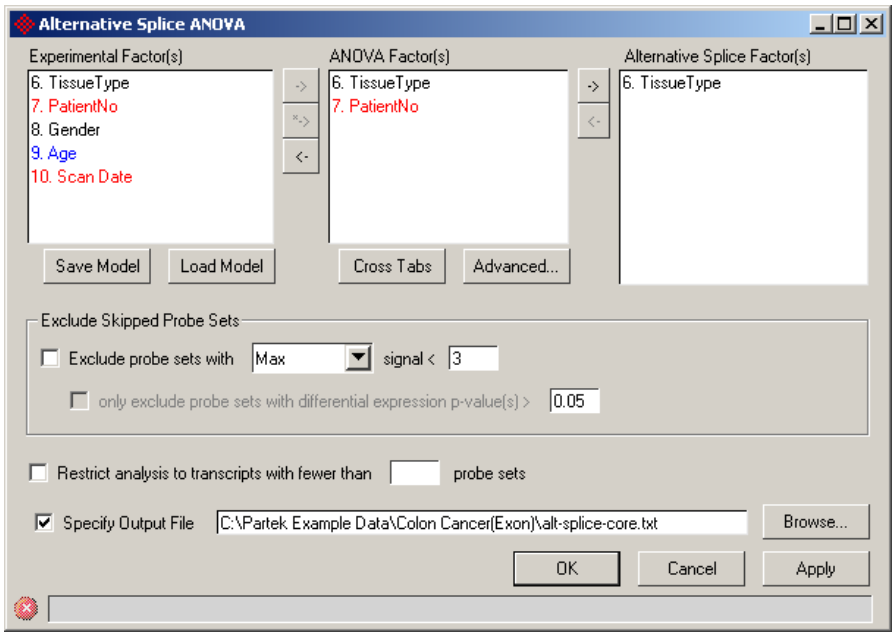


Figure 11. 46: Alt-splicing dialog, specify 2 way ANOVA model including tissue type and patient number, detect alt-splicing event depend on tissue type

Note: the alternative splice factor(s) have to be in the ANOVA model, and more than one factor can be specified in as alternative splice factors.

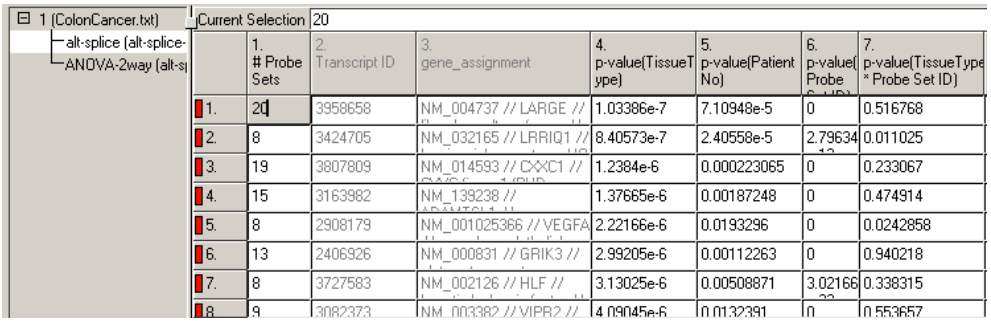
Use the **Browse…** button to specify the out put file name, by default the output file is called *alt-splice.txt*, stored in the same folder as the original file.

If an exon probe set has low expression and does not exhibit differential expression, then it is likely that it is simply skipped in all present samples. By setting the parameters in *Exclude Skipped Probe Sets*, you can avoid false positives in the alt-splicing result.

The analysis can be restricted to transcripts with fewer probe sets by select the check box and specify the maximum number of probe sets in a transcript (Figure 11. 47).

Figure 11. 47: Restrict the analysis to transcripts that have fewer than 40 probe sets

Click **OK** or **Apply** to compute. By default, the results will be displayed in a new child spreadsheet (Figure 11. 48). Each row represents a gene, number of exons (probe sets) in the gene, p-values and F ratios of all the factors and interactions in the ANOVA model for the gene are presented on the columns, and genes are sorted by the first p-values column.



	1. # Probe Sets	2. Transcript ID	3. gene_assignment	4. p-value(TissueType)	5. p-value(Patient No)	6. p-value(Probe	7. p-value(TissueType * Probe Set ID)
1.	20	3958658	NM_004737 // LARGE //	1.03386e-7	7.10948e-5	0	0.516768
2.	8	3424705	NM_032165 // LRRIQ1 //	8.40573e-7	2.40558e-5	2.79634	0.011025
3.	19	3807809	NM_014593 // CXXC1 //	1.2384e-6	0.000223065	0	0.233067
4.	15	3163982	NM_139238 //	1.37665e-6	0.00187248	0	0.474914
5.	8	2908179	NM_001025366 // VEGFA	2.22166e-6	0.0193296	0	0.0242858
6.	13	2406926	NM_000831 // GRIK3 //	2.99205e-6	0.00112263	0	0.940218
7.	8	3727583	NM_002126 // HLF //	3.13025e-6	0.00508871	3.02166	0.338315
8.	9	3082373	NM_003382 // VIPR2 //	4.09045e-6	0.0132391	0	0.553657

Figure 11. 48: Alternative splicing ANOVA result spreadsheet

To visualize gene that most significantly shows alt-splicing, right click on the column header of the p-value of interaction of exon with the factor, choose **Sort Ascending**, and then right click the first row header and choose **Gene View** from the pop-up menu (Figure 11. 51)

Note: the alt-splice result must be a child of the exon expression spreadsheet in order to invoke the gene view.

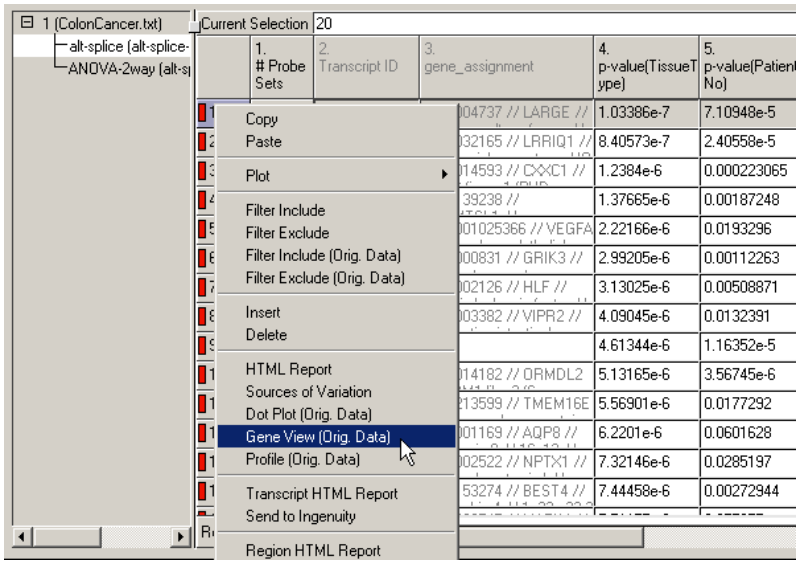


Figure 11. 49: Plot gene view of the gene

There are two types of information showing about the gene in the view. In the bottom part, exons are on the x-axis equally spaced by default, gene expression is on y-axis; each dot represents the average expression of the exon in a subgroup, the error bar represents standard error; all the variant isoforms in the location retrieved from UCSC browser are displayed in the top part of the viewer.

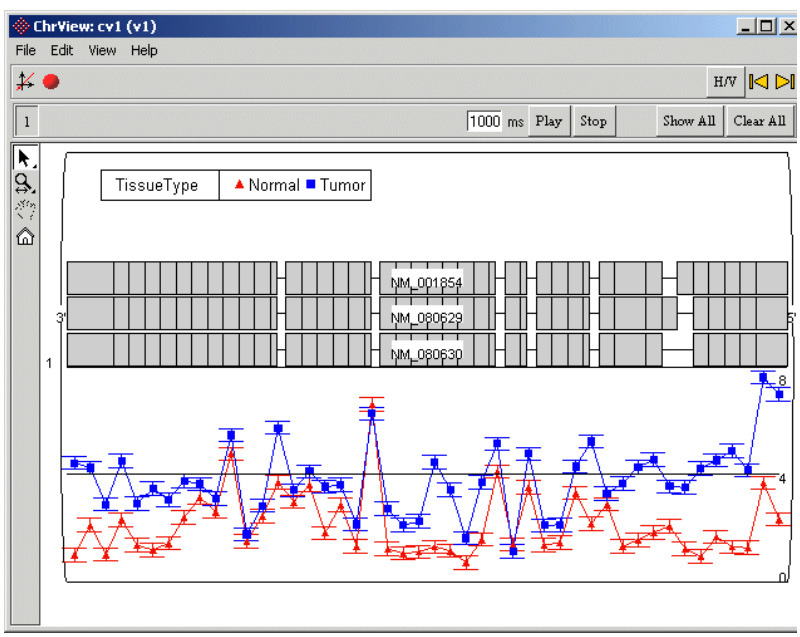


Figure 11. 50: Gene view of exons expression depending on tissue types, there are 46 exons in the gene on the x-axis showing at the bottom part of the graph, the expression is on y-axis. Each red dot represents ls mean expression of an exon in normal tissue; each blue dot represents ls mean expression of an exon in tumor tissue. The top part of the graph showing all the versions of isoform in this region retrieved from UCSC browser

Association Tests

Introduction

This section describes the available tools for conducting SNP association studies. The following topics are discussed:

- $\chi 2$ Association Tests
- Hardy-Weinberg Equilibrium
- Linkage Disequilibrium

Partek provides a number of tests useful in conducting association studies across hundreds of thousands of SNP calls. All SNP columns in a spreadsheet will automatically be processed when invoking these options. If processing all SNPs is not desired, filters can be used to only select the columns of interest.

The menu choices for these operations can be found by selecting **Stat > Association Study**. If this menu is not visible, make sure the target spreadsheet is active and has the associated "genotype" property to notify Partek that the spreadsheet has data appropriate for association studies. Properties can be added to a spreadsheet by selecting **File > Properties...**.

$\chi 2$ Association Tests on SNP Data

Introduction

Partek provides a set of statistic calculations for determining the single SNP association between the all genotype calls and a categorical column.

- Statistics
- Pearson $\chi 2$
- Likelihood ratio
- Models
    - Allele
    - Genotype
    - Recessive/Dominant

Implementation Details

Multiple statistics and models can be chosen for any run of tests. These will be described below.

Statistics

The Pearson χ^2 and likelihood ratio statistics test the null hypothesis of row and column independence. They are computed by the formulas:

- $$CH(x) = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\left(x_{ij} - n_i m_j / N\right)^2}{n_i m_i / N}$$

- $$LI(x) = 2\sum_{i=1}^{r}\sum_{j=1}^{c} x_{ij}\log\frac{x_{ij}}{m_i n_j / N}$$

In the above x_{ij} is the frequency for the cell at the ith row and j^{th} column of the contingency table, m_i and n_j are the marginal row and column totals (respectively), and N is the total frequency for the whole table. Both statistics follow a χ^2 distribution with (r-1) (c-1) degrees of freedom. The number of columns in the contingency table will be determined by the model used for the test. The number of rows will be the number of levels for the column of study.

Models

Partek provides multiple choices for the model used when performing the association test.

The Allele model considers each SNP call to contribute two alleles to the frequency table. For example, a call of AA will contribute two A counts while a call of AB will contribute one A and one B. Samples with no call (NC) for the given allele are excluded from the test at that SNP. The contingency table created in an allele association test has dimension r x 2, where r is the number of unique values for the chosen categorical variable.

The genotype model uses the four genotype calls (AA, AB, BB, NC) in comparison to the chosen categorical. Each sample provides exactly one contributed frequency. It is worth noting that unlike the allele model, the genotype still uses the no call SNPs when generating the frequency table. The dimension of the resulting table is r x 4.

Dominant/recessive models divide the genotype calls into two categories. The first group of genotypes has the “dominant” allele present as one of the alleles. The second group is a homozygous “recessive” call. Samples with no call for a locus are not used for that SNP. Since the assignment of A and B alleles may be arbitrary, the dominant/recessive model analyzes two different tables (one with a “dominant” A, the other a “dominant” B) at each SNP. Samples with no call for a given SNP are not included in the analysis for that location. The contingency tables generated by the dominant/recessive model are r x 2.

Invoking the χ^2 Test Dialog

The *SNP Chi Square Test* dialog is invoked by selecting **Stat > Association Study > Chi Square on SNP**.

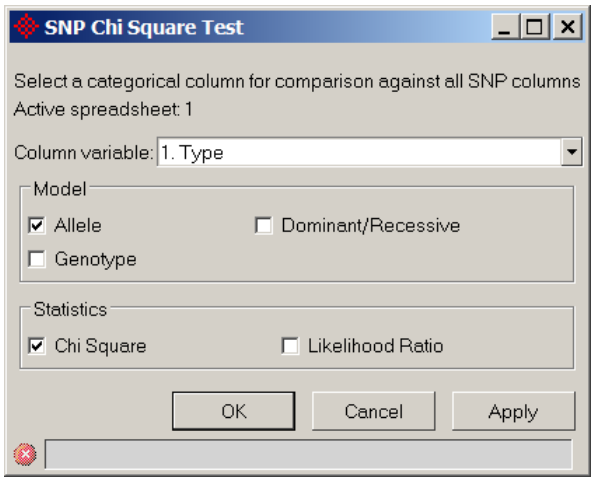


Figure 11. 51 : Configuring the SNP Chi Square Test dialog

Once invoked, you can choose the combination of models and statistics that you would like to compute. The result spreadsheet's column labels will show the statistic and model used in each calculation.

Hardy-Weinberg Equilibrium

Introduction

Hardy-Weinberg Equilibrium is an expectation of genotype frequencies given allele frequencies for a given SNP. Calculating the Hardy-Weinberg Equilibrium (HWE) may be desired for quality control or examining population statistics. Partek provides the following information in the result spreadsheet for each SNP.

- Exact p-value
- Pearson's Chi Square and asymptotic p-value
- Frequency of alleles

Implementation Details

The exact p-value for a given SNP is determined by efficiently enumerating all possible allele combinations for a fixed number of alleles. The probabilities of equally or more extreme genotype frequencies are summed to determine the exact p-value.

The Pearson chi square test calculates the expected genotype frequencies for the allele frequencies. The differences between expected and observed frequencies are then used to calculate chi square.

In all statistical calculations of HWE, Partek ignores No Call (NC) SNPs. The allele frequencies are also provided in the result spreadsheet. These may be useful in examining the minor allele frequency for a SNP.

Invoking the HWE Test

The Hardy-Weinberg Equilibrium can be found on the main menu by selecting **Stat > Association Study > Hardy-Weinberg Equilibrium** as seen in Figure 11. 52.

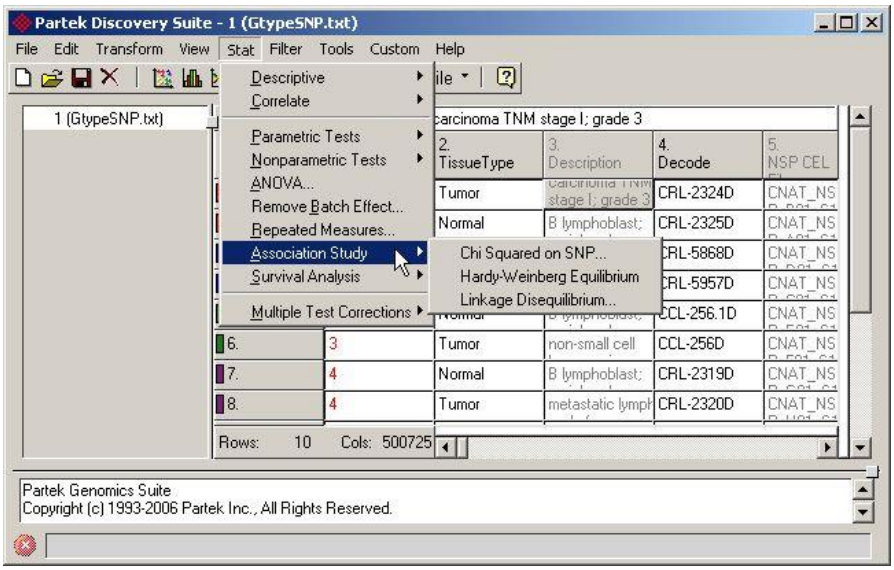


Figure 11. 52: Selecting Association Study > Hardy-Weinberg Equilibrium

For the *Association Stud*y menu to appear under the *Stat* menu, the currently active spreadsheet should have the "genotype" property to notify Partek that the spreadsheet contains data appropriate for genotype analysis.

Linkage Disequilibrium

Linkage Disequilibrium (LD) may be of interest in determining which alleles may be transferred together in a population. Partek provides a windowed approach to determining the LD around a given SNP to efficiently calculate LD across the whole genome. Filters can be used to exclude SNPs that are not desired (when only concerned with LD on one chromosome, for example).

Partek will generate one of three statistics when calculating LD.

- D
- D'
- r^2

Implementation Details

The windowed approach used by Partek allows you to specify the number of SNPs upstream and downstream that will have LD statistics generated for any loci. The window size does not relate to genomic distance, only the closest measured SNPs.

Partek also assumes the spreadsheet to be operated has columns in the correct genomic order.

Linkage Disequilibrium provides information regarding the strength of the relationship between two loci for the data's haplotypes. Partek uses an EM algorithm to infer haplotypes for samples with heterozygous calls at both loci. Samples with no call (NC) at either locus are not considered in LD calculations for that pair of SNPs.

The three statistics of LD calculated by Partek are calculated as:

- $D = Observed(A1, A2) - P(A1)P(A2)N$
- $D' = \frac{D}{D_{max}}$
- $r^2 = \frac{D^2}{P(A1)P(B1)P(A2)P(B2)}$

In the above formulations, N is the total number of alleles, and Ai and Bi are allele calls A or B at locus i, respectively.

Invoking the Linkage Disequilibrium Dialog

The LD dialog can be found from the main menu by selecting **Stat > Association Tests > Linkage Disequilibrium**.

Figure 11. 53: Configuring the Linkage Disequilibrium dialog

From the dialog seen in Figure 11. 53, you can choose an integer value for the number of surrounding SNPs to consider for LD for each SNP and the desired LD statistic.

LD Plot

Partek also provides visualization of the LD statistics once they are generated. For this visualization to represent the data properly, the LD result spreadsheet should not be sorted, but rather retain the genomic ordering from the parent spreadsheet.

Once a LD result spreadsheet has been created, you can invoke an LD plot by right clicking on a row in the LD result spreadsheet and choosing **LD Plot** (Figure 11. 54).

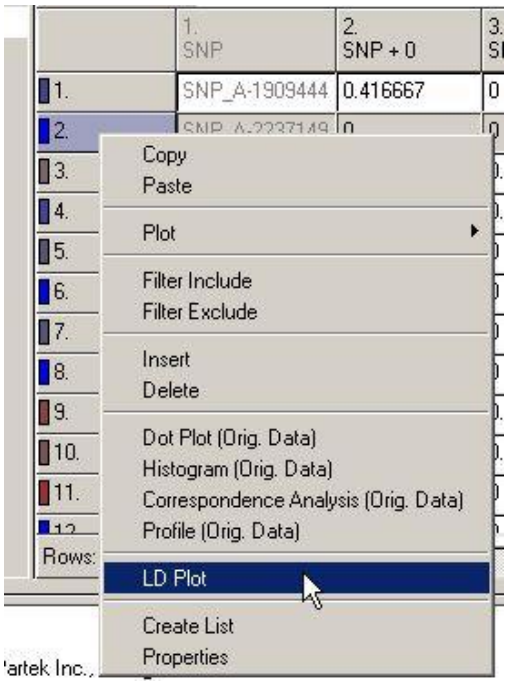


Figure 11. 54: Invoking the LD plot from the result spreadsheet

An intensity plot similar to Figure 11. 55 will be created. The intensity plot will be centered on the SNP on which the plot was invoked. In addition to the visualization, a spreadsheet containing the data displayed in the plot is created for viewing in tabular format. You can also view these values directly from the intensity plot by hovering the mouse over the SNP pair of interest.

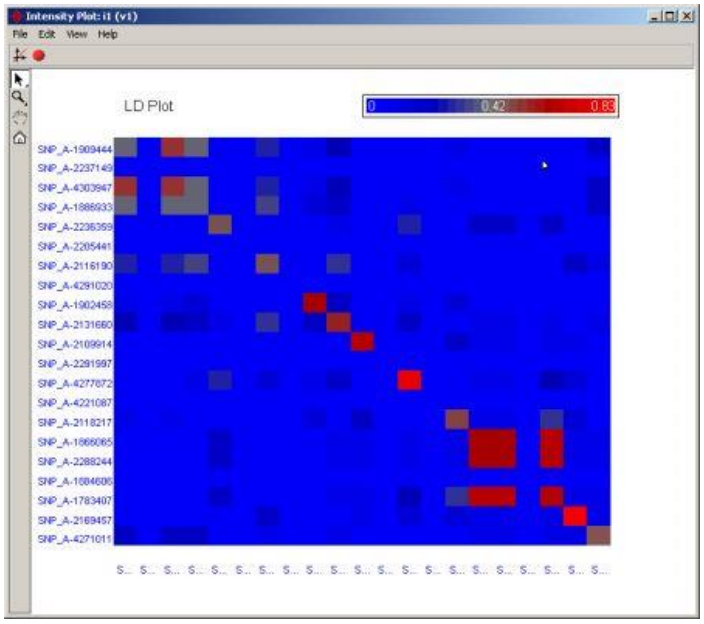


Figure 11. 55: Viewing the Linkage Disequilibrium plot

Logistic Regression Analysis

Introduction

Linear regression is often used to analyze the relationship between a numerical response variable with one or multiple explanatory variables (numerical or categorical). However, when the response variable is binary, logistic regression is often used to model the relationship between the response variable and the explanatory variables (Agresti, 2002).

Implementation Details

Logistic Regression:

Let P be the probability of an event and m be the number of the events and n be the total number of trails. The odds of the event is defined as

$$Odds = \frac{P}{1-P} = \frac{m}{n-m}$$

Logit is the natural log of the odds, that is,

$$\log(odds) = \log it(P) = \ln(\frac{P}{1-P})$$

In logistic regression, logit(P) is as dependent variable and X is as explanatory variable. The linear relationship is described as

$$\log it(P) = a + bX$$

Maximum likelihood estimation (MLE):

The coefficients of logistic regression model a and b are estimated by maximum likelihood estimation (MLE). This is different from linear regression estimation which uses ordinary least squares (OLS) estimation. OLS is to minimize the sum of squared distances of the data points to the regression line. MLE is to maximize the Log Likelihood function (LL), which reflects how likely it is (the odds) that the observed values of the response variable may be predicted from the observed values of the explanatory variables. For the jth observation, let $\hat{p}_j$ be the estimated probability of the observed response, the log likelihood function is as

$$-2LogL = -2\sum_j w_j f_j \{r_j \log(\hat{p}_j) + (n_j - r_j)\log(1-\hat{p}_j)\}$$

Where w_j and f_j are the weight and frequency values of the jth observation, r_j is the number of events, n_j is the number of trials. Newton-Raphson Algorithm is used to get MLE. Let $\beta' = (\beta_0, \beta_1, ..., \beta_k)$, the gradient vector and the Hessian matrix are

$$g_\beta = \sum_j w_j f_j \frac{\partial l_j}{\partial \beta}$$

$$H_\beta = \sum_j - w_j f_j \frac{\partial^2 l_j}{\partial \beta^2}$$

The maximum likelihood estimate $\hat{\beta}$ of β is obtained iteratively by the following function until LL converge.

$$\beta_{m+1} = \beta_m + H_{\beta_m}^{-1} g_{\beta_m}$$

The Logistic Regression Dialog

To open the *Logistic Regression* dialog, select **Stat > Logistic Regression**, The *Logistic Regression* dialog (Figure 11. 60) is used to specify the explanatory variable, weight variable and encode method. Figure 11. 60 is configured to predict the event using all the response variables – one response variable per test at a time.

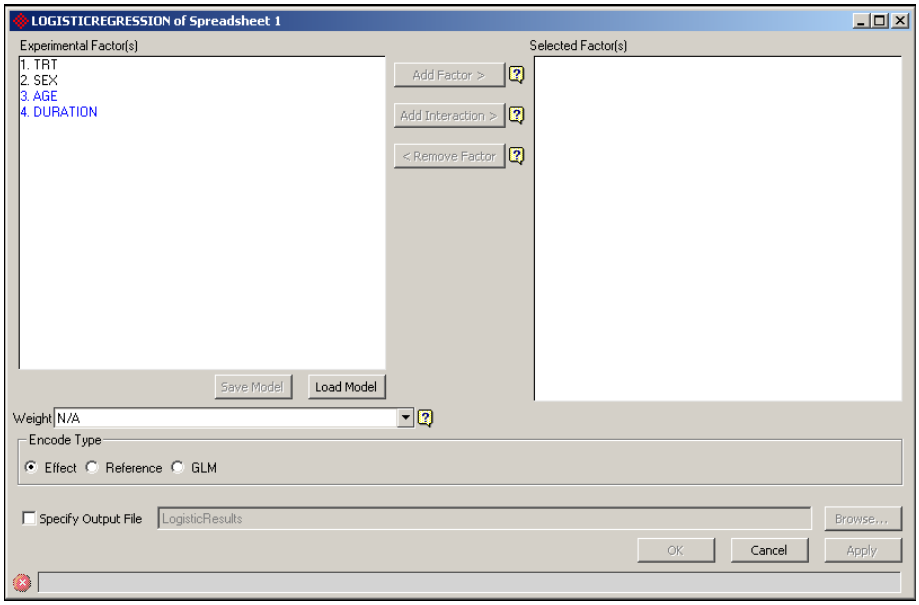


Figure 11. 56: Configuring the Logistic Regression model

Selecting the Weight Variable and Encode Type

The *Weight Variable* must be selected from the drop down list (Figure 11. 61), which contains integer response variables in the spreadsheet. This variable measures the number of the trails. If the weight variable is N/A, the weight will set as 1. There are three encode type which are Effect, Reference and GLM. Effect encode will allow an explanatory variable contribute to levels-l columns in design matrix. This explanatory variable is a categorical variable and levels are the levels of this categorical variable. The *ith* level will set the *ith* diagonal element as 1 and the last level will set the last row as -1. Reference encode is similar as effect encode expect that the last level will set the last row as 0. GLM encode will allow an explanatory variable contribute to levels columns in a design matrix. The *ith* level will set the *ith* diagonal element as 1.

Settings for the Logistic Regression Model

Highlight the factor(s) and click Add Factor > button to add the selected factor(s) into the model.

Select at least two factors on the *Experimental Factor(s)* panel, and click Add Interaction > button to add the interaction of the selected *Factor(s)*. To remove an interaction, select it on the *Factor(s)* panel and click the < *Remove Factor(s)* button. Click Apply or OK button, an analysis is performed on all response variables, the results will be stored in a child spreadsheet (Figure 11. 64). Each row of the result spreadsheet corresponds to one of the numeric columns in the parent spreadsheet.

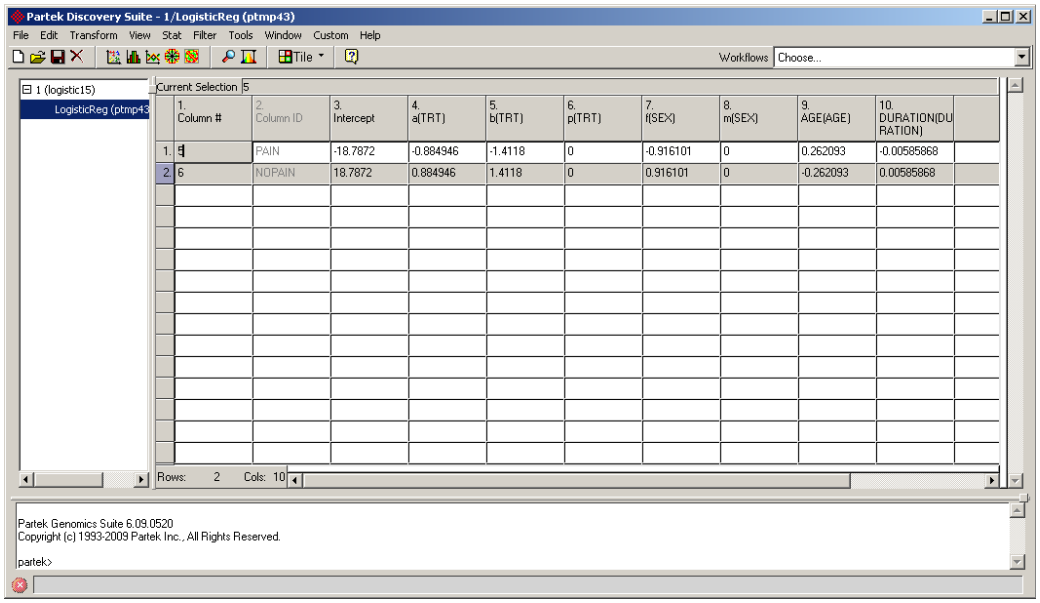


Figure 11. 57: Result of the Logistic Regression on all response variables

In the result spreadsheet, the test result of each response variable is in a row and is summarized by the column number and name of the variable. It is followed by summary statistics, such as coefficient estimate.

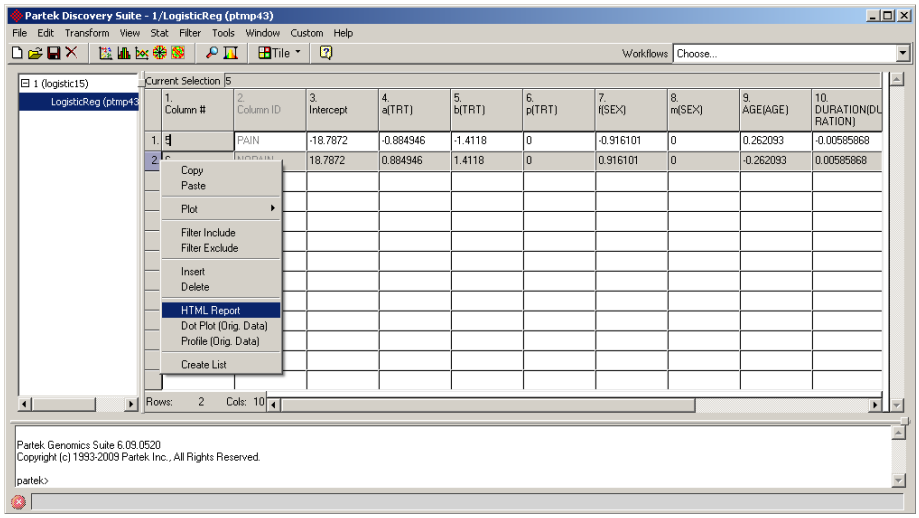


Figure 11. 58: Selecting an HTML Report from the Logistic Regression results spreadsheet

Detailed reports about individual response variables can be viewed by right-clicking on the row label corresponding to the response variable and select **HTML Report** option in the pop-up menu (Figure 11. 65). Figure 11.66 is an example of HTML result report.

Logistic Regression of "logistic15" on PAIN

Table of Contents

Parameter Estimates go to top

Name	Estimate	DF	StdErr	Chisquare	p value
Intercept	-19.223610	1	7.131533	7.266145	0.007027
a(TRT)	-0.848264	1	0.550157	2.377321	0.123109
b(TRT)	-1.494909	1	0.662241	5.095615	0.023986
f(SEX)	-0.917302	1	0.398058	5.310452	0.021198
AGE(AGE)	0.268753	1	0.099645	7.274431	0.006994
DURATION(DURATION)	-0.005230	1	0.033296	0.024676	0.875177
a * f(TRT * SEX)	0.201040	1	0.556791	0.130371	0.718048
b * f(TRT * SEX)	-0.048729	1	0.556328	0.007672	0.930202

Effect Information go to top

Name	DF	Chisquare	p value
TRT	2	11.988605	0.002493
SEX	1	5.310452	0.021198
AGE	1	7.274431	0.006994
DURATION	1	0.024676	0.875177
TRT * SEX	2	0.141175	0.931846

* p-value is less than 0.05

Figure 11. 59: Viewing the example report for a single Logistic Regression test

Survival Analysis

Introduction

Survival Analysis was partly developed in the medical and biological sciences. The most important feature of survival data is the presence of "censored" data. For example, in the medical research we may study the survival of patients from after treatment for a disease, including death rates (time to death), etc. In each case, by the end of the study period, while we may know the "survival time" for some patients, some will still be alive, and others will have dropped out during the study period; thus, those patients represent "censored" observations.

The information from censored data is valuable because while it does not fully measure survival time, it does measure at least a minimum length of survival prior to the time the study ends or the subject drops out of the study (this is a special type of censored data referred to as "right-censored"). Special tests are developed to correctly use the censored observations together with the uncensored observations. Kaplan-Meier Test and Cox Regression analyze right-censored data due either to withdrawal of subjects or termination of the experiment. Kaplan-Meier Test gives the estimation of survival function, rank test, Log-Rank, Wilcoxon, and univariate Chi-Square test. Cox Regression provides the coefficient estimates of the Cox proportional hazards model and model fit statistics.

Implementation Details

Survival Function:

Let $t_1, t_2, \ldots, t_n$ be the exact survival times of the n individuals under study. We first re-label the n survival times in order of increasing magnitude such that $t_{(1)} \leq t_{(2)} \leq t_{(3)} \ldots \leq t_{(n)}$. Then survival function at t

$$S(t) = \prod_{t_{(r)} \leq t} \frac{n-r}{n-r+1}$$

where $t_{(r)}$ is uncensored (Kaplan, E. L., and Meier, P. 1958).

Log-Rank Test and Wilcoxon Test:

These statistics are used to test homogeneity of survival functions from strata. Strata are variables to classify the samples into different groups.

Let vector $v=(v_1,v_2, \ldots, v_c)'$ with

$$v_j = \sum_{i=1}^{k} w_i (d_{ij} - n_{ij} d_i / n_i),$$

where c is the number of strata. The estimated covariance matrix, $V=(V_{jl})$, is given by

$$V_{jl} = \sum_{i=1}^{k} w_i^2 (n_i n_{il} \delta_{jl} - n_{ij} n_{il}) d_i s_i / (n_i^2 (n_i - 1)) ,$$

where i labels the distinct event times, δ_{jl} is 1 if j=l and 0 otherwise, n_{ij} is the size of the risk set in the jth stratum at the ith event time, d_{ij} is the number of events in the jth stratum at the ith time, $n_i = \sum_{j=1}^{c} n_{ij}, d_i = \sum_{j=1}^{c} d_{ij}$, $s_i = n_i - d_{i.}$ The term w_i is 1 for the log rank test and n_i for the Wilcoxon test (Peto, R., and Peto, J. 1972).

Univariate Test of Covariates:

The index a labels all observations, a=1,2,…,n, and the indices i, j will mark the observations that correspond to events, i, j=1,2 …, k. The ordered event times are denoted as $t_{(i)}$, the corresponding vectors of covariates are denoted $z_{(i)}$, and the ordered times, both censored and event times, are denoted t_a. The rank test statistics have the form

$$v = \sum_{a=1}^{n} c_{a,\delta_a} z_a$$

where n is the total number of observations, c_{a,δ_a} are rank scores, δ_a is 1 if the observation is an event and 0 if the observation is censored, and z_a is the vector of the test variable.

$$c_{a,\delta_a} = \sum_{(j:t_{(j)} \le t_a} (1/n_j) - \delta_a$$

where n_j is the number at risk just prior to $t_{(j)}$. The estimated covariance matrix is given by

$$V = \sum_{i=1}^{k} (\sum_{(a:t_a \ge t_{(i)}} (z_a - \overline{z_i})'(z_a - \overline{z_i})) / n_i ,$$

where $\overline{Z_i} = \sum_{(a:t_a \ge t_{(i)})} z_a / n_i$.

The univariate tests for each covariate are formed from each component of v and the corresponding diagonal elements of V as v_i^2 / V_{ii} .

Cox Regression:

Cox regression (also called Cox proportional-hazards regression) allows analyzing the effect of several risk factors on survival. The probability of the event is called the “hazard”. The hazard is modeled as

$$H(t) = H_0(t) * \exp(b_1 X_1 + b_2 X_2 + b_3 X_3 + ... + b_k X_k) ,$$

where $X_1 \ldots X_k$ are the set of predictor variables and $H_0(t)$ is the baseline hazard at time t when all predictor variables are zero. Dividing both sides of the above equation by $H_0(t)$ and taking logarithms, we obtain:

$$\ln\left(\frac{H(t)}{H_0(t)}\right) = b_1 X_1 + b_2 X_2 + b_3 X_3 + \ldots + b_k X_k ,$$

$H(t)/H_0(t)$ is the "hazard ratio". To estimate the coefficients $b_1, \ldots, b_k$, Cox (1972) proposed a partial likelihood function based on the conditional probability of failure, assuming that there are no tied values among the survival times. Later Cox's partial likelihood function was modified to handle ties (Efron, 1977). The maximum partial likelihood estimator $\hat{b}$ of b can be obtained by solving the following simultaneous equations:

$$\frac{\partial(l(b))}{\partial b} = 0 \quad \text{and} \quad \hat{Cov}(\hat{b}) = \left[-\frac{\partial^2 l(\hat{b})}{\partial b \partial b'} \right]^{-1} .$$

The Cox Regression Dialog

To open the *Cox Regression* dialog, select **Stat > Survival Analysis > Cox Regression…** The *Cox Regression* dialog (Figure 11. 60) is used to specify the time variable, event variable, predictor, and strata. Figure 11. 60 is configured to predict the event happens using all the response variables – one response variable per test at a time.

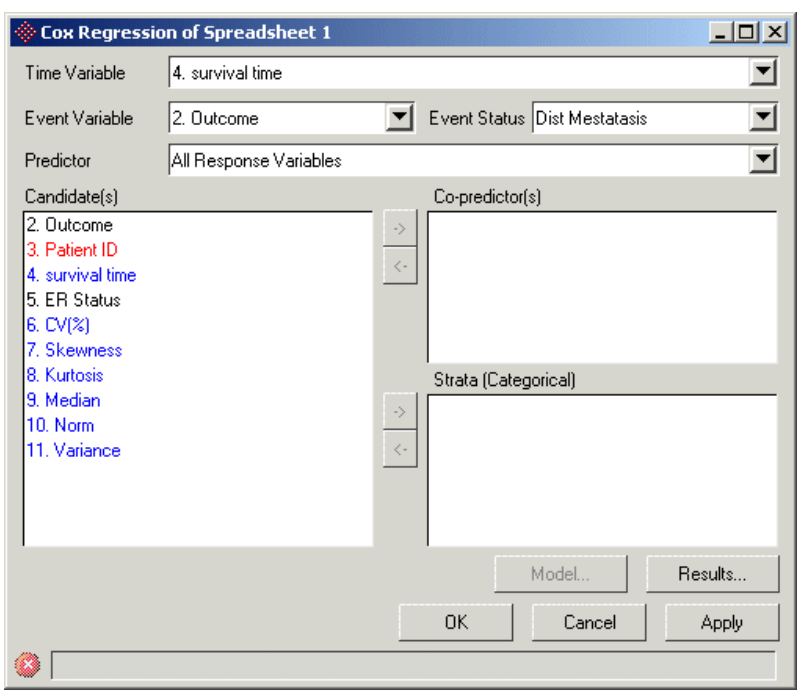


Figure 11. 60: Configuring the Cox Regression model

Selecting the Time Variable

The *Time Variable* must be selected from the drop down list (Figure 11. 61), which contains numeric factor variables in the spreadsheet. This variable measures the time when the event occurs.

Figure 11. 61: Selecting the Time variable

Selecting the Event Variable

The *Event Variable* tells if the time in the *Time Variable* is event time or censor time. It can be selected from the drop down list (Figure 11. 62), which contains categorical variables that have only two subgroups in the spreadsheet, for example metastasis vs. non-metastasis, or dead vs. alive. After the *Event Variable* is specified, the corresponding subgroups of the variables will be listed in the *Event Status* drop down. The selected subgroup name is the event, for example metastasis or dead, and the other subgroup name is censor, e.g. non-metastasis or alive. Choose **NA** if there is no censor.

Figure 11. 62: Selecting the Event Variable

Selecting the Predictor Variable

By default, if there is at least one numeric response variable in the spreadsheet, *All Response Variable(s)* will be shown to test all of the variables, one test per predictor (Figure 11. 63). To choose a specific response as a predictor, select the variable name from the drop-down list. The response variable here is as an input or predictor variable in a Cox Regression model.

Figure 11. 63: Selecting the Predictor variable

When an analysis is performed on all response variables, the results of the all the tests will be stored in a child spreadsheet (Figure 11. 64). Each row of the result spreadsheet corresponds to one of the numeric columns in the parent spreadsheet..

1 (survivalanalysis)
Cox (37)

Current Selection 18

	1. Column #	2. Column ID	3. p-value(gene)
1.	18	1294_at	0.0104908
2.	20	1320_at	0.1082
3.	14	1053_at	0.121463
4.	16	121_at	0.499367
5.	17	1255_g_at	0.552331
6.	19	1316_at	0.56293
7.	13	1007_s_at	0.637993
8.	15	117_at	0.771791

Figure 11. 64: Result of the Cox Regression on all response variables

In the result spreadsheet, each response variable is tested in a row and is summarized by the column number and name of the variable. It is followed by summary statistics, such as p-values for each predictor. Detailed reports about individual response variables can be viewed by right-clicking on the row label corresponding to the response variable and select **HTML Report** option in the pop-up menu (Figure 11. 65).

1 (survivalanalysis)
Cox (37)

Current Selection 18

	1. Column #	2. Column ID	3. p-value(gene)
1.	18	1294_at	0.0104908
2.			0.1082
3.			0.121463
4.			0.499367
5.		_at	0.552331
6.			0.56293
7.		_at	0.637993
8.			0.771791

Copy
Paste
Plot
Filter Include
Filter Exclude
Insert
Delete
HTML Report
Dot Plot (Orig. Data)
Histogram (Orig. Data)
Profile (Orig. Data)

Figure 11. 65: Selecting an HTML Report from the Cox Regression results spreadsheet

If one response variable is selected as a predictor from the drop down list, only one Cox Regression test is performed, the result will be displayed in a HTML report (Figure 11. 66).

Cox Regression Result

Model Information go to top

Test	Chi Square	DF	p-value
Likelihood Ratio	0.221906	1	0.637591
Wald	0.221377	1	0.637993
Score	0.221269	1	0.638075

Coefficient Information go to top

Name	DF	Estimate	Std Error	W (Wald Chi Square)	p-value (W)	Hazard Ratio
1007_s_at	1	0.0989362	0.210276	0.221377	0.637993	1.103996

Model Fit Statistics go to top

	Without Predictor(s)	With Predictor(s)
-2logL	1152.682179	1152.460273
AIC	1152.682179	1154.460273
SBC	1152.682179	1157.123712

Figure 11. 66: Viewing the example report for a single Cox Regression test

To specify a multivariate Cox Regression model, the covariates need to be specified as co-predictors.

Selecting a Co-predictor Variable

The co-predictor can be any number of categorical variables and/or numeric factor variables, e.g. tumor size, age etc. It is as an input variable in a Cox Regression model. Co-predictors must be selected from the *Candidate(s)* list (Figure 11. 67). When an item in the *Candidate(s)* list box is selected, the -> button will be enabled. Click on that button to move the selected item to the *Co-Predictor(s)* list box. To remove a co-predictor, select it in the *Co-Predictor's* list box and click on the <- button to move it back to the *Candidate(s)* list box.

If **NA** is selected from the *Predictor* drop down list, there must be at least one co-predictor specified. The total predictors of one model include one *Predictor* and one or more *Co-Predictor(s)*.

When there are more than two predictors (including co-predictors), the interactions among the predictors can be added to the model.

Settings for the Cox Regression Model

Click on the **Model** button to invoke the *Model Configure* dialog (Figure 11. 67).

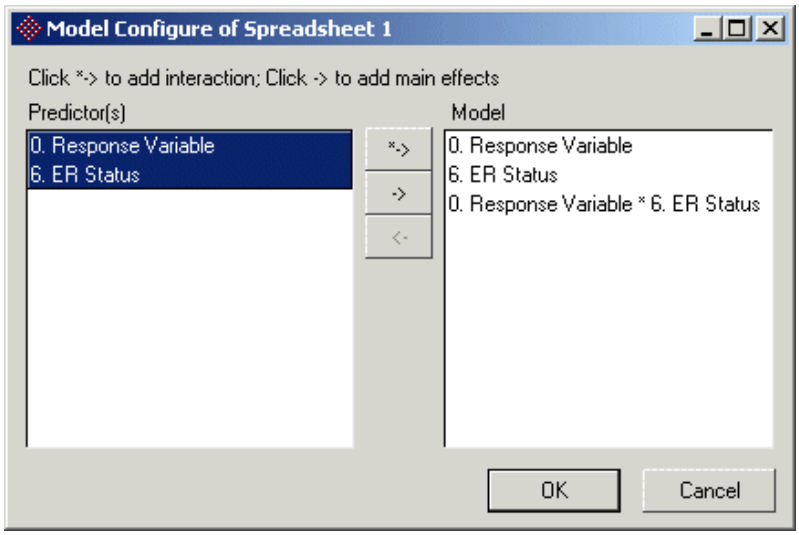


Figure 11. 67: Adding interaction of each response variable with ER Status in each Cox Regression model, ER Status was specified as a Co-Predictor

All the predictors (including co-predictors) specified in the *Cox Regression* dialog are listed in the *Predictor(s)* panel. If *All Response Variable(s)* was selected in the *Predictor(s)* drop down list, "**0. Response Variable**" appears on the top of the list representing each response numeric variable that will be computed in each test (Figure 11. 67); if *NA* was selected in the *Predictor(s)* drop down list, the model will only include variables in the *Co-Predictor* list; if a specific response numeric variable was selected in the *Predictor(s)* drop down list, that variable will appear on the top of the *Predictor(s)* panel list.

Select at least two predictors on the *Predictor(s)* panel, and click the *-> button to add the interaction of the selected predictors. T remove an interaction, select it on the *Model* panel and click the <- button.

Selecting the Stratified Variables

The stratified analysis is to test if the regression models are identical for different group, e.g. ER status (ER+, ER-). The strata can be any number of categorical variables; they must be selected from the *Candidate(s)* list (Figure 11. 68). When a categorical variable in the *Candidate(s)* list box is selected, the -> button next to the *Strata (Categorical)* list box will be enabled. Click on that button to move the selected item to the *Strata (Categorical)* list box. To remove a stratifying variable, select it in the *Strata (Categorical)* list box, and click the <- button.

Configuring the Result Spreadsheet for Multiple Tests

By default, the result spreadsheet of tests on *All Response Variables* only displays the p-values of each predictor in the model. To display more statistical results for each test, click the **Result** button. Select the corresponding checkbox to display the values in the result spreadsheet.

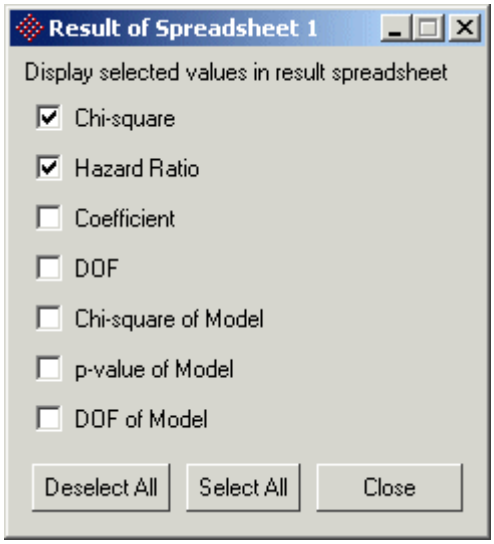


Figure 11. 68: Configuring the statistics results displayed on the results spreadsheet

The Kaplan-Meier Dialog

To open the *Kaplan-Meier* dialog, select **Stat > Survival Analysis > Kaplan-Meier…** The *Kaplan-Meier* dialog (Figure 11. 69) is used to specify the time variable, event variable, Test Variables, and strata. Figure 11. 69 is configured. The test variable has to be a numerical variable to test the association of survival time with covariates. The use of the strata analysis here is to test equality of survival curves across strata.

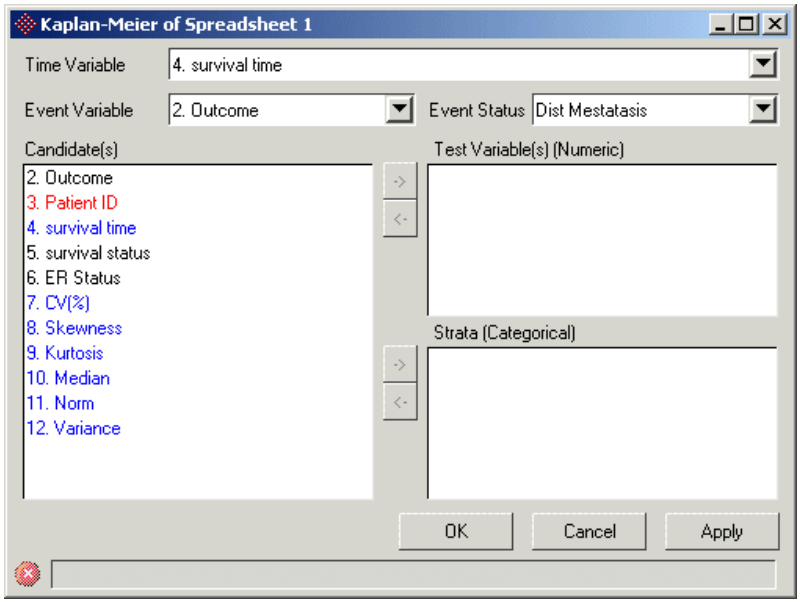


Figure 11. 69: Configuring the Kaplan-Meier dialog

Running the Computations

OK will perform the configured computation and dismiss the dialog.

Apply will perform the configured computation, but the dialog will remain to perform another computation.

Cancel will close the dialog without doing any computations.

Non-parametric Tests

Mann-Whitney Test

Introduction

The Mann-Whitney test is a nonparametric test used to compare two groups. Unlike parametric tests, it makes no assumptions about the distribution of the data, e.g. normality or homogeneity of variance. It is the nonparametric alternative to a two sample *t*-test and is useful when the assumptions of normality or equality of variance are not met. However, if the assumptions of normality and homogeneity of variance are valid, this test is less powerful than the parametric *t*-test. The Mann-Whitney test uses the ranks of the data (including tied rank values when appropriate) rather than the original values to compute the 'U' statistic, which is used to calculate the probability that neither group is "stochastically larger" than the other. This test can be loosely viewed as a test for a difference in medians.

Implementation Details

The implementation of the Mann-Whitney test is only valid when the number of samples in each group is greater than or equal to five. Partek ranks all the values from two groups. If two values are the same, both of them get the same rank, which is the average of the two ranks. The smallest number of values is ranked as 1 and the largest number is N. Rank Sum is the sum of the ranks. Mean Rank is the mean of the ranks. Median is the middle value of the ranks. The statistics to test two groups are different include U Statistics, σ and σ_{adj} (adjusted for ties), z and z_{adj}.

They are given as the following:

$$U = n_1 n_2 + n_1(n_1+1)/2 - RankSum(1),$$

$$\sigma = \sqrt{n_1 n_2 (n+1)/12},$$

$$\sigma_{adj} = \frac{n_1 n_2}{12}\left(n+1-\frac{\sum_{i=1}^{g} t_i^3 - t_i}{n(n-1)}\right),$$

$$z = U - \frac{n_1 n_2}{\sigma},$$

$$z_{adj} = U - \frac{n_1 n_2}{\sigma_{adj}},$$

Where $n = n_1 + n_2$

g = the number of groups of ties

t_i = the number of tied ranks in group i.

The normal approximation of p-value and p-value (corrected for ties) are z test probabilities.

Configuring the Mann-Whitney Dialog

To open the *Mann-Whitney* dialog, select **Stat > Nonparametric Tests > Mann-Whitney…**. Figure 1 shows the main dialog for the Mann-Whitney test. This dialog is used to specify the grouping variable (factor), the response variable(s) to be tested, and any multiple test corrections. Figure 11. 70 is configured to test for a difference between ALL and AML on all the numeric variables.

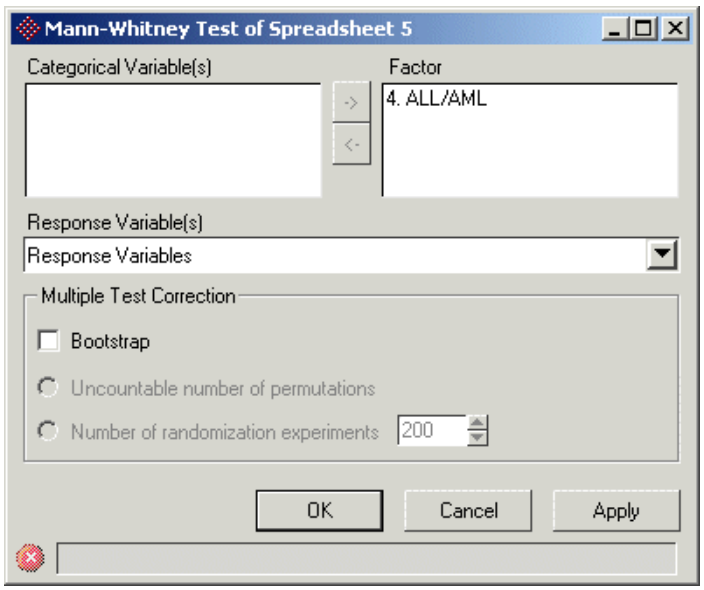


Figure 11. 70: Configuring the Mann-Whitney test dialog for multiple tests

Selecting Grouping Variables

The *Grouping Variable* (or factor) must be selected from the *Categorical Variable(s)* list, which contains variables that have only two categories (levels) in the spreadsheet. There can be only one grouping variable in Mann-Whitney computations. When an item in the *Categorical Variable(s)* list box is selected, the -> button next to *Grouping Variable* list box will be enabled; click on it to move the selected item to the *Grouping Variable* list box. To remove a factor, select it in the *Grouping Variable* list box and the <- button next to it will be enabled; click on the <- button and the item selected in the *Grouping Variable* list box will be moved back to the *Categorical Variable(s)* list box.

Selecting Response Variables

By default, if there is more than one numeric variable in the spreadsheet, *Response Variables* will be shown as the *Response Variable(s)* to test all of them at one time. To choose a specific response variable to test, select the variable name from the

drop-down list; however, if there is only one numeric variable in the spreadsheet, the variable name will be selected as the *Response Variable* by default.

When an analysis is performed on all numerical variables, the results will be summarized in a new spreadsheet that is a child of the original. In the results spreadsheet, each variable tested in a row is summarized by the column number and name of the variable, and followed by summary statistics including the p-values, means, and standard deviations for each factor. The rows are automatically sorted by the first column of p-values. To sort by a different p-value, right click on the column heading and select **Sort Ascending** in the pop-up menu. Detailed reports about individual test variables can be viewed by right-clicking on the row label corresponding to that variable and selecting the **HTML Report** option on the pop-up menu (Figure 11. 71).

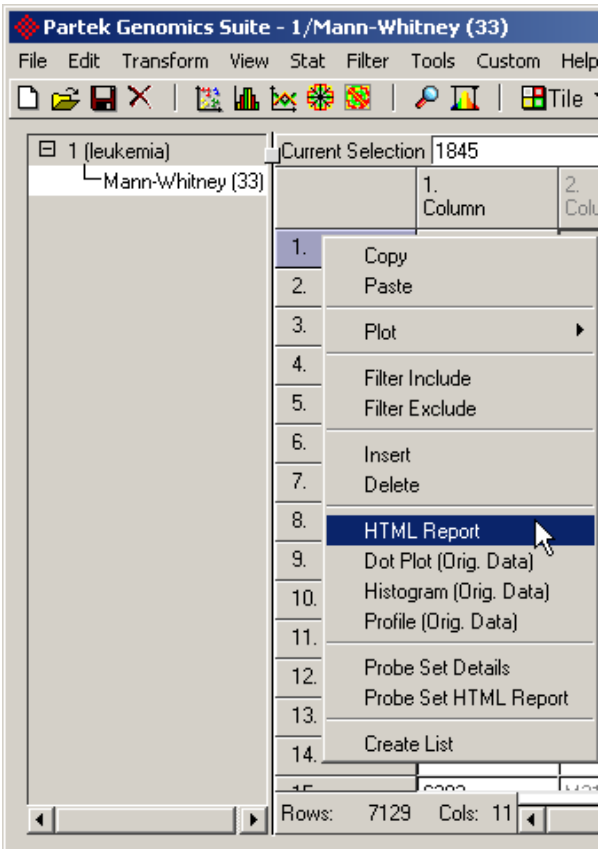


Figure 11. 71: Selecting an HTML report based on the result spreadsheet of the multiple tests with Mann-Whitney

To test a single response variable, select the variable from the drop-down list, the result will be displayed in a HTML report (Figure 11. 72).

Mann-Whitney report of variable M23197_at

Grouping Variable: ALL/AML,
Response Variable: M23197_at

Factor Level Information

Factor	Levels	Level Values
ALL/AML	2	ALL, AML

Descriptive

ALL/AML	N	Median	Mean Rank	Rank Sum	U Statistic
ALL	47	154	24.2766	1141	1162
AML	25	769	59.48	1487	13

Mann-Whitney

	U Statistic	p-value(chisqr approximation)
Uncorrected	1162	2.07921e-011
Corrected for Ties	-6.78926	2.07566e-011

Multiple Test correction

Original p-value	Bonferroni
2.07566e-011	1.47974e-007

Figure 11. 72: Viewing the HTML report of the Mann-Whitney, single test

Running the Computations

OK will perform the configured Mann-Whitney computation and dismiss the dialog.

Apply will perform the configured Mann-Whitney computation, but the Mann-Whitney dialog will remain to perform another computation.

Cancel will close the dialog without doing any computation.

Kruskal-Wallis Test

Introduction

The Kruskal-Wallis test is a nonparametric test used to compare two or more groups of sampled data. Unlike the parametric test, it makes no assumptions about the distribution of the data, e.g. normality or homogeneity of variance. It is the nonparametric alternative to the one-way ANOVA, and is useful when the assumptions of normality or equality of variance are not met. However, if the assumptions of normality and homogeneity of variance are valid, this test is less powerful than the parametric ANOVA. The Kruskal-Wallis test uses the ranks of the data rather than the original values to calculate the ‘H’ statistic. This test can be loosely viewed as a test for a difference in medians.

Implementation Details

The implementation of the Kruskal-Wallis statistic is only recommended when the number of samples in each group is greater than or equal to five. For samples of this size, the sampling distribution of the H statistic can be approximated by the chi-square distribution for the degrees of freedom being equal to the number of groups minus one (dof = k -1).

Configuring the Kruskal-Wallis Dialog

To open the *Kruskal-Wallis* dialog, select **Stat > Nonparametric Tests > Kruskal-Wallis…**. The *Kruskal-Wallis* dialog (Figure 11. 73) is used to specify the grouping variable (factor), the response variable(s) to be tested, and the multiple test corrections. Figure 11. 73 is configured to test for a difference between ALL and AML on all the numeric variables.

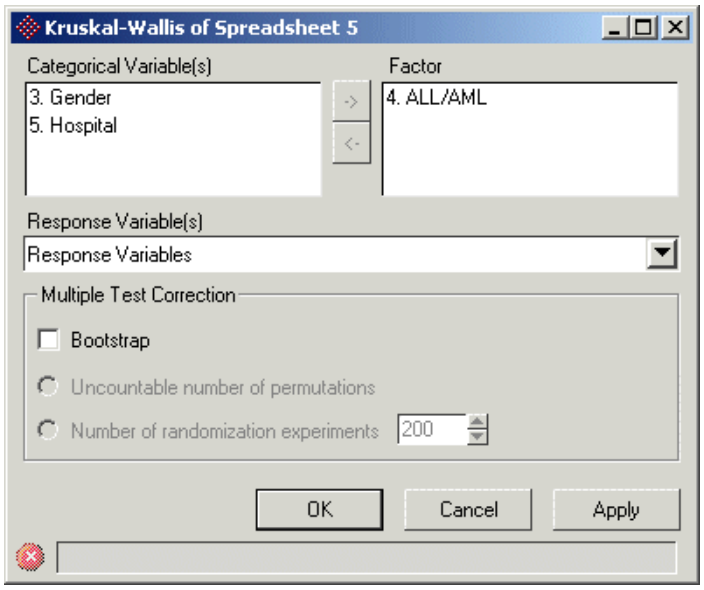


Figure 11. 73: Configuring the Kruskal-Wallis dialog for multiple tests

Selecting Grouping Variables

The *Grouping Variable* (or factor) must be selected from the *Categorical Variable(s)* list, which contains all categorical variables in the spreadsheet. There can be only one grouping variable in Kruskal-Wallis computations. When an item in the *Categorical Variable(s)* list box is selected, the -> button next to the *Grouping Variable* list box will be enabled, click on it to move the selected item to the *Grouping Variable* list box. To remove a factor, select it in the *Grouping Variable* list box and the <- button next to it will be enabled; click on the <- button and the item selected in the *Grouping Variable* list box will be moved back to the *Categorical Variable(s)* list box.

Selecting Response Variables

By default, if there is more than one numeric variable in the spreadsheet, *All* will be shown as the *Response Variable(s)* to test all of them at one time. To choose a specific response variable to test, select the variable name from the drop-down list; however, if there is only one numeric variable in the spreadsheet, the variable name will be selected as the *Response Variable* by default.

When an analysis is performed on all numerical variables, the results will be summarized in a new spreadsheet that is a child of the original. In the results spreadsheet, each variable tested in a row is summarized by the column number and name of the variable, and followed by summary statistics including the p-values, means, and standard deviations. The rows are automatically sorted by the first column of p-values; however, to sort by a different p-value, right click on the column heading and select **Sort Ascending** in the pop-up menu. Detailed reports about individual test variables can be viewed by right-clicking on the row label corresponding to the variable and selecting the **HTML Report** option on the pop-up menu (Figure 11. 74).

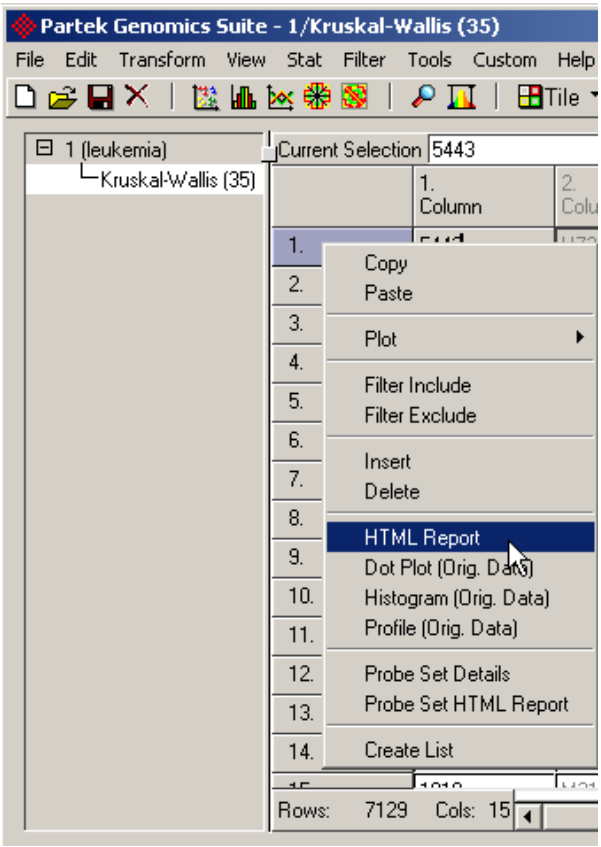


Figure 11. 74: Selecting an HTML report based on the result spreadsheet of the multiple tests with Kruskal-Wallis

To test a single response variable, select the variable from the drop-down list, the result will be displayed in a HTML report (Figure 11. 75).

Kruskal-Wallis Test of "leukemia" on AFFX-BioB-5_at

Grouping Variable: ALL/AML,
Response Variable: AFFX-BioB-5_at

Factor Level Information

Factor	Levels	Level Values
ALL/AML	2	ALL, AML

Descriptive

ALL/AML	N	Mean Rank	Rank Sum
ALL	47	34.9574	1643
AML	25	39.4	985

Kruskal-Wallis

	H Statistic	p-value(chisqr approximation)
Uncorrected	0.735354	0.391153
Corrected for Ties	0.735425	0.39113

Figure 11. 75: Viewing the HTML report of the Kruskal-Wallis, single test

Running the Computations

OK will perform the configured Kruskal-Wallis computation and dismiss the dialog.

Apply will perform the configured Kruskal-Wallis computation, but the Kruskal-Wallis dialog will remain to perform another computation.

Cancel will close the dialog without doing any computation.

Friedman and Quade Tests

Introduction

The Friedman test is a nonparametric test used to compare several measures repeated on the same subjects. Unlike the parametric test, it makes no assumptions about the distribution of data, e.g. normality or homogeneity of variance. It is the nonparametric alternative of repeated measures ANOVA, and it useful when the assumptions of normality or equality of variance are not met.

The Quade test is very similar to the Friedman test, the difference being that the Quade test takes the order of the subject (block) into account and the rank of the subject is determined by the size of the sample range.

The user interface of Friedman and Quade is the same, and they both are under *Stat > Nonparametric Test* menu. The following details will use the Friedman dialog as the example.

Implementation Details

The data has to be balanced, meaning every subject must have all of the measurements. If for any subject, there are missing measurements, you can filter based on either the subjects or the measurements to make it balance.

The random variable X_{ij} is the rank in subject i under treatment j. i= 1,2,…b. j=1,2,…k. The sum of the ranks for each treatment to obtain:

$$R_j = \sum_{i=1}^{b} R(X_{ij}).$$

The Friedman test statistic:

$$T_1 = \frac{12}{bk(k+1)} \sum_{j=1}^{k} \left(R_j - \frac{b(k+1)}{2} \right)^2.$$

If there are ties present, an adjustment needs to be made. Let A_1 be the sum of the squares of the ranks and average ranks.

$$A_1 = \sum_{i=1}^{b} \sum_{j=1}^{k} \left[R(X_{ij}) \right]^2$$

Also, compute the "correction factor" C_1 given by:

$$C_1 = bk(k+1)^2 / 4$$

Then the statistic T_1 adjusted for the presence of ties becomes:

$$T_1 = \frac{(k-1)\left[\sum_{j=1}^{k} R_j^2 - bC_1 \right]}{A_1 - C_1}$$

Current research indicates the preferred statistic, because of its more accurate approximate distribution, is the two-way analysis of variance statistic computed on the ranks $R(X_{ij})$, which simplifies to the following function of T_1 given above.

$$T_2 = \frac{(b-1)T_1}{b(k-1) - T_1}$$

Quade test statistic is

$$T_3 = \frac{(b-1)B}{A_2 - B}$$

Let $A_2 = \sum_{i=1}^{b} \sum_{j=1}^{k} S_{ij}^2$, $B = \frac{1}{b} \sum_{j=1}^{k} S_j^2$

and $Q_i = \max imum\{X_{ij}\} - \min imum\{X_{ij}\}$

as the range of the ranks in subject i. $S_{ij} = Q_i \left[R(X_{ij}) - \frac{k+1}{2} \right]$ and $S_j = \sum_{i=1}^{b} S_{ij}$ for j=1,2,…,k.

Configuring the Friedman Test Dialog

Open the Friedman dialog by selecting **Stat > Nonparametric Tests > Friedman…** from the Partek main window. You will use this dialog to specify the

subject variable, the grouping variable (factor), the response variable(s) to be tested, and the multiple test corrections. Figure 11. 76 is configured to compare different treatments on each animal for all response variables.

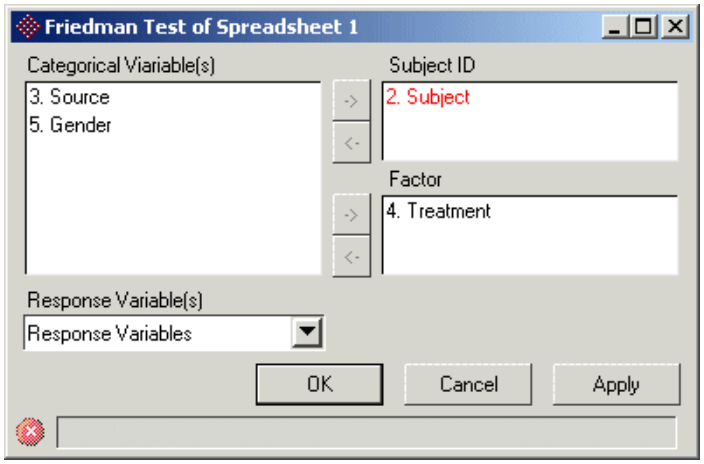


Figure 11. 76: Configuring the Friedman test dialog, multiple tests

Selecting Grouping Variables

The *Subject ID* and *Factor* must be selected from the *Categorical Variable(s)* list, which contains variables that have more than two categories (levels) in the spreadsheet. There can be only one subject variable and one factor variable in a Friedman test computation. When an item in the *Categorical Variable(s)* list box is selected, the -> button next to the *Grouping Variable* list box will be enabled. Click on it to move the selected item to the *Subject ID* or *Factor* list box. To remove a subject or a factor, select it and the <- button next to it will be enabled. Click on the <- button and the item selected will be moved back to the *Categorical Variable(s)* list box. The *Factor* variable must have two subgroups (levels).

Selecting Response Variables

By default, if there is more than one numeric variable in the spreadsheet, *Response Variables* will be shown as the *Response Variable(s)* to test all of variables at one time. To choose a specific response variable to test, select the variable name from the drop-down list.

When an analysis is performed on all numerical variables, the results will be summarized in a new spreadsheet that is a child of the original. In this child (results) spreadsheet, each variable tested in a row is summarized by the column number and the name of the variable, and followed by summary statistics including the p-values, means, and standard deviations. The rows are automatically sorted by the first column of p-values.

To sort by a different p-value, right click on the column heading and select **Sort Ascending** in the pop-up menu. Detailed reports about individual test variables can be viewed by right-clicking on the row label corresponding to the variable and selecting the **HTML Report** option on the pop-up menu (Figure 11. 77).

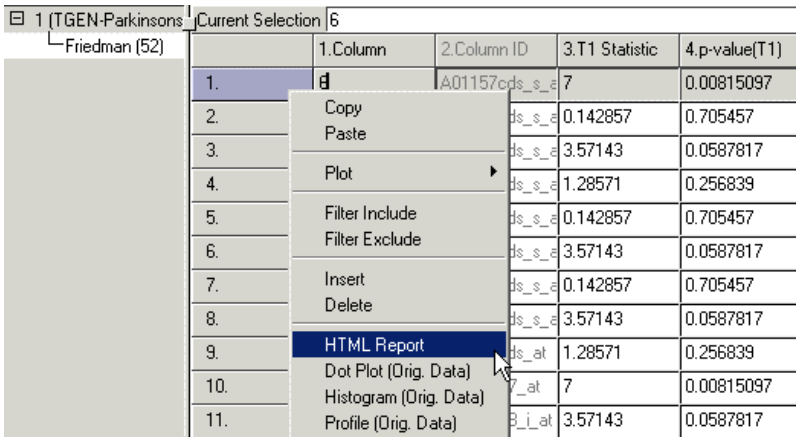


Figure 11. 77: Selecting an HTML report based on the result spreadsheet of the multiple tests with Friedman

To test a single response variable, select the variable from the drop-down list, the result will be displayed in a HTML report (Figure 11. 77).

Friedman test on A03913cds_s_at

Factor Level Information

Factor	Levels	Level Values
Type	2	substantia, ventral

Descriptive

Type	N	Mean Rank	Rank Sum	Median
substantia	7	1.571429	11.000000	2.000000
ventral	7	1.428571	10.000000	1.000000

Friedman

	T1	p-value(T1)	T2	p-value(T2)
Uncorrected	0.142857	0.705457	0.125	0.735765
Corrected for Ties	1.000000	0.317311	1.000000	0.355918

Figure 11. 78: Viewing the HTML report for a paired sample t-test, single test

Running the Computations

OK will perform the configured *t*-test computation and dismiss the dialog.

Apply will perform the configured *t*-test computation, but the paired sample *t*-test dialog will remain to allow for another computation.

Cancel will close the dialog without doing any computation.

Kolmogorov-Smirnov Test

Introduction

The Kolmogorov-Smirnov test is a nonparametric test used to compare the distribution of two variables. It makes no assumptions about the distribution of the data, e.g. normality or homogeneity of variance. It tests the maximal difference of the two distributions as well as the locations of the distribution. If you want to just compare the location of two distributions, which means comparing the ranks, you can use the Mann-Whitney test instead.

Implementation Details

In the One Sample Kolmogorov-Smirnov Test, there are n observations $X_1,...,X_n$.

Assumptions:

The sample is a random sample.

Hypothesis $H_0 : F(x) \le F^*(x)$ for all x from -∞ to +∞.

Procedure:

Let S(x) be the empirical distribution function based on the random sample $X_1,...,X_n$.

Define, for all x,

$$T = \max_x | F^*(x) - S(x) |$$

P value of T is calculated from 5000 simulation results.

In the Two Sample Kolmogorov-Smirnov Test, there are N = m + n observations $X_1,...,X_m$ and $Y_1,...,Y_n$.

Assumptions:

A1. The N X's and Y's are mutually independent.

A2. Each X comes from the same continuous population I_1.

A3. Each Y comes from the same continuous population I_2.

Hypothesis H_0: P(X ≤ a)=P(Y ≤ a), for all a.

Procedure:

1. Reorder the combined sample of N observations $X_1,\ldots,X_m,Y_1,\ldots Y_n$ with increasing in magnitude. Denote these ordered values by $Z_{(1)} \le Z_{(2)} \le ... \le Z_{(N)}$.
2. Define, for all a,

$$F_m(a) = \frac{\# X's \le a}{m}$$

$$G_n(a) = \frac{\# Y's \le a}{n}$$

$$d = \max_{i=1,...,20} \{ | F_{10}(Z_{(i)}) - G_{10}(Z_{(i)}) | \}$$

3. P value of d is calculated from 5000 simulation results.

Configuring the Kolmogorov-Smirnov Dialog

The two tested variables should be on 2 columns in the spreadsheet. To open the *Kolmogorov-Smirnov* dialog, select **Stat** > **Nonparametric Tests** > **Kolmogorov-Smirnov…** from the Partek main menu. The *KS* dialog (Figure 11. 79) is used to compare the distribution of one numeric variable to a certain distribution or distribution of another numeric variable.

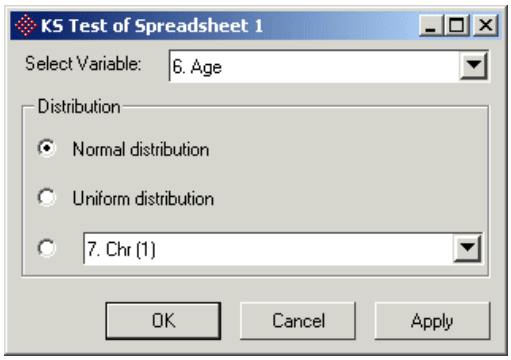


Figure 11. 79: Configuring the Kolmogorov-Smirnov dialog

Selecting a Variable

The *Select Variable* drop-down list contains all the numeric type of variables; simply click on the drop-down arrow and select the variable to test.

One-Sample KS Test

To test if the selected variable has a specified distribution, choose *Normal distribution* or *Uniform distribution*.

Two-Sample KS Test

To test the location and shape of the two variables, choose a variable from the drop-down list from the *Distribution* panel (Figure 11. 80).

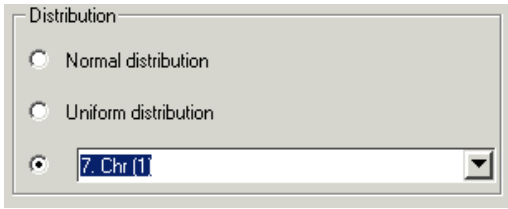


Figure 11. 80: Configuring the Two-Sample Kolmogorov-Smirnov test

Running the Computations

OK will perform the configured KS test computation and dismiss the dialog.

Apply will perform the configured KS test computation, but the dialog will remain to perform another computation.

Cancel will close the dialog without doing any computation.

Multiple Test Correction for P-Values

Introduction

A *p*-value is the probability that the observed values could have occurred by chance. It indicates the probability that one could obtain a test statistic that is as extreme as or more extreme than the observed one if the null hypothesis is true. *P*-values provide a sense of the strength of the evidence against the null hypothesis. The lower a *p*-value is, the stronger the evidence to reject the null hypothesis.

When multiple tests are performed, the probability of incorrectly rejecting a single null hypothesis ("false positive" or "Type I error") increases. There are several methods to correct Type I error for multiple tests in Partek. Options include Bonferroni, Dunn-Sidak, Bootstrap, and False Discovery Rate.

Bonferroni and Dunn-Sidak

The *Dunn-Sidak* method is used control the "experiment-wise Type I error rate" (α_e), which is the probability of making a single Type I error among all the hypotheses tested. Suppose you test K *independent* hypotheses, each at the comparison-wise significance level α_c. If all the null hypotheses are true, the probability of making zero Type I errors is $(1- \alpha_c)^K$. Hence the overall significance level (adjusted for multiple tests) is $\alpha_e=1- (1- \alpha_c)^K$. The overall significance level can be adjusted, or the individual p-values can be adjusted by using $p_S=1-(1-p_c)^{1/K}$, where p_S is the Dunn-Sidak corrected p-value and p_c is the unadjusted p-value.

An approximation to the Dunn-Sidak can be computed using the Bonferroni correction. If testing a K *independent* hypotheses, the expected number of Type I errors would be given by $K*\alpha$, thus the significance level of each individual test α' should be adjusted to $\alpha_e =\alpha_c /K$. Alternatively, the *p*-values may be adjusted as $p_B= p_c *K$, where p_B is the Bonferroni corrected *p*-value and p_c is the unadjusted *p*-value. If p_B is greater than 1, it is set to 1.

For corrected *p*-values that are below 0.05, Dunn-Sidak and Bonferroni are nearly identical. For adjusted *p*-values greater than 0.05, the Dunn-Sidak is more conservative and more correct.

Both the Dunn-Sidak and Bonferroni corrections are generally considered overly conservative for two reasons:

- They assume all the tests are independent, which may not be true in many real world applications. The result is a corrected *p*-value that may be larger than it should be.
- They protect against even a single false positive, which may be too strict if thousands of tests are being conducted.

Both Bonferroni and Dunn-Sidak methods either reduce the alpha level or adjust the *p*-value for each individual test. When either of these methods is selected, Partek augments the test report for each hypothesis to include the corrected *p*-value in addition to the adjusted *p*-value.

The Bonferroni and Dunn-Sidak methods can be performed in any spreadsheet that contains p-value columns. Select **Stat > Multiple Test Correction** from the Partek main menu to invoke the dialog (Figure 11. 81).

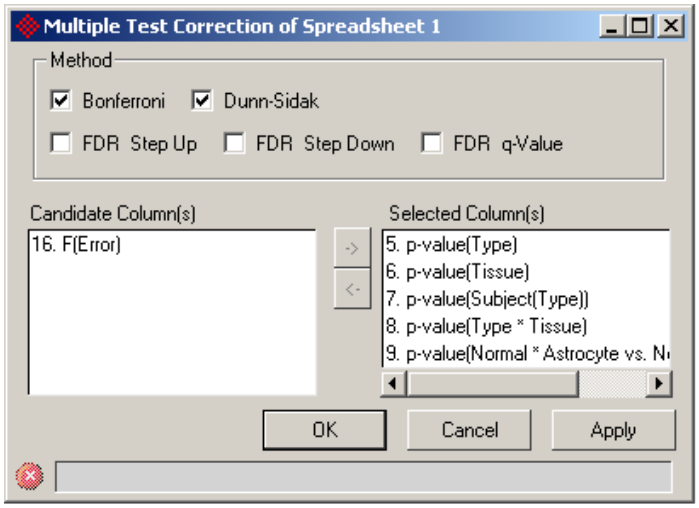


Figure 11. 81: Bonferroni and Dunn-Sidak dialog

Selecting Columns

The Bonferroni and Dunn-Sidak methods can be computed only on columns whose values are between 0 and 1 (inclusive). These are called *Candidate Column(s)* in the dialog. If the label of a candidate column has a prefix of "*p*-value", the column will be automatically selected and will appear in the *Selected Column(s)* list box, by default. Corrections can be computed on more than one p-value column at a time. To select or deselect the columns to compute, click on the item in the *Candidate Column(s)* or *Selected Column(s)* list box, and click on the corresponding arrow to move the item to the opposite box. In the *Candidate Column(s)* box, the items are sorted by the column number; in the *Selected Column(s)* box, the items are sorted by the order of selection.

Method

You need to select at least one method to compute by selecting the checkbox. Click **OK** or **Apply** to compute, the adjusted p-value will be inserted to the right of the corresponding p-value columns on the spreadsheet.

Bootstrap

Partek provides a bootstrap method to perform multiple tests correction in the two-sample t-Test, Mann-Whitney, and Kruskal-Wallis tests. The bootstrap is used to determine the probability of obtaining a particular *p*-value by chance. For instance,

suppose you do a t-test on 7,129 variables and are interested in knowing if an unadjusted *p*-value of 5.48787e-6 is significant when considering the multiple tests. This example corresponds to the 122nd ranked gene on an leukemia dataset. In this example, Partek will perform 1,000 iterations of the bootstrap (Figure 11. 82).

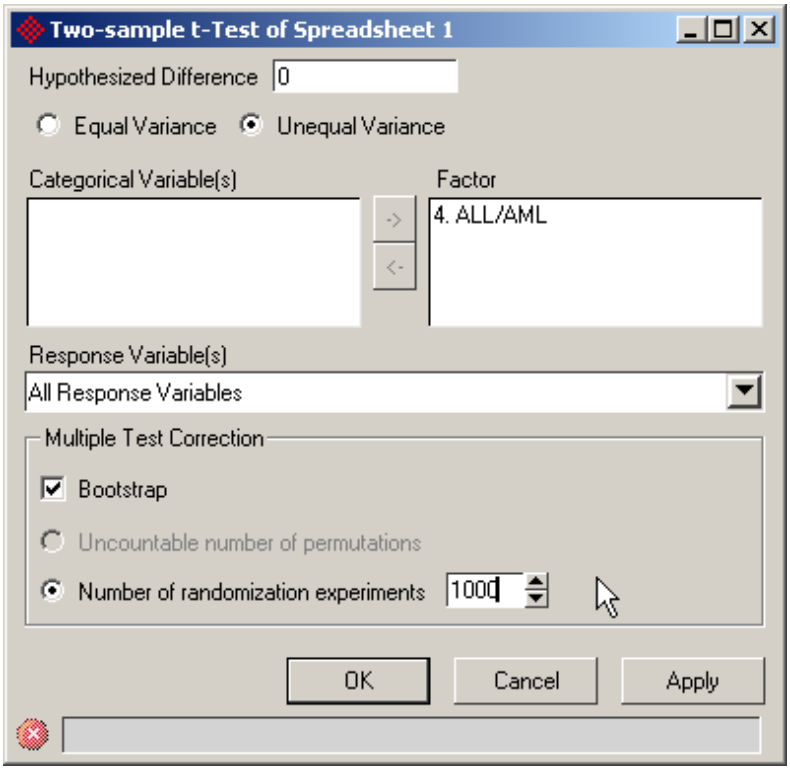
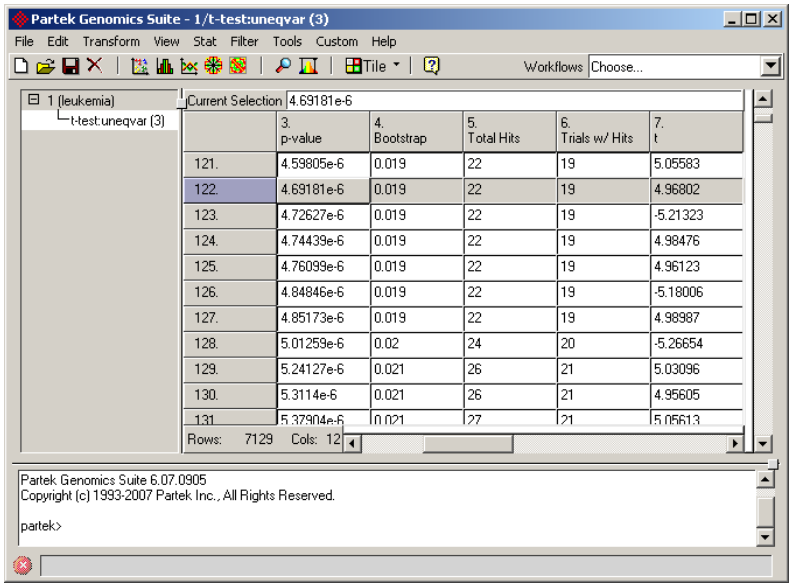


Figure 11. 82: Configuring the two-sample t-Test with Bootstrap correction

Each iteration of the bootstrap does the following:

- Randomly reassigns the 72 labels for each row so that there are always 47 ALL's and 25 AML's
- Runs a t-test on all 7129 genes
- Increments by one the column labeled “Trials w/ Hits” if *any* of the 7129 p-values is less than or equal to 4.69181e-6 (Figure 11. 83)



	3. p-value	4. Bootstrap	5. Total Hits	6. Trials w/ Hits	7. t
121.	4.59805e-6	0.019	22	19	5.05583
122.	4.69181e-6	0.019	22	19	4.96802
123.	4.72627e-6	0.019	22	19	-5.21323
124.	4.74439e-6	0.019	22	19	4.98476
125.	4.76099e-6	0.019	22	19	4.96123
126.	4.84846e-6	0.019	22	19	-5.18006
127.	4.85173e-6	0.019	22	19	4.98987
128.	5.01259e-6	0.02	24	20	-5.26654
129.	5.24127e-6	0.021	26	21	5.03096
130.	5.3114e-6	0.021	26	21	4.95605
131.	5.37904e-6	0.021	27	21	5.05613

Figure 11. 83: Viewing the result of the two-sample t-Test with 1,000 iterations of the Bootstrap correction

The 72 labels are randomly reassigned again and again, for a total of 1,000 random assignments. It is possible that the same assignment will be chosen more than once, causing this to be a bootstrap (random sampling *with replacement*).

In the end, 16 of the 1000 trials gave a *p*-value at least as small as 4.69181e-6, therefore the probability of getting a *p*-value as small as 4.69181e-6 by chance is 16/1000 or .016.

By comparing this result to a Dunn-Sidak or Bonferroni, notice that the bootstrap is not as conservative as the other corrections because it does not assume that the tests are independent. However, the bootstrap is still conservative because it still protects against a single false positive.

False Discovery Rate (FDR)

False Discovery Rate is the most lenient multiple test adjustment available in Partek. It is a compromise between the uncorrected analysis of the multiple tests and family-wise error rate. FDR is the proportion of false positives among all positives. Partek implements the step up (Benjamini & Hochberg, 1995), step down (Benjamini and Liu, 1999) and q-Value (Storey, J.D., 2003) methods to control the false discovery rate.

In the step up method, there are n p-values; they are sorted by ascending order, and m represents the rank of a p-value. The calculation compares p-value*(n/m) with the specified alpha level, and the cut-off p-value is the one that generates the last product that is less than the alpha level.

The goal of step up method is to find:

$$k^* = \max\left\{m:\ P_m \le \frac{m}{n}\cdot\alpha\right\} = \max\left\{m:\ P_m \cdot \frac{n}{m} \le \alpha\right\}$$

Define the step-up value as:

$$S_m = P_m \frac{n}{m}$$

Then, an equivalent definition for k* is:

$$k^* = \min\{j: S_m > \alpha \ for\ all\ m\ between\ (j+1)\ and\ n\}$$

So when $S_m > \alpha$ and $S_{m-1} > S_m$, then $S_{m-1} > \alpha$, the step up value is:

$$S_n^* = P_n$$

$$S_{m-1}^* = \min\{S_{m-1}, S_m^*\}$$

In order to find k^*, start with S_n^* and then go up the list until you find the first step up value that is less or equal to alpha.

In the step down method, the p-values are sorted in descending order, and the calculation compares p-value*n/(n+1-m). The cut-off p-value is the one that generates the first product that is less than alpha level.

In the q-Value method, q-value is the minimum "positive false discovery rate" (pFDR) that can occur when rejecting a statistic.

For an observed statistic T = t and a nested set of rejection area {C},

$$q-value(t) = \min_{\{C: t\in C\}} pFDR(C)$$

$$pFDR(C) = \frac{\pi_0 \cdot \Pr ob(T \in C \mid H = 0)}{\Pr ob(T \in C)} = \Pr ob(H = 0 \mid T \in C)$$

To calculate the FDR on a result spreadsheet containing unadjusted *p*-values, select **Stat > Multiple Test Correction** from the Partek main menu. Use this dialog to compute FDR on specific *p*-value columns (Figure 11. 84).

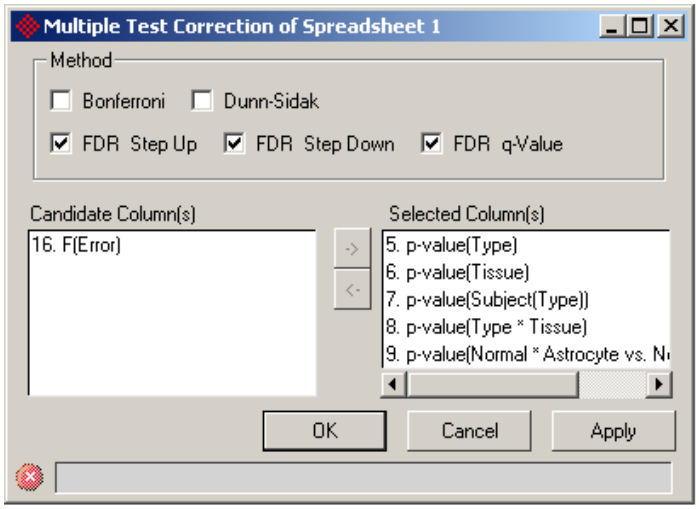


Figure 11. 84: Configuring the False Discovery Rate dialog

Selecting Columns

FDR can only be computed on columns whose values are between 0 and 1 (inclusive). These are called *Candidate Column(s)* in the dialog. By default, if the label of a candidate column has a prefix of "*p*-value", the column will be automatically selected and will appear in the *Selected Column(s)* list box. Partek can compute FDR on more than one p-value column at a time. To select or deselect the columns to compute FDR, click on the item in the *Candidate Column(s)* or *Selected Column(s)* list box, the corresponding arrow will be enabled, and clicking on it will move the item to the opposite box. In the *Candidate Column(s)* box, the items are sorted by the column number; in the *Selected Column(s)* box, the items are sorted by the order of selection.

Method

You need to select at least one method to compute by selecting the checkbox. Click **OK** or **Apply** to compute, the FDR values will be inserted to the right of the corresponding p-value columns on the spreadsheet.

Power Analysis

Power analysis procedure conducts prospective analysis which is used to

- Determine the minimum sample size to achieve adequate power on a given fold change
- Determine what fold change could be acquired on the given sample size to achieve the specified power

Implementation Details

Input for *Power Analysis* includes:

- Experimental design
- Statistical model (ANOVA)
- Comparison (contrast) on which to do power analysis
- Effect size (fold change)
- Sample size
- Significance level (alpha)
- Power (1-beta)

Power Analysis obtains the experimental design, statistical model (ANOVA) and comparison (contrast) from the current study.

Let Y be the response vector, X be the design matrix, β be the model parameter vector, so the underlying function for the ANOVA model can be written in the form of $Y = \beta X + \varepsilon$ where ε is the error term which is normally and independently distributed with mean 0 and standard deviation σ. *Comparison (contrast)* was set in the *Estimate Gene Significance (ANOVA)* step to test the null hypothesis $H_0 : \mathrm{L}\beta = 0$ where L is the contrast matrix.

For the four parameters like effect size, sample size, significance level and power, each can be obtained by solving the following power analysis formulas when fixing the other three.

Power Analysis Formulas

$power = P(F(r_L, N - r_x, \lambda) \geq F_{1-\alpha}(r_L, N - r_x))$ (Muller and Peterson 1984)

Where r_L is the rank of contrast L, r_x is the rank of design matrix X, N is the total sample size, α is the significance level and λ is the non-central parameter of F statistic under alternative hypothesis $H_A : L\beta \neq 0$.

$$\lambda = N(L\beta)'(L(\ddot{X}'diag(w)\ddot{X})^{-1}L')^{-1}(L\beta)\sigma^{-2}$$

Where $\ddot{X}$ is composed of the unique rows of design matrix X, w is a vector of weights which reflect the proportion of each unique row in the whole design matrix X. σ is the ANOVA model standard deviation.

Configuring Power Analysis Dialog

To invoke the dialog, select ***Stat>Power Analysis*** *form the menu to open (Figure 11. 89)*

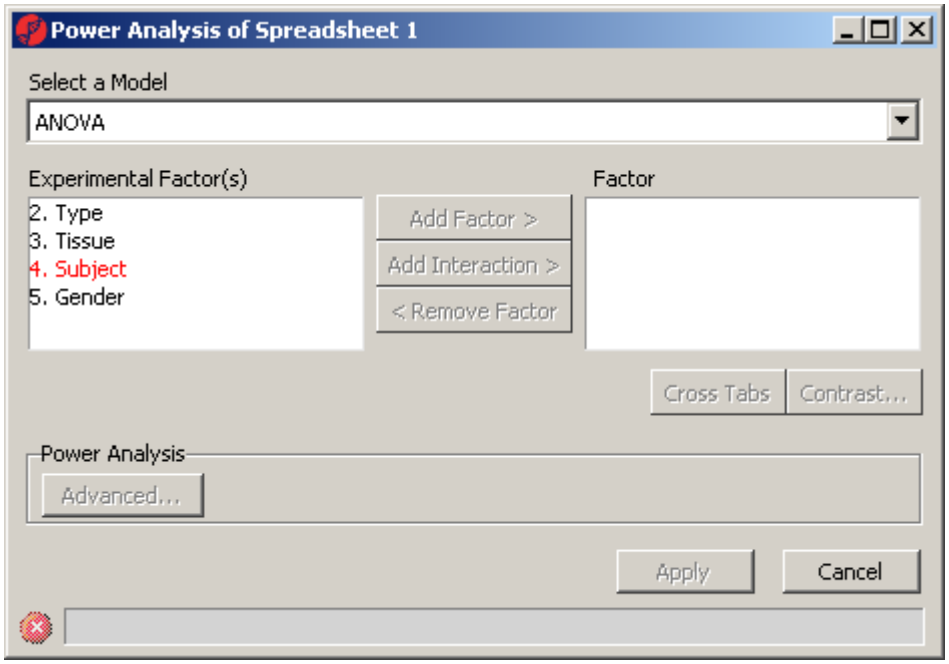


Figure 11. 85: Configuring Power Analysis Model and Factors

Select a model allows you to select *One Sample t-test*, *Two Sample t-test* and *ANOVA*. If ANOVA is chosen, *Contrast* needs to be specified, only one comparison can be selected to do power analysis at a time. One sample t-test doesn't need any factor to be selected; Two sample t-test doesn't need contrast to be specified.

Configuring the Effect Size

Selecting the **Advanced...** button in the *Power Analysis* frame will open the *Power Analysis Configuration* dialog (Figure 11.90) to configure the parameters of effect size, sample size, significance and power.

Specify the range and step size for effect size in this dialog so that the *Power Analysis* will produce the minimum sample sizes (the newly produced sample size is supposed to be assigned to each comparison group with the same proportion as the original dataset) required to achieve each of the specified effect sizes, respectively. Effect size (fold change here) must be greater than or equal to one. Decreasing the effect size will probably require more samples. For better viewing, 10 points of effect size can be accommodated in the specified range by the specified step size.

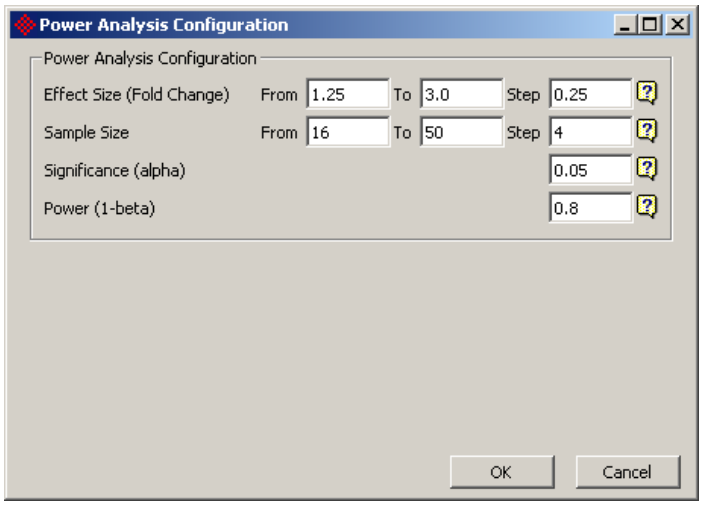


Figure 11. 86: Configuring Power Analysis

Configuring the Sample Size

Specify the range and step size for the sample size so that the *Power Analysis* will produce a fold change that is set by the given sample sizes. Sample size should be larger than model's degree of freedom. For better viewing, 10 points of sample size can be accommodated in the specified range by the specified step size.

Configuring the Significance

The significance level is the probability to reject the null hypothesis when the null hypothesis is actually true. A commonly used significance level of 0.1 is set as the default. The range for significance level is between 0 and 1. Decreasing the significance level will probably require more samples to achieve the same fold change.

Configuring the Power

The power level is the probability to reject the null hypothesis when the null hypothesis is actually false. A commonly used power of 0.8 is set as the default. The range for power is between 0 and 1. Increasing the power will probably require more samples to achieve the same fold change.

Saving the Power Analysis Configuration

Selecting **OK** in the *Power Analysis Configuration* dialog (Figure 11.90) will save all the parameters configured and dismiss the dialog; selecting **Cancel** will close the dialog without saving.

Running the Power Analysis

Selecting **OK** in the *Power Analysis* dialog (Figure 11.89) will perform the configured power analysis and dismiss the dialog; selecting **Cancel** will close the dialog without doing any computation.

Visualizing the Data: Box plot

The box plot provides a way to graphically view the numeric data through five numbers in summary. The five numbers, 10th percentile, 25th percentile, 50th percentile, 75th percentile and 90th percentile of the power analysis, result in the gene level. *Partek Express* Power Analysis will generate two box plots, *Fold Change to Sample Size* and *Sample Size to Fold Change*. These two box plots can be invoked by selecting the radio button on the *Power Analysis* tab in the *Partek Express* main window.

Box Plot: Fold Change to Sample Size

The *Fold Change to Sample Size* box plot indicates the sample size (in Y axis) to achieve the adequate power of the given fold change (in X axis).

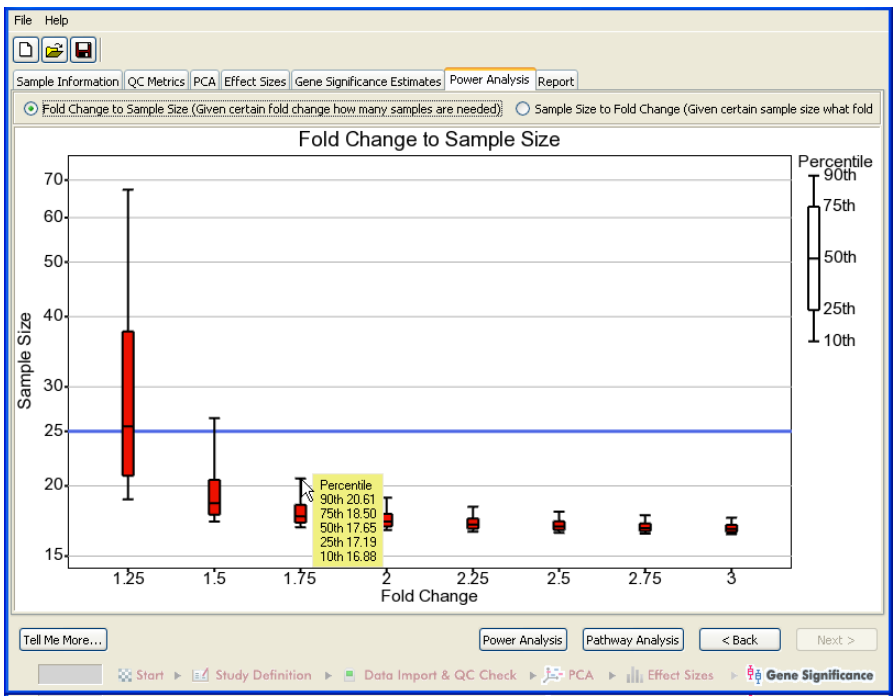


Figure 11. 87: Box plot of Fold Change to Sample Size

Note: Y axis tick marks are in log (base 2) scale. The current study sample size is marked with a blue reference line. Moving the mouse over a box-whisker will show a more detailed sample size report. In the examples shown in figure 3, to detect 50% of genes with a fold change of at least 1.75 would require 17.65 (round up to 18) samples.

Note: Power Analysis for a specific fold change assumes the proportion of samples in each category is similar to that of the existing samples. Table 11. 1 shows the number of samples needed for a fold change of at least 1.75.

# of Samples	Percent of Genes	Fold Change
16.88	10%	1.75

17.19	25%	1.75
17.65	50%	1.75
18.50	75%	1.75
20.61	90%	1.75

Table 11. 1: Viewing the number of samples needed to achieve a fold change of at least 1.75

Box Plot: Sample Size to Fold Change

The *Sample Size to Fold Change* box plot shows what fold change (in X axis) could be acquired on the given sample size (in Y axis).

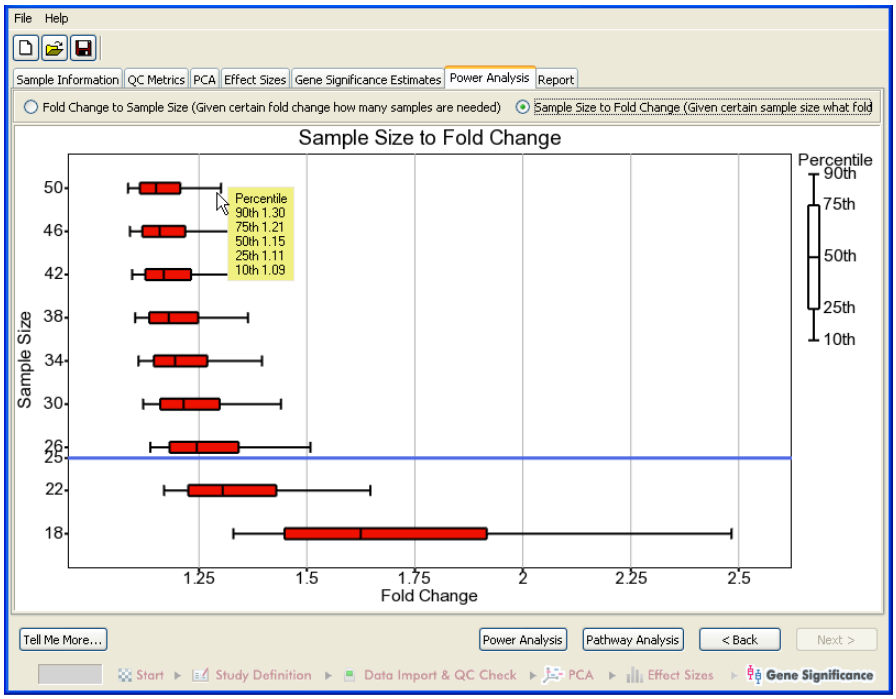


Figure 11. 88: Box plot of Sample Size to Fold Change

The blue line marks the number of samples used in the analysis of the study. Moving the mouse over a box-whisker will bring up the detailed fold change report for the respective samples size. In the example shown in **Error! Reference source not found.**, using 50 samples in the study would detect 50% of genes with a fold change of at least 1.15.

Note: Power Analysis for a specific sample size assumes the proportion of samples in each category is similar to that of the existing samples. Table 11. 2 shows the fold change of 50 samples at varying percentages.

# of Samples	Percent of Genes	Fold Change
50	10%	1.09
50	25%	1.11
50	50%	1.15
50	75%	1.21

50	90%	1.30

Table 11. 2: Viewing the fold change of 50 samples at differing percentages

Correspondence on Threshold

- Open the dialog by selecting **Tools > Correspondence on Threshold**

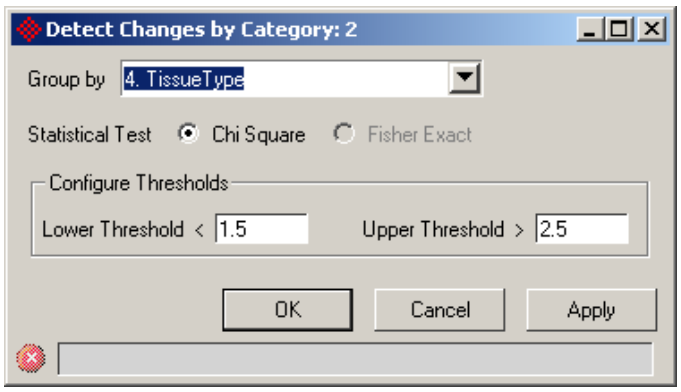


Figure 11. 89: Test Counts from threshold dialog

From this dialog, you can run *Chi Square* or *Fisher Exact* tests on contingency tables built using a categorical column and two thresholds.

The Fisher Exact test is only available on columns with two categories. For each column, two tests are run (resulting in two p-values). For both tests, the x-axis of the contingency table is determined by the categories of the Group by column.

For the test to get the first (upper) p-value the top cells in the contingency tables are the counts of the values in the spreadsheet that are greater than the upper threshold. The lower cells are those that are less than or equal to the upper threshold.

For the test to get the second (lower) p-value, the top cells in the contingency tables are the counts of the values in the spreadsheet that are less than the lower threshold. The lower cells are those that are greater than or equal to the lower threshold.

Along with the p-values, the result spreadsheet contains (for each class) the counts of the samples that exceed the lower threshold and those that exceed the upper threshold.

The Chi Square test results in a missing value if any row or column total is zero.

The html report will have the Chi-square value for the Chi Square test and the left and right tail p-values for the Fisher Exact test.

Test on Genomic Window

- Open the *Test on genomic window* dialog by selecting **Tools > Test on Genomic Window**. Note: this option only appears for spreadsheets that have genomic features on columns

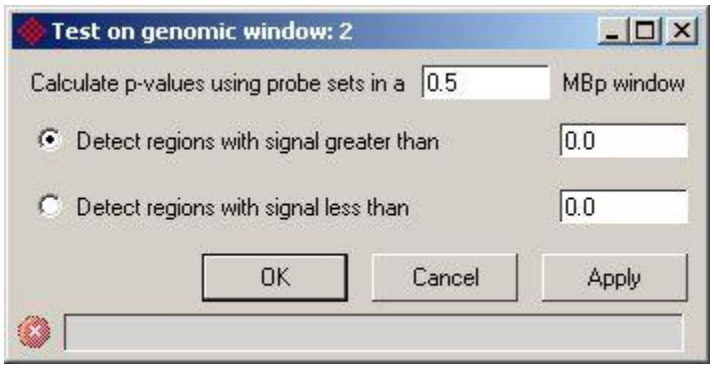


Figure 11. 90: Test on genomic window dialog

This test results in a p-value for each sample and each probe set. The test for a given probe set uses the probe sets on the same chromosome that are in a window centered on it (up to half the window size in each direction).

You can choose to test for signals greater than a fixed value or less than a fixed value.

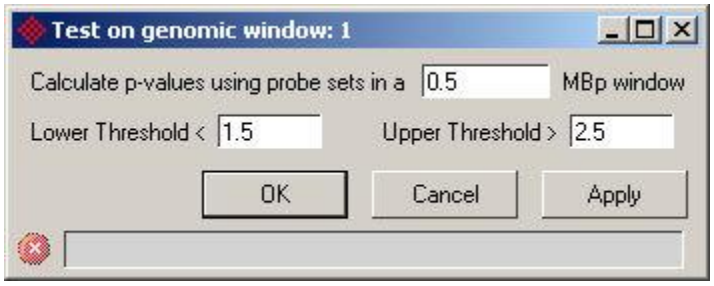


Figure 11. 91: Test on genomic window dialog on copy number

If the spreadsheet has the "copy number" property (**File > Properties**) then two tests are run and the lower p-value is used. If the mean of the region is above the lower threshold and below the upper threshold then the p-value will be 1.

Detect Significant Regions

Open the *Detect Significant Region* dialog by selecting **Tools > Detect Significant Regions**. Note: this option only appears for spreadsheets that are the result of a statistical test.

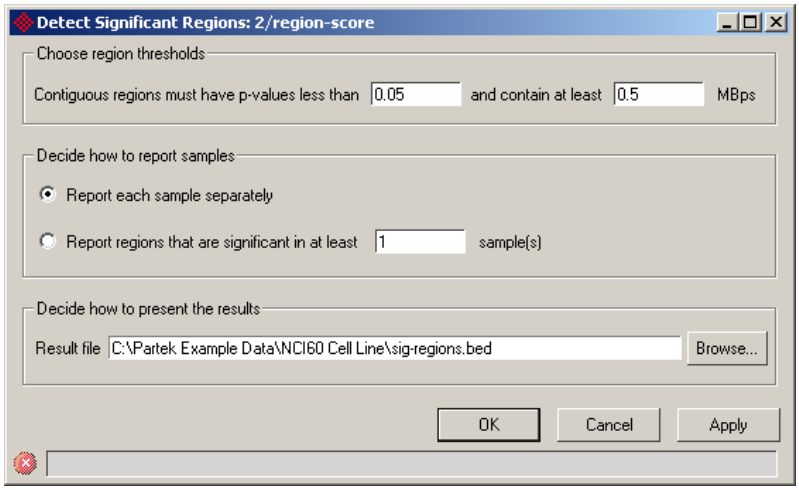


Figure 11. 92: Detect Significant Regions dialog from test on genomic window

If the spreadsheet has more than one p-value column, then choose the column from which you want to detect regions. If the spreadsheet is the result of **Stat > Test on genomic window** you will not need to specify the p-value column.

The result file will be saved in the bed format and will also be opened as a child spreadsheet. More information about the bed format can be found at http://genome.ucsc.edu/FAQ/FAQformat.

References

Agresti, Alan (2002). Categorical Data Analysis, New York: John Wiley & Sons, Inc., 165 -196.

Benjamini, Y., Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing, JRSS, B, 57, 289-300.

Benjamini, Y. and Liu, W. (1999). A step-down multiple hypotheses testing procedure that controls the false discovery rate under independence. *J. Statist. Plann. Inference* 82: 163--170.

Efron, B.(1977). The efficiency of Cox's likelihood function for censored data, Journal of the American Statistical Association 72:557-565.

Eisenhart, C. (1947). The assumptions underlying the analysis of variance. Biometrics, 3: 1-21.

Fisher, R.A. (1925), *Statistical Methods for Research Workers*, Edinburgh: Oliver & Boyd.

Greenhouse, S.W. and Geisser, S. (1959), On Methods in the Analysis of Profile Data, Psychometrika, 32, 95-112.

Huynh, H. and Feldt, L.S. (1970), Conditions under Which Mean Square Ratios in Repeated Measurements Designs Have Exact F-Distributions, Journal of the American Statistical Association, 65: 1582-1589.

Huynh, H. and Feldt, L.S. (1976), Estimation of the Box Correction for Degrees of Freedom from Sample Data in the Randomized Block and Split Plot Designs, Journal of Educational Statistics, 1: 69-82.

Kaplan, E.L. and Meier, P. (1958), Nonparametric estimation from incomplete observations, Journal of the American Statistical Association, 53: 457-481.

Muller, K.E. and Peterson, B. L. (1984), Practical methods for computing power in testing the multivariate general linear hypothesis. *Computational Statistics and Data Analysis*, 2: 143-158.

Muller, K.E. and Benignus, V.A. (1992), Increasing scientific power with statistical power. *Neurotoxicology and Teratology*, 14: 211-219.

Muller, K.E., LaVange, L.M., Ramey, S.L. and Ramey, C.T. (1992), Power calculations for general multivariate models including repeated measures applications. *Journal of the American Statistical Association*, 87: 1209-1226.

Peto R, Peto J. Asymptotically efficient rank invariant test procedures. J Poyal Stat Sco A 1972;135:185-206.

Rao, C.R. (1971). Estimation of Variance and Covariance Components-MINQUE Theory. Journal of Multivariate Analysis I, 257-275.

Scheffé, H. (1959). *The analysis of variance*. New York: Wiley.

Searle, S.R. (1971), *Linear Models*, New York: John Wiley & Sons, Inc.

Sidak, Z. (1967), "Rectangular Confidence Regions for the Means of Multivariate Normal Distributions," Journal of the American Statistical Association, 62, 626 -633.

Snedecor, G.W. and Cochran, W.G. (1980), *Statistical Methods*, Seventh Edition, Ames, IA: Iowa State University Press.

Steel R.G.D. and Torrie, J.H. (1980), *Principles and Procedures of Statistics*, Second Edition, New York: McGraw-Hill Book Co.

Storey JD. (2003) The positive false discovery rate: A Bayesian interpretation and the q-value. *Annals of Statistics*, 31: 2013-2035.

Thompson, W.A., Jr (1962). The Problem of Negative Estimates of Variance Components. Ann. Math. Stat. 33: 273-289.

Diagnostic & Predictive Modeling

Introduction

Partek offers discriminate analysis, classification model selection, and variable selection for diagnostic and predictive modeling.

Classification Model Selection

The *Classification Model Selection* tool can be used to evaluate multiple models in one run to select an optimal model and to produce an unbiased estimate of prediction accuracy. In order to find the optimal model, the following questions must be answered:

- For variable selection, how many and which variables are going to be used by the classification model?
- What are the optimal parameters for the classifier? For example, for a K-Nearest Neighbor (KNN) classifier, what number of neighbors and which type of distance measure will be used? For a neural network, many more parameters must be determined, such as the number of hidden layers, number of neurons on each layer, the learning rate, training iterations, etc.

The number of variables, classifier types considered, and the parameters for each classifier type define the model space, which will be searched to find the best predictive model. Partek uses a two-level, nested cross-validation (CV) to solve this problem. Background on the technique of cross-validation can be found in various sources such as Stone, 1974; Geisser, 1975; and Efron & Tibshirani, 1993.

Simple, Single-Level Cross-Validation

Figure 12. 1 shows the partitions of a 10-fold cross-validation. The data is first divided into 10 random partitions. At each iteration, 1/10 of the data is held out for testing while the remaining 9/10 of the data is used to fit the parameters of the model. Simple, single-level cross-validation can be used to obtain an estimate of prediction accuracy for a single model.

Figure 12. 1: 10-fold 1-Level Cross-Validation

Two-Level, Nested Cross-Validation for Model Selection

While simple single-level cross-validation can estimate the error for a single model, it cannot at the same time be used to select between multiple candidate models. When multiple models are considered, it is not valid to report the error estimate of the best model as determined via single-level cross-validation (using cross-validation results to select an optimal model makes use of the test data for model selection).

In Partek, two-level nested cross-validation is used to select an optimal classifier and estimate the accuracy of the optimal classifier when multiple classifiers are considered for a single problem. In this approach, an "outer" cross-validation is performed to produce an unbiased estimate of prediction error (by holding out samples as an independent test set). In this way, the outer cross-validation serves the same purpose as simple single-level cross-validation. An additional "inner" level cross-validation is performed on the training data (the data not held out as test data by the outer cross-validation) to select the optimal model to be applied to the held out test set.

Consider the case of a 10x10 two-level nested cross-validation. In the outer cross-validation, with 10% of samples held out as test cases, the remaining 90% are used in a 10-fold cross-validation to determine the optimal predictor variables and other classifier parameters. The model that performed the best on the inner cross-validation is applied to the held-out test samples in the outer cross-validation. In the case when multiple models are tied for the best on inner cross-validation, all of the tied models will be applied to the held out samples and an average classification rate of the tied models will be used for the error estimate for the outer cross-validation pass. Thus, an "inner" cross-validation is performed in order to select predictor variables and optimal model parameters, and an "outer" cross-validation is used to produce overall accuracy estimates for the classifier. This is referred to as a "10x10 two-level nested cross-validation". The number of partitions in the inner and outer cross-validations does not have to be the same.

Opening the Model Selection Dialog

- To open the *Model Selection* dialog (Figure 12. 2), select **Tools > Predict > Model Selection** from the Partek main window

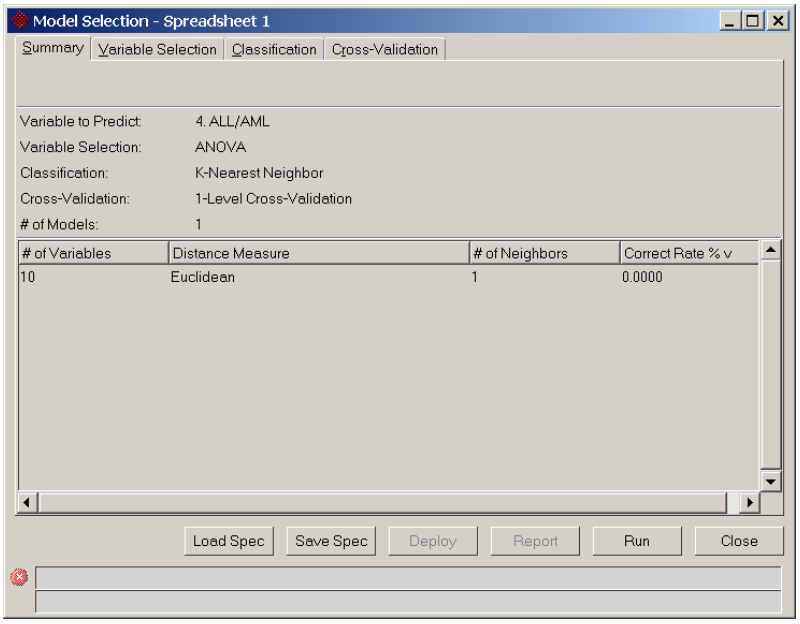


Figure 12. 2: Configuring the Model Selection dialog

- On the *Summary* tab (Figure 12. 3), the *Variable to Predict* field shows the categorical variable in the current spreadsheet that will be predicted

The *Variable Selection*, *Classification* method, and the level of *Cross Validation* are also displayed in the panel and will be discussed in further detail later.

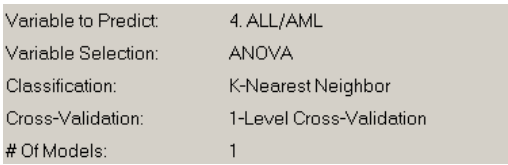


Figure 12. 3: Viewing the summary panel of the Model Selection dialog

The lower panel of the dialog shows all the models (in this example there is only one) with their parameters. A K-Nearest Neighbor model is identified by *# of Variables*, *Distance Measure*, and *# of Neighbors*. During and after running a test experiment, the *Correct Rate %* field shows the test score of that model.

Deploying a Selected Model

When there is a highlighted model in this panel, the *Deploy* button will be enabled. Clicking on the *Deploy* button will train the model with all samples and save the trained model to a *Partek Black Box (.pbb)* file.

The *# of Models* field shows how many models are defined. *Run* executes the *Model Selection*. *Report* shows a detailed HTML report of the results.

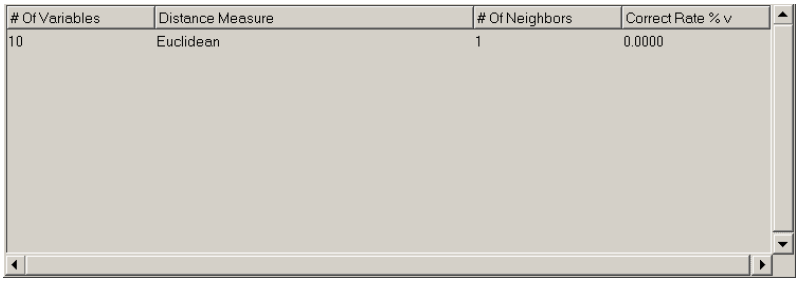


Figure 12. 4: Viewing the lower panel of the Model Selection dialog

Variable Selection

Variable selection is beneficial in a classification problem when the number of variables is greater than the number of samples/observations. It improves the classification accuracy due to noise reduction and cuts the cost of acquiring a large set of variables. It is especially helpful to those classifiers, which suffer from the "Curse of Dimensionality", such as discriminant analysis and neural networks.

There are two approaches of variable selection depending on whether a classifier is used. The first technique is called the filter approach, which does not have a classifier. The second technique is called the wrapper approach, which internally requires a classifier.

The *# of Predictor Variables* in the current spreadsheet is shown at the top of the *Variable Selection* page. Choose a variable selection method from the list shown in Figure 12. 5.

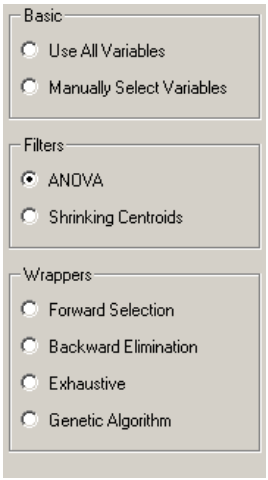


Figure 12. 5: Configuring the Variable Selection choices

Basic methods are *Use All Variables* and *Manually Select Variables*. Filter approach methods are *ANOVA* and *Shrinking Centroids*. Wrapper approach methods are *Forward Selection, Backward Elimination, Exhaustive,* and *Genetic Algorithm.* Depending on the variable selection method chosen on the left, the large panel on the right will allow you to configure more parameters.

Basic Methods

Using All Variables

In Figure 12. 6, *Use All Variables* was selected, so there are no additional parameters in the panel. Instead, a message of *All 7129 Predictor Variables will be used* is shown, so all the variables are in one group. This group will be used in the variable selection.

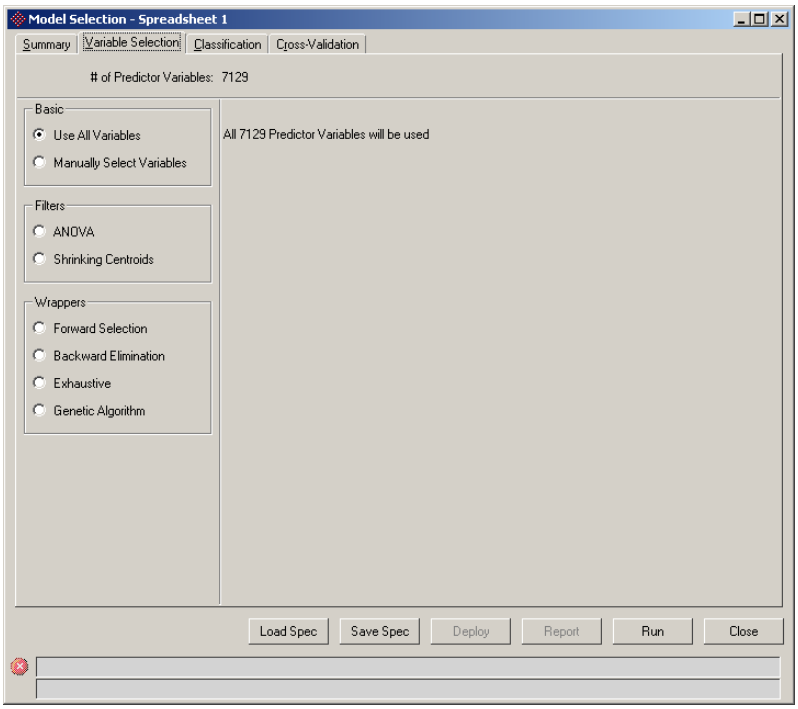


Figure 12. 6: Configuring the Variable Selection - Use All Variables dialog

Manually Select Variables

You can manually select variables by choosing the *Manually Select Variables* option. Two list boxes will appear in the panel (Figure 12. 7). The *Variable Candidates* list box includes all appropriate numerical variables in the spreadsheet; select the desired variables in the *Variable Candidates* list box, and click on the "->" button to move them into the *Always include these variables* list box (select). Similarly, to remove variables from the *Always include these variables* list box, select them and click the <- button to move them back (deselect). The manually included variables are placed in one group. This group will be used in the variable selection.

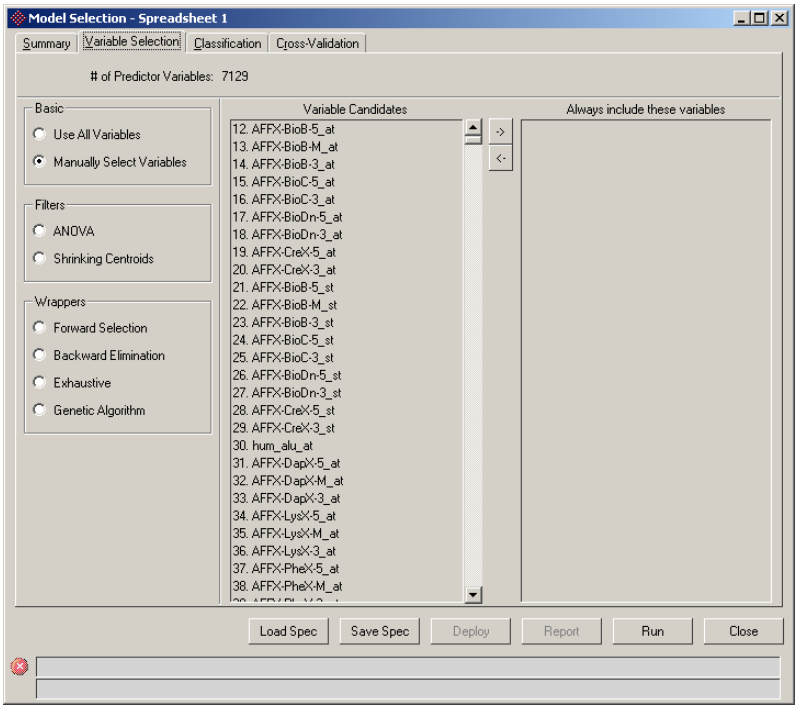


Figure 12. 7: Configuring the Variable Selection - Manually Select Variables dialog

Filters

ANOVA

By default, a one way ANOVA on the designated *Class* variable is used to select variables; however, to use a different ANOVA method, click the **Configure** button. The *ANOVA* dialog will appear. Select variables by using the -> <- buttons to move them into and out of the selected variable list. After specifying the parameters, click **Save**. The new ANOVA specification will be shown accordingly. The panel in Figure 12. 8 shows that a 2-way ANOVA will be used and the ALL/AML p-values will be examined.

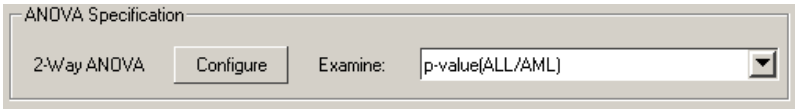


Figure 12. 8: Configuring the ANOVA Variable Selection Specification panel

The *p*-values to be calculated by the ANOVA are shown in the *Examine* drop-down list. The variables selected are determined by examining a specific p-value. By default, the *p*-value on the class variable will be chosen. Variables with the smallest *p*-values are selected as predictors.

Specifying the Number of Variable Groups to Try

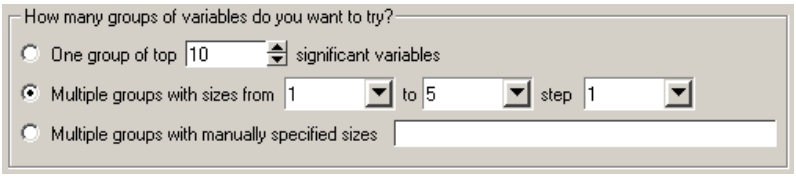


Figure 12. 9: Selecting the Number of Variables panel

Selecting a Fixed Number of Variables

To use fixed number of variables, select the *One group of top significant variables* and designate a number. By giving a number here, e.g. 10, the top 10 variables will be used for classification.

Multiple Groups with Sizes From-To-Step

Multiple groups with sizes from- to-step and *Multiple groups with manually specified sizes* are useful when the optimal number of variables for the classification is not known, because both options try multiple groups. In *Multiple groups with sizes from-to-step*, three numbers can be specified: from, to, and step. For example, you may want to try the top 10, top 20, top 30, …, up to the top 100 significant variables by giving the parameters of *Multiple groups with sizes from 10 to 100 step 10* (10 groups).

Multiple Groups with Manually Specified Sizes

Multiple groups with manually specified sizes allow you to arbitrarily specify the sizes of the variable subsets to evaluate. Table 12. 1 shows examples of different configurations for this option.

Manual Specification	Number of Genes Evaluated
5-10 50 100	5, 6, 7, 8, 9, 10, 50, 100
10-100-10	10, 20, 30, 40, 50, 60, 70, 80, 90, 100
100-1000-100	100, 200, 300, 400, 500, 600, 700, 800, 900, 1000
10-100-10 100-1000-100	10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000

Table 12. 1: Manually Specified Groups

Figure 12. 10: Configuring the manually specifying 19 sizes of variable subsets panel

In addition to the variables selected by ANOVA, the *Add these manually selected variables* option allows you to manually add variables.

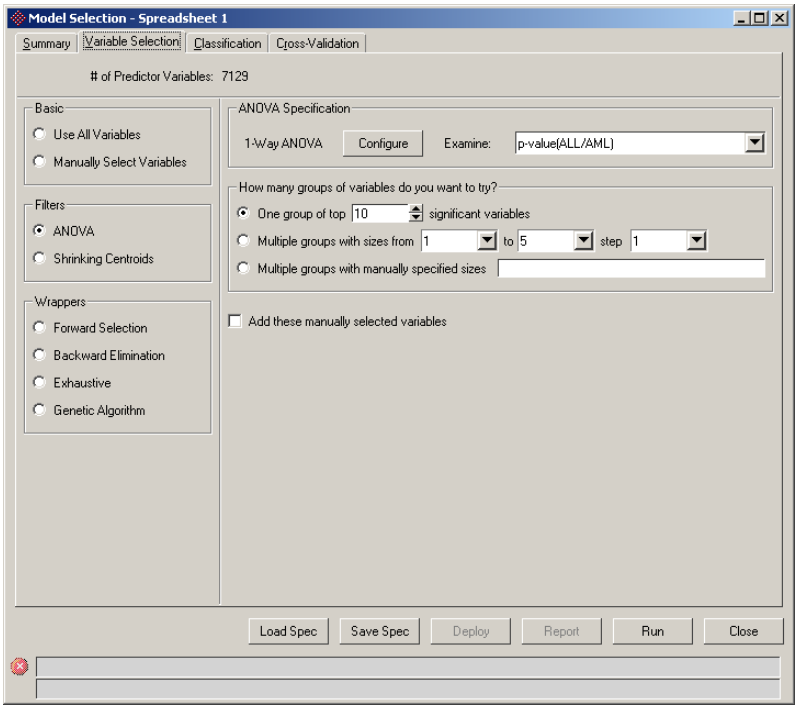


Figure 12. 11: Configuring the Variable Selection – ANOVA dialog

Shrinking Centroids

By shrinking the class centroids towards the variable overall centroid, insignificant variables can be filtered out. Hence, the *Shrinking Centroids* can be used as a variable selection method. For additional information, see Tibshirani, Hastie, Narasimham, & Chu, 2002 and Tibshirani, Hastie, Narasimham, & Chu, 2003. Figure 12. 12 shows the dialog configurations for *Shrinking Centroids* in Partek.

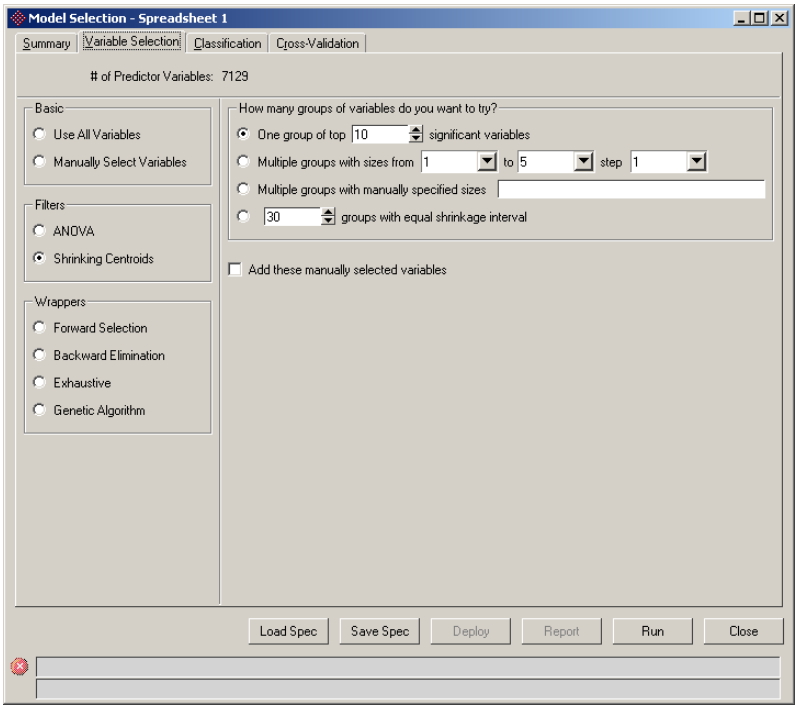


Figure 12. 12: Configuring the Variable Selection – Shrinking Centroids dialog

Specifying the Number of Variable Groups to Try

Selecting a Fixed Number of Variables, Multiple Groups with Sizes From-To-Step, and Multiple Groups with Manually Specified Sizes are the same as in the ANOVA panel.

# of Groups with Equal Shrinkage Interval

 # Groups with equal shrinkage interval allows you to use equal shrinkage to specify the variable groups. The first group uses 0 shrinkage thus will use all of the variables. The last group uses the largest shrinkage thus will keep only the most significant variable analyzed by the *Shrinking Centroids* method. During cross-validation, the numbers of variables are not fixed in the middle groups. Figure 12. 13 shows the number of groups is 30, which is the default. The minimum number is 1, which uses all of the variables.

Figure 12. 13: Configuring 30 groups with equal shrinkage interval

Wrappers

The variable selection *wrapper* approach can be divided into two components, the evaluation criteria used to assess the "goodness" of a particular subset of variables, and the search method used to select the subsets of variables to be evaluated. The evaluation component requires a classifier. The lower the score (usually the apparent error rate or posterior probability) evaluated by the classifier, the better the subset of variables. Techniques, such as discriminate analysis, can be used as the evaluation classifier, because some classifiers, such as 1-nearest neighbor,

intrinsically have no apparent error and may not be the best technique to use. Search methods include *Forward Selection, Backward Elimination, Exhaustive,* and *Genetic Algorithm.* In Partek, the evaluation classifiers are listed in the *Classification* page, which will be introduced later in the chapter.

Forward Selection

With *forward selection*, each variable is evaluated by itself and the variable, which is the best predictor of the outcome variable, is selected. The algorithm proceeds by pairing the first selected variable with each of the N-1 remaining variables. The variable that when combined with the first variable produces the best result (as determined by the evaluation criteria) is then selected as the second variable. This process is continued to some pre-determined number of variables or until all variables have been selected. If no predetermined upper bound on the number of variables is set, then forward selection will continue until it has placed all variables in the candidate set. In this case, forward selection will cause the evaluation criteria to be evaluated N(N+1)/2 times. The first variable chosen is guaranteed to be the best single variable (as measured by the evaluation criteria), but subsequent variable subsets are not guaranteed to be optimal.

Figure 12. 14 shows the settings for the forward selection *Early Stopping Criteria*. The forward selection will stop when the error rate is less than the specified value (by default, it is 0.0). The *# of variables* can be specified so that it stops when reaching the specified number. The real number of variables, which is the result of the lowest evaluation score, will not necessarily be the specified number.

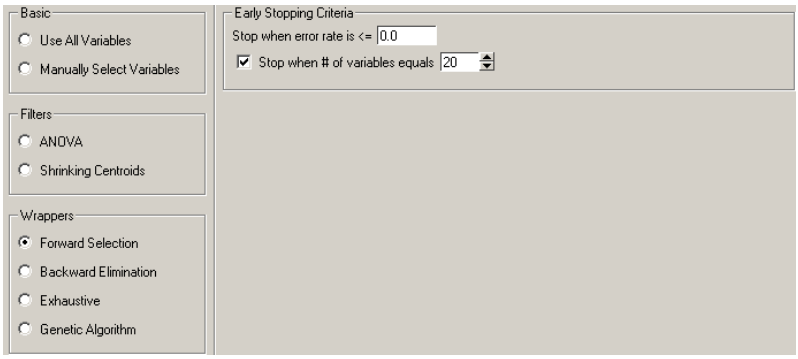


Figure 12. 14: Configuring the Forward Selection panel

Backward Elimination

Backward elimination is much like forward selection, except with backward elimination, each of the N subsets of N-1 variables is evaluated first. After this, the variable, which is not one of the N-1 variables in the subset that scored the best on the evaluation criteria, is eliminated. This process is repeated until some pre-determined number of variables is reached or until one variable is reached. Like forward selection, if you start by evaluating all variables and proceed until a single variable is found, execution requires N(N+1)/2 evaluations. Also like forward selection, the variables selected for all but the N-1 stage are not guaranteed to be the optimal choice.

Figure 12. 15 shows the setting for backward elimination *Early Stopping Criteria*. The backward elimination will stop when the error rate is less than the specified value.

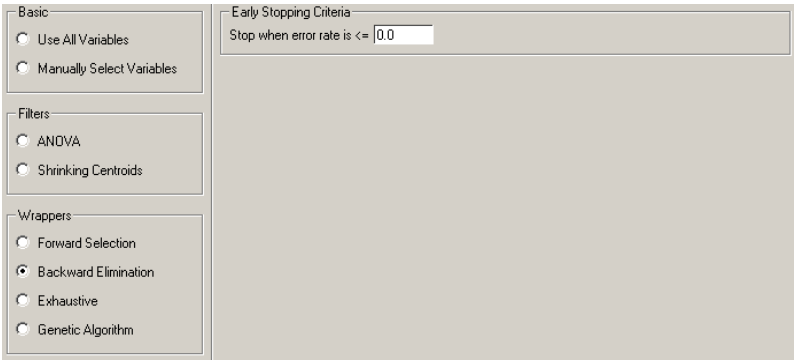


Figure 12. 15: Configuring the Backward Elimination panel

Exhaustive

It would be ideal to evaluate every possible combination of the candidate variables to determine which subset of variables is the most effective. In practice, however, there are many times when this is simply not possible on today's computers. For example, consider the modest problem where there are 60 variables. If you want to find only the best 15 variables out of the 60, there are 60/ 45/15 or 53,194,089,192,720 (53 trillion) possible unique combinations of variables to evaluate! Even on today's fastest computers, this would be a formidable task. However, there are cases when you start with a relatively small number of variables and can evaluate all possible combinations. In these cases, you absolutely should do an exhaustive search.

In Figure 12. 16, the *Searching Criteria* panel allows you to specify the *# of Variables*. The *Early Stopping Criteria* is the same as in backward elimination.

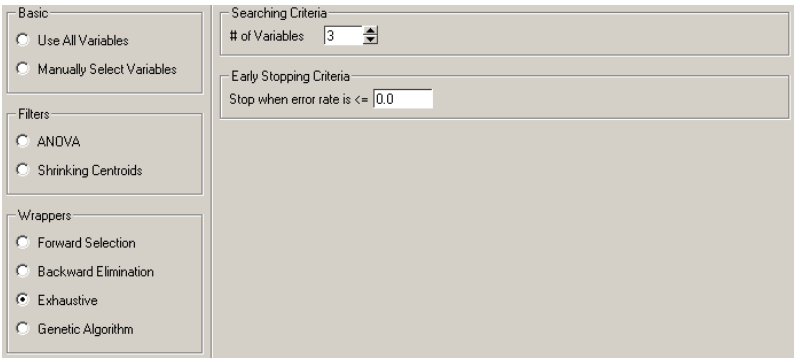


Figure 12. 16: Configuring the Exhaustive panel

Genetic Algorithms have recently seen widespread use in optimization problems and often, evaluating all possible combinations of variables cannot be afforded. In this case, the genetic algorithm provides an effective way to search the space of possible variable subsets without trying every possibility.

Figure 12. 17 shows the settings for genetic algorithm. In the *Searching Criteria* panel, you can specify the *# of Variables*, the *Population Size*, the number of *Generations*, and the Mutation *Probability*. There are two ways to specify the Population Size. The first one is to give the size directly. For large number of total variables and a relatively small population size, some variables would not appear in the initial population thereby might have less change to be selected. You can *Specify How Many Times a Variable Appears in the Initial Population* (by default, it is 1). By doing it that way, the *Population Size* will be automatically calculated.

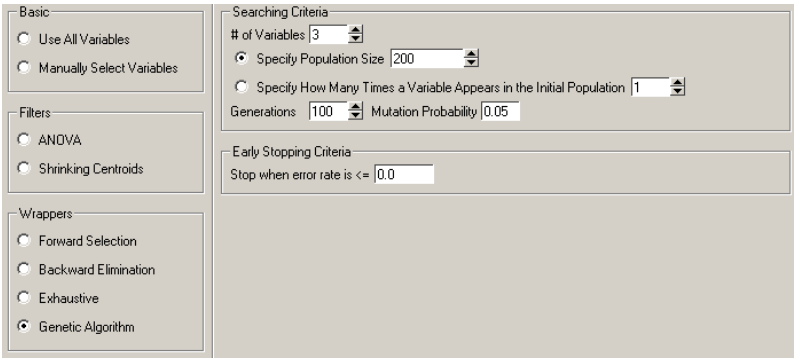


Figure 12. 17: Configuring the Genetic Algorithm panel

Classification

The *Classification* page allows the configuration of the *Variable to Predict*. Use the *Variable to Predict* combo box to choose the variable that will be used to do the prediction. All categorical variables are listed here. By default, the class variable is selected. Use the left mouse button to choose multiple classification methods from the list shown in the left panel of Figure 12. 18. Use the right mouse button to navigate among those classification methods without enabling or disabling the classification methods. Depending on the classification method chosen, the large panel on the right will allow you to configure more parameters. In Figure 12. 18, *K-Nearest Neighbor* is selected.

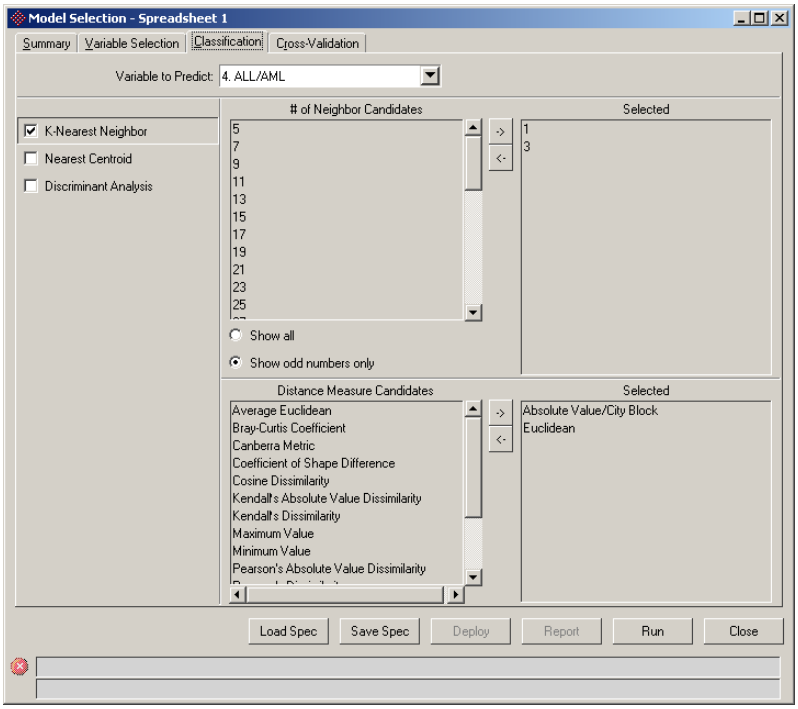


Figure 12. 18: Configuring the K-Nearest Neighbor dialog

K-Nearest Neighbor

# of Neighbor Candidates

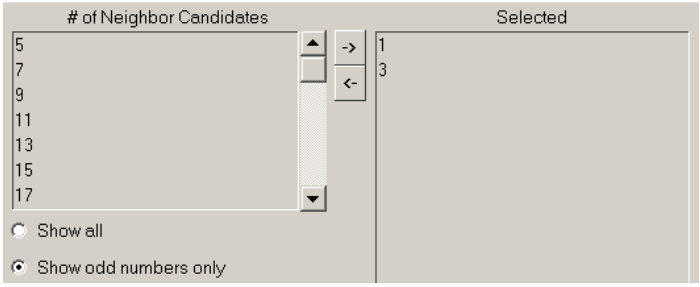


Figure 12. 19: Configuring the Number of Neighbors panels

Use the *# of Neighbor Candidates* to choose the number of neighbors (K). You can try multiple numbers here. If the number of samples is less than K, all samples will be used in the classification. The operation of the list boxes is the same as the operation in the *Manually Select Variables* dialog box. By default, to avoid ties in a two-class situation, the odd numbers of neighbors are shown, but it can be turned off by selecting the *Show all* radio button.

Distance Measure Candidates

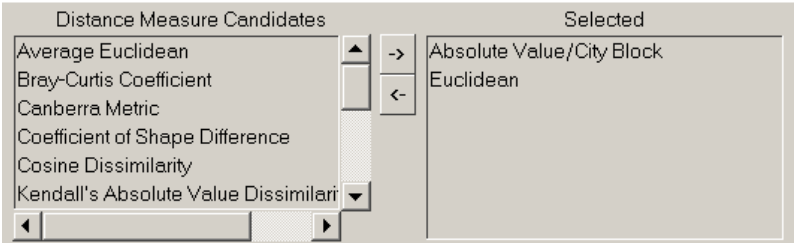


Figure 12. 20: Configuring the Distance Measure Candidates panels

The *Distance Measure Candidates* allows you to choose multiple distance measures that will be used by the K-Nearest Neighbor classifier.

Tip: To determine the number of models with K-Nearest Neighbor classification method do the following:
# of models = # of groups in the variable selection * # of selected Ks * # of selected distance measures

In Figure 12. 11, there is one group in the variable selection page (top 10 significant variables). In Figure 12. 18, there are two values of K selected (1 and 3). 2 *Distance Measures* are selected (*Absolute Value/City Block* and *Euclidean*), so the model space defined by Figure 12. 11 and Figure 12. 18 has 1x2x2 = 4 models.

Nearest Centroid

For information on the Nearest Centroid classification method, see the reference [4]. Figure 12. 21 shows the prior probabilities configuration for *Nearest Centroid.* Here you can select *Equal*, *Proportional*, or *Specified Prior Probabilities*. To specify prior probabilities, list all the categories followed by the prior probabilities. Category names are case sensitive and should be quoted if there are 2 or more words. An example of specifying prior probabilities is as follows: Infected 0.3 "Not Infected" 0.7

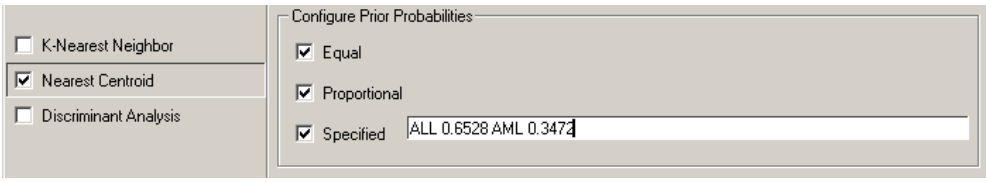


Figure 12. 21: Configuring the Nearest Centroid panel

Tip: To determine the number of models with the Nearest Centroid classification method do the following:
# of models = # of groups in the variable selection * # of selected prior probabilities

In Figure 12. 12, there are 30 groups in the variable selection page. In Figure 12. 21, three types of prior probabilities are selected so the model space defined by Figure 12. 12 and Figure 12. 21 has 30x3 = 90 models.

Discriminant Analysis

The *Discriminant Analysis* method can do predictions based on the class variable. When it is enabled, the *Variable to Predict* box will be disabled and the class variable will automatically be selected (Figure 12. 22).

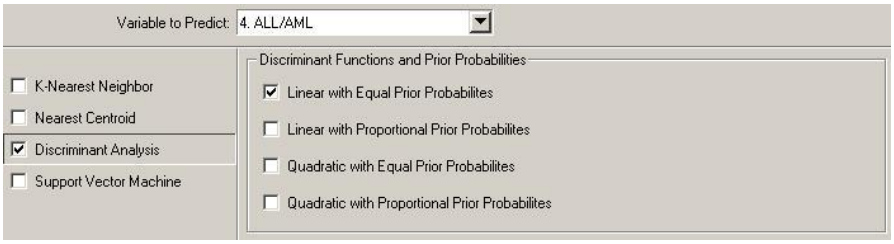


Figure 12. 22: Configuring the Predict on Class Variable with Discriminant Analysis panel

Figure 12. 22 also shows the settings for the *Discriminant Analysis* classification method. Here *Linear* and/or *Quadratic Discriminant Functions* can be selected, as well as *Equal* and/or *Proportional Prior Probabilities.*

Tip: To determine the number of models with Discriminant Analysis classification method do the following:
# of models = # of groups in the variable selection * # of selected discriminant functions * # of selected prior probabilities

In Figure 12. 17, there is 1 group of variables (the genetic algorithm will search for the best 3 variables). In Figure 12. 22, two discriminant functions and 2 types of prior probabilities are selected; therefore, the model space defined by Figure 12. 17 and Figure 12. 22 has 1x2x2 = 4 models.

Support Vector Machine

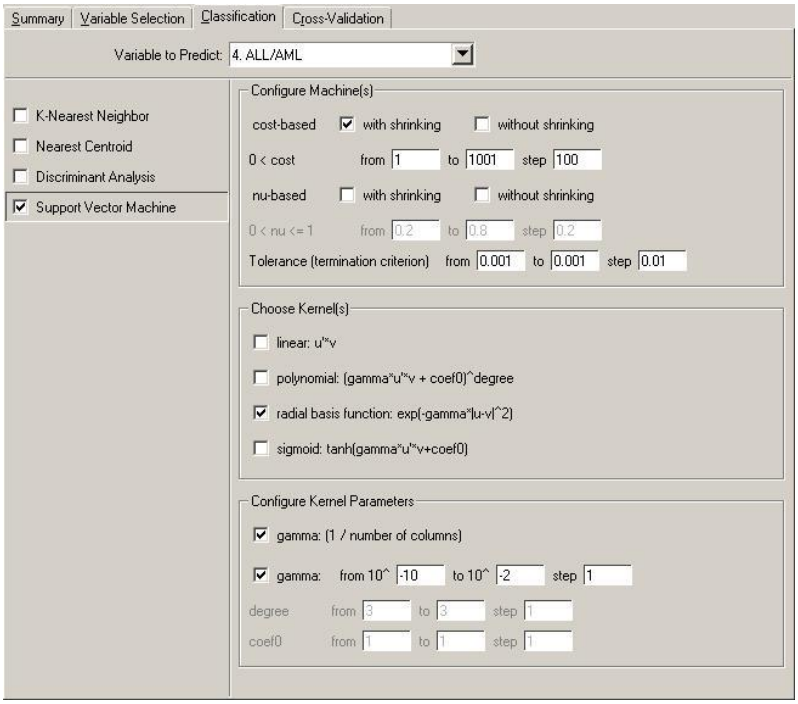


Figure 12. 23: Configuring the SVM panel

To run model selection with SVM you must choose at least one machine (cost or nu based, with or without shrinking) and at least one kernel (linear, polynomial, radial basis function, or sigmoid). The kernel parameters that will be used depend on the kernels that are checked.

The cost and nu parameters adjust the balance between over and under fitting the model to the data. The cost parameter must be greater than zero but has no upper bound. The nu parameter must be greater than zero and less than or equal to one.

The gamma parameter is stepped through as an exponent.

Partial Least Squares

The *Partial Least Squares* method can do predictions by doing PLS first as components selection and doing Discriminant Analysis second as classification method (Figure 12. 224).

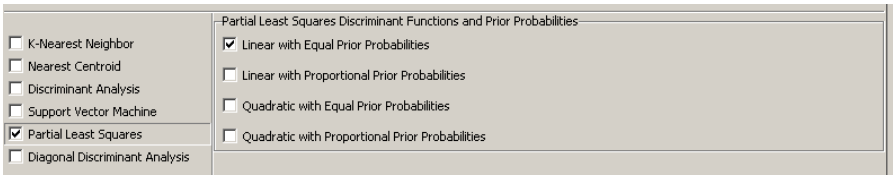


Figure 12. 244: Configuring the PLS panel

Figure 12. 224 shows the settings for the *Partial Least Squares Discriminant Analysis* (PLS-DA) classification method. Here *Linear* and/or *Quadratic*

Discriminant Functions can be selected, as well as *Equal* and/or *Proportional Prior Probabilities.*

Diagonal Discriminant Analysis

Figure 12. 25 shows the settings for the *Diagonal Discriminant Analysis* classification method. Here *Linear* and/or *Quadratic Discriminant Functions* can be selected, as well as *Equal* and/or *Proportional Prior Probabilities.*

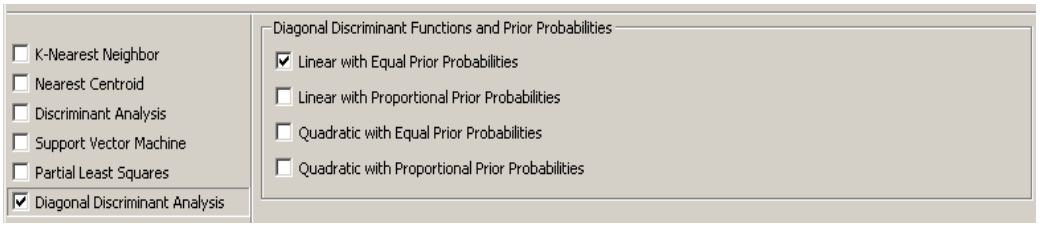


Figure 12. 255: Configuring the Diagonal Discriminant Analysis panel

Random Forest

The Random Forest method uses votes from multiple random trees' classifications to get the final classification. Figure 12. 26 shows the settings for the *Random Forest* classification method. *# of trees* and *Bootstrap Seed* need to be configured for the Random Forest model. In each tree, *# of Random Candidates* and *# of Linear Combinatins* should be specified to compose the tree. More than one *# of Random Candidates* and *# of Linear Combinations* can be selected to generate multiple models.

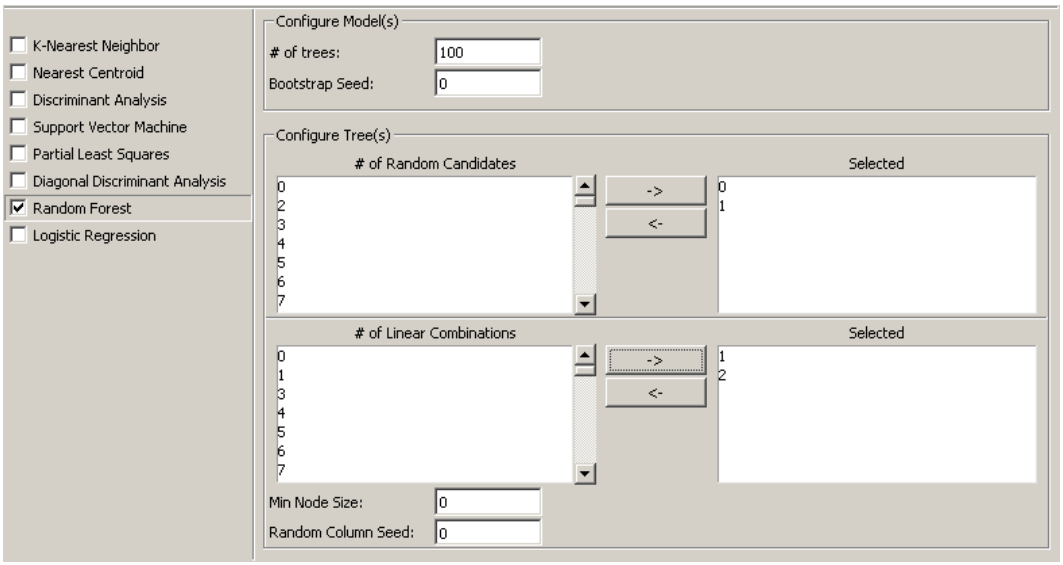


Figure 12. 266: Configuring the Random Forest panel

Logistic Regression

Figure 12. 27 shows the settings for the *Logistic Regression* classification method. There is no configuration needed for Logistic Regression method. The underlying logistic regression model is:

$$f(z) = \frac{e^z}{e^z + 1}$$

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k$$

Where $x_1,\ x_2,\ x_3, \ldots,\ x_k$ are the variables selected from Variable Selection step

Figure 12. 277: Configuring the Logistic Regression panel

SVM Regression

SVM Regression is Support Vector Machine Regression which is used to predict on numeric variables eg. age, height, wight and so on. Figure 12. 28 shows the settings for the SVM Regression method. It is almost the same as Support Vector Machine configuration except that the machine types can be selected for SVM Regression are *epsilon_svr* and *nu_svr*, but the machine types for Support Vector Machine are *cost* and *nu.*

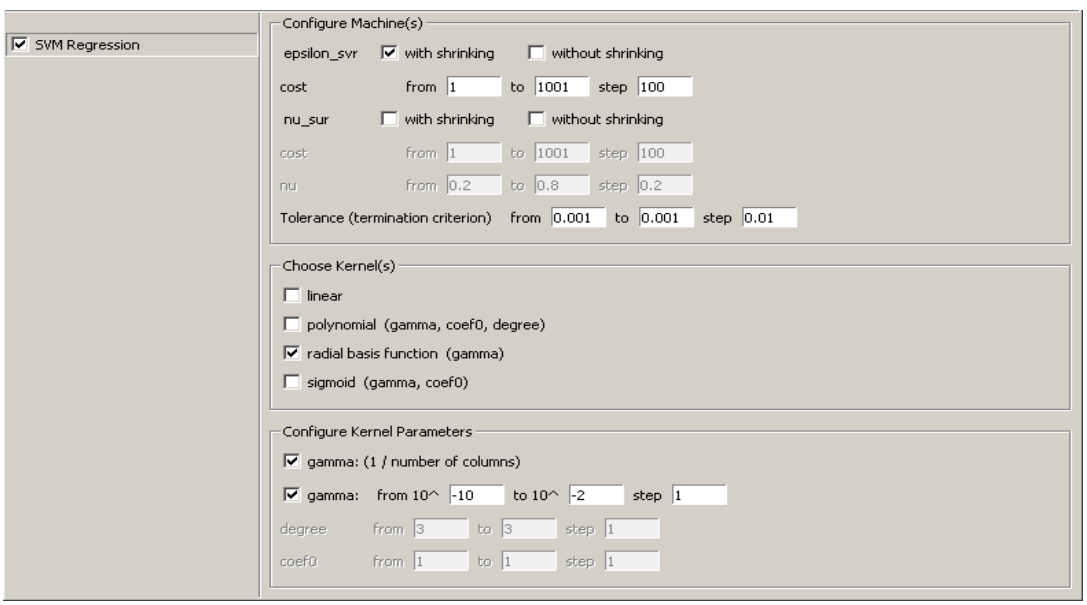


Figure 12. 288: Configuring the SVM Regression panel

Cross-Validation

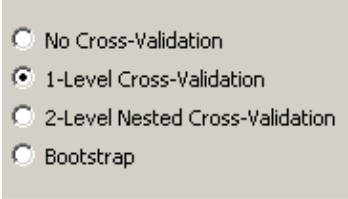


Figure 12. 299: Configuring the Cross-Validation options

The *Cross-Validation* tab allows you to choose the cross-validation method. The choices are *No Cross-Validation, 1-Level Cross-Validation,* and *2-Level Cross-Validation.* In the example below, *No Cross-Validation* was selected and can be used to do the re-substituted test that uses the training samples to produce what is called the "apparent error rate". *1-Level* and *2 –Level Nested Cross-Validation* will be explained in detail below.

Presentation Order

Figure 12. 30: Configuring the Random Seed panel

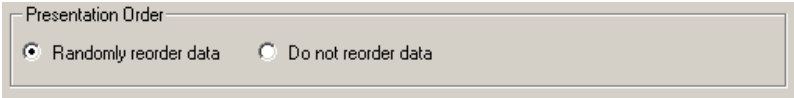


Figure 12. 31: Configuring the Presentation Order panel

Do Not Reorder Data

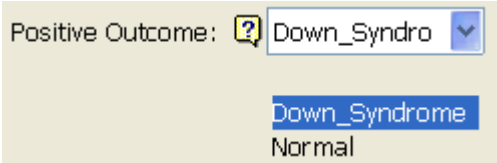
Choosing **Do Not Reorder Data** will not randomize the data's original presentation order when partitioning the data for cross-validation.

Randomly Reorder Data

Choosing **Randomly Reorder Data** will randomize the original data prior to partitioning for cross-validation. The random seed can be specified as in Figure 12. 30.

Positive Outcome

Specifies which category is considered as positive. For example, in a Disease/Normal study, specify Disease as the **Positive Outcome**. This field is needed when using **Matthews Correlation Coefficient** or **Area Under Curve** as the **Selection Criterion**.



Model Selection Criterion

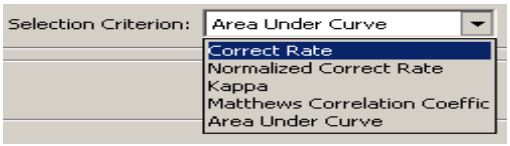


Figure 12.32: Configure the Selection Criterion to Predict on Categorical Variable

To predict on Categorical Variable, the Model Selection tool uses confusion matrix based scores to evaluate model performance. Those scores include *Correct Rate, Normalized Correct Rate, Kappa, Matthews Correlation Coefficient* and *Area Under Curve.* Although some of them can be calculated from any size of confusion matrix, here the 2x2 matrix is used as an example:

	Actual positive	**Actual negative**
Predicted positive	TP	FP
Predicted negative	FN	TN

TP,FP, FN, and TN are counts. N = TP+FP+FN+TN.

Score	**Calculation**
Correct Rate	(TP+TN)/N
Normalized Correct Rate	(TP/(TP+FP) + TN/(FN+TN)) / 2
Kappa	$\frac{(TP+TN)-(((TP+FN)(TP+FP)+(FP+TN)(FN+TN))/N)}{N-(((TP+FN)(TP+FP)+(FP+TN)(FN+TN))/N)}$
Matthews Correlation Coefficient	$\frac{TP\times TN-FP\times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$
Area Under Curve	$(\frac{TP}{TP+FN}+\frac{TN}{FP+TN})/2$
Prevalence	(TP+FN)/N

Comparison of these scores:

Score:	**Dealing with unbalanced data:**	**Ease of Explanation:**	**Note:**
Correct Rate	Poor	Good	Doesn't use all the confusion matrix information, sensitive to prevalence
Normalized Correct Rate	Fair	Fair	
Kappa	Poor – Fair	Poor	Measures agreement while correcting for chance classification. Kappa can range between -1 and 1, with values close to 1 reflecting highest

			agreement. 0 means no improvement over chance. A negative Kappa means that there is less agreement than would be expected by chance. The scale for accessing the kappa agreement: Kappa < 0.4 poor, 0.4 < Kappa < 0.75 good, and Kappa >0.75 excellent. Fails when the size of one class far exceeds the other. See reference Cohen, Jacob (1960) for more information on Kappa.
Matthews Correlation Coefficient	Poor – Fair	Poor	Measures the quality of classification. MCC can range between -1 and 1, with value of 1 reflecting a perfect prediction, 0 meaning an average random prediction and -1 representing an inverse prediction. Note: **Positive Outcome** needs to be specified to use **Matthews Correlation Coefficient** as the **Selection Criterion**
Area Under Curve	Fair	Fair	Measures the ability to distinguish 2 classes. The range is between 0 and 1. 1 means the false positive rate is 0 and true positive rate is 1 (a perfect classifier); 0.5 means the classifier couldn't distinguish 2 classes, the performance is not better than flipping a fair coin; < 0.5 means an inverse prediction. Note: **Positive Outcome** needs to be specified to use **Area Under Curve** as the **Selection Criterion**

To predict on Numeric Variable, the Model Selection tool uses Mean Square Error to evaluate model performance.

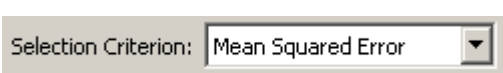


Figure 12.33: Configure the Selection Criterion to Predict on Numeric Variable

Score	Calculation
Mean Square Error	$\frac{\sum(\hat{X}-X)^2}{N}$, where $\hat{X}$ is the predicted values and X is the real values

Correct Rate, *Normalized Correct Rate*, *Kappa*, *Matthews Correlation Coefficient*, *Area Under Curve* and *Mean Square Error* will be reported in the cross-validation results (see examples in **Bootstrap**
Figure 12. 39 shows the dialog configurations for *Bootstrap*.

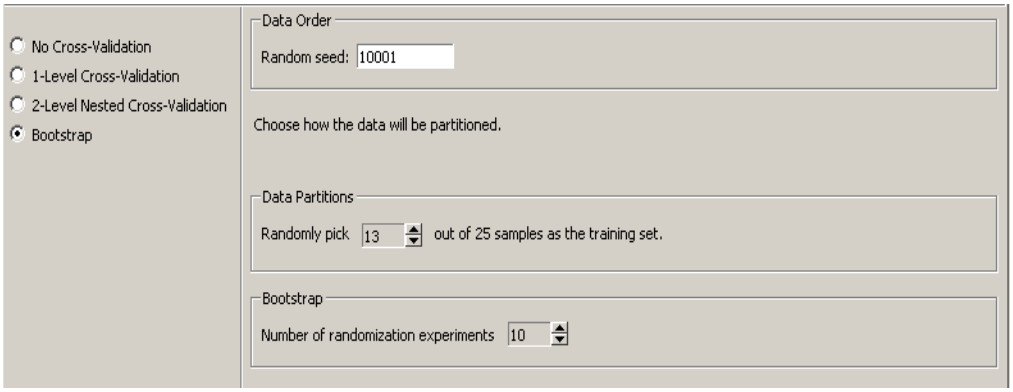


Figure 12. 39: Configuring the Bootstrap dialog

In.the *Data Order* panel (Figure 12. 40), *Random seed* is set to randomly pick the specified number of samples from the original data set as the training set.

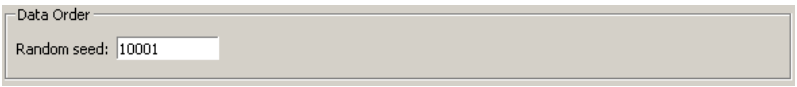


Figure 12. 40: Set the Bootstrap Random Seed

In the *Data Partitions* panel (Figure 12. 3041), you can specify the size of training set.

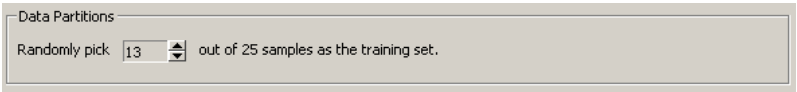


Figure 12. 41: Specify the Size of Training Set.

In the *Bootstrap* panel (Figure 12. 42), you can specify how many runs of randomization experiments by changing the number in the spinbox. In each run. Partek Model Selection will check whether this experiment includes all of the categories of the predicted variable. If not, it will keep generating a randomization experiment until it meets the requirement.

Figure 12. 42: Configure Bootstrap Method.
Running the Model Selection section below). Most importantly, in a 2-Level cross-validation, models will be selected based on one of these scores (selectable in the Model Selection Criterion [Figure 12.32] combo box) after each round of inner cross-validation.

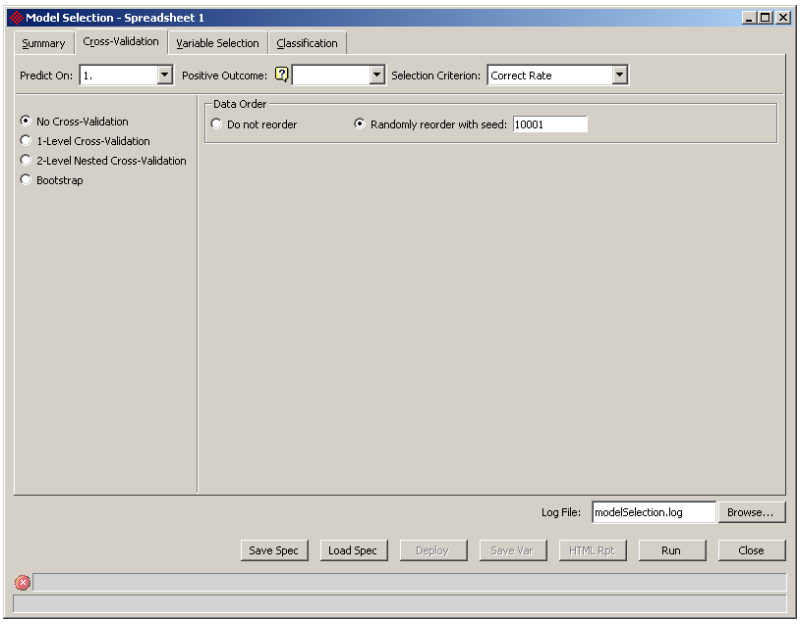


Figure 12.34: Configuring the Cross-Validation tab

1-Level Cross-Validation

Figure 12.35 shows the dialog configurations for *1-Level Cross-Validation*.

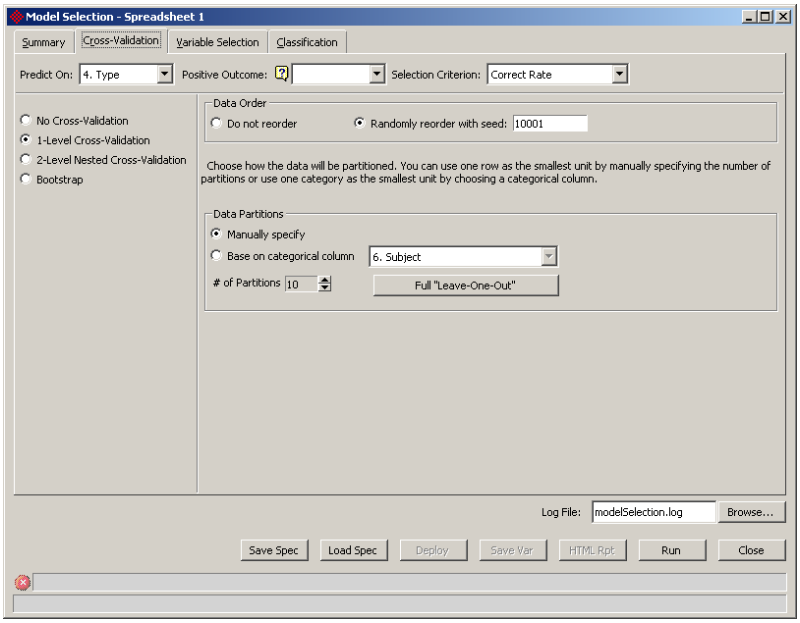


Figure 12.35: Configuring the 1-Level: Single Cross-Validation dialog

In the *Data Partitions* panel (Figure 12. 306), you can choose *Manually*, or *Base on a categorical column*, which allows you to use a categorical column to specify data partitions. To choose the *Manually specify* option, click the radio button and specify the number of data partitions (folds) in the *# of Partitions* box. The minimum number is **2**. The default number is **10**. The *Full "Leave-One-Out"* button is

enabled when *manually specify* is selected. It allows you to do the largest number of cross-validations (# of partitions equals the number of samples).

To choose the *Base on categorical column* option, click the radio button, and choose a column from the drop-down list. Samples from the same category will stay together in a partition. The *Full 'Leave-One-Category-Out'* button is enabled when *Base on categorical column* is selected.

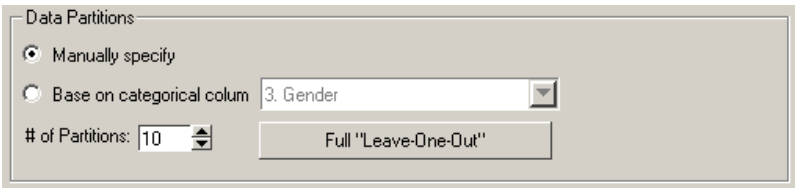


Figure 12. 306: Configuring the 1-Level Cross-Validation Manual Configuration Data Partitions panel

2-Level Cross-Validation

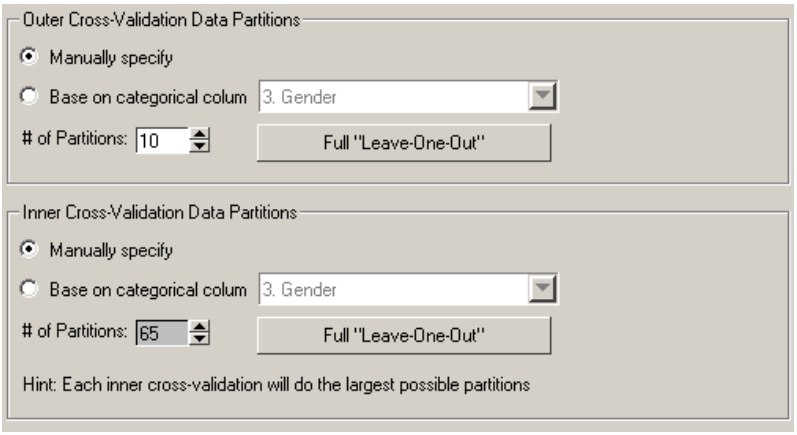


Figure 12. 317: Configuring the Nested Cross-Validation panel

Figure 12. 317 shows the configurations of the 2-Level Nested Cross-Validation panel. They are similar to the *1-Level Cross-Validation*. Here you can specify 2 nested levels of cross-validation. The partitioning methods of the inner and outer cross-validations do not have to be the same. If the number of partitions specified is larger than the number of samples, full leave-one-out partitioning will be performed. Figure 12. 328 shows there will be 10-fold for the outer cross-validation and 10-fold for the inner cross-validation.

Note: For 2-Level Cross-Validation some outer '*# of Partitions'* and inner '*# of Partitions*' combinations are not valid. For example, you have 4 male samples and 6 female samples. You want to have a 2-fold outer cross-validation based on gender and specify a 6-fold inner cross-validation. The invalid situation will come up when the 6 female samples are held out in the outer cross-validation while the inner cross-validation wants to perform 6-fold partitions with 4 male samples. In such cases, the *Partek Classification Model Selection* will use the best-effort strategy. Namely, if there are 5 samples, it will do a 5-fold cross-validation, if there are 4 samples it will

do a 4-fold cross-validation etc. Also in these cases the inner '*# of Partitions*' entry box is grayed out. That indicates the number could not be exactly performed during all passes of cross-validation.

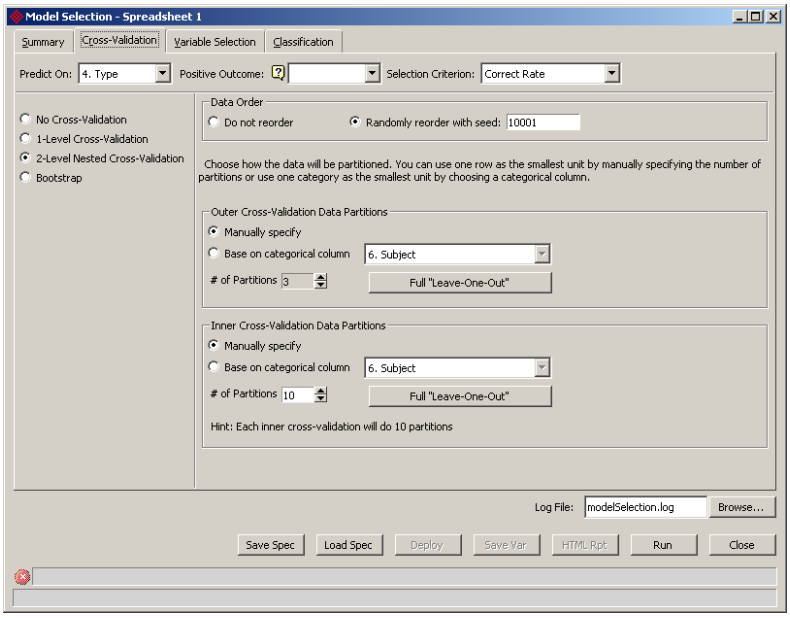


Figure 12. 328: Configuring the 2-Level Nested Cross-Validation dialog

Bootstrap

Figure 12. 39 shows the dialog configurations for *Bootstrap*.

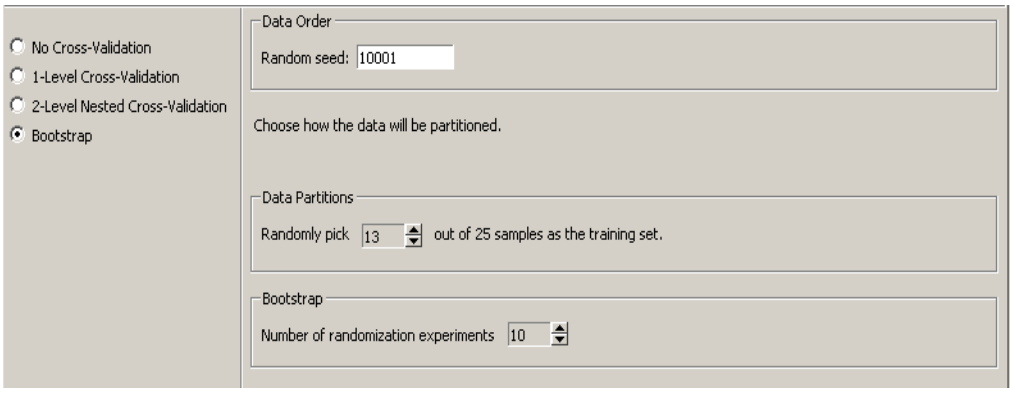


Figure 12. 39: Configuring the Bootstrap dialog

In.the *Data Order* panel (Figure 12. 40), *Random seed* is set to randomly pick the specified number of samples from the original data set as the training set.

Figure 12. 40: Set the Bootstrap Random Seed

In the *Data Partitions* panel (Figure 12. 3041), you can specify the size of training set.

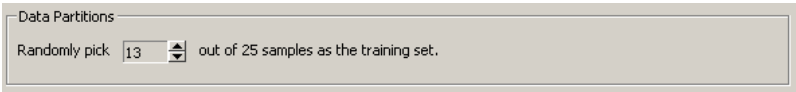


Figure 12. 41: Specify the Size of Training Set.

In the *Bootstrap* panel (Figure 12. 42), you can specify how many runs of randomization experiments by changing the number in the spinbox. In each run. Partek Model Selection will check whether this experiment includes all of the categories of the predicted variable. If not, it will keep generating a randomization experiment until it meets the requirement.

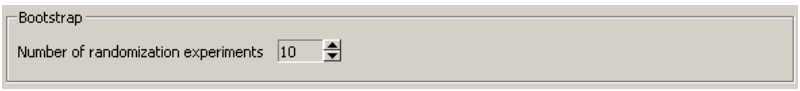


Figure 12. 42: Configure Bootstrap Method.

Running the Model Selection

Click the **Run** button to start running the computations. The summary page will be brought to the front (Figure 12.43). After clicking **Run** to start running the test, the *Deploy*, *Report*, and *Run* buttons will be disabled. The live report list box will show all the models and their current correct rates. If there are multiple models, click on the column header to sort increasingly or decreasingly by any particular field.

Clicking  will stop the test. The progress bar shows the percentage of the computations completed.

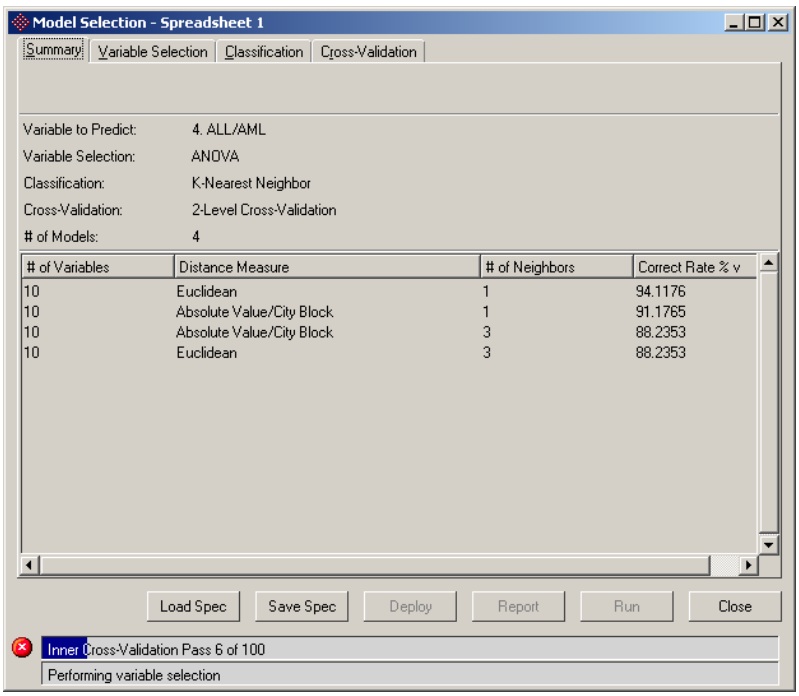


Figure 12.43: Running a Test in the Summary tab

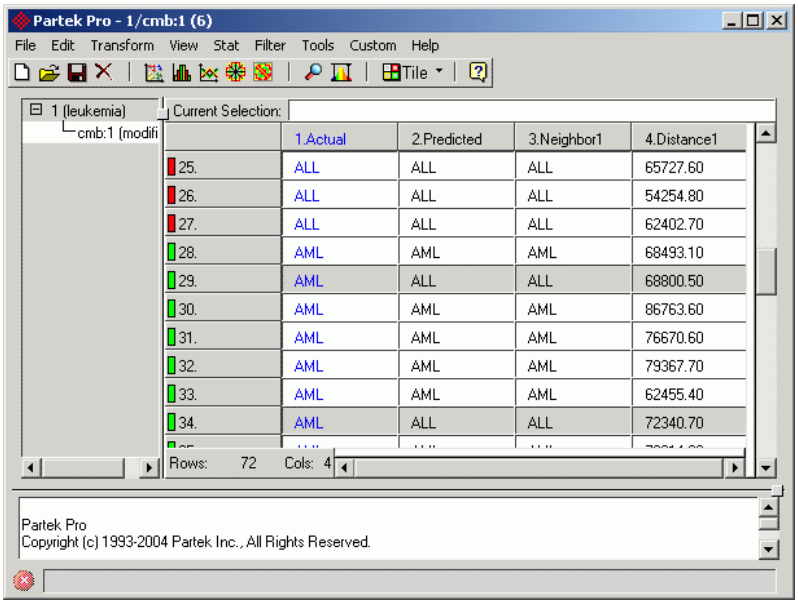
Reports

A report is available after processing is finished. Depending on the configuration, you will get a different spreadsheet and/or an HTML report. The rest of the reports are all from K-Nearest Neighbor analysis. Nearest Centroid reports are similar.

A Single Model with 1-level Cross-Validation

Spreadsheet Report

A single model with 1-level cross-validation will give a spreadsheet report. Figure 12.44 shows the spreadsheet that contains the results of a single K-Nearest Neighbor model with 1-level cross-validation. Column 1 shows the actual class, column 2 shows the predicted class, and columns 3 and 4 show the prediction and the distance from the 1-nearest neighbor. Depending on the number of neighbors evaluated, there could be more columns like column 3 and 4. The rows in the spreadsheet that are highlighted indicate those samples were misclassified during the cross-validation.



	1.Actual	2.Predicted	3.Neighbor1	4.Distance1
25.	ALL	ALL	ALL	65727.60
26.	ALL	ALL	ALL	54254.80
27.	ALL	ALL	ALL	62402.70
28.	AML	AML	AML	68493.10
29.	AML	ALL	ALL	68800.50
30.	AML	AML	AML	86763.60
31.	AML	AML	AML	76670.60
32.	AML	AML	AML	79367.70
33.	AML	AML	AML	62455.40
34.	AML	ALL	ALL	72340.70

Figure 12.44: Viewing the spreadsheet report on Single K-Nearest Neighbor model with 1-Level Cross-Validation

HTML Report

To get the HTML report click **Report** button. An HTML Report will include:

- Date
- Time
- Title
- Comments
- Experiment Summary
- Variable Selection Parameters
- Classification Parameters

- Confusion Matrix
- Classification Summary
- Table of Selected Variables
- Frequency Table of Selected Variables

An example of an HTML report is shown below.

Pattek Inc.
www.pattek.com

February 13, 2006
01:20:10 PM

K-Nearest Neighbor Analysis Test Summary for "leukemia"

Report invoked from model selection

Summary

Variable to Predict	4. ALL/AML
# of Predictor Candidates	7129
# of Samples	72
# of Models	1
Random Seed	10001
Presentation Order	Randomly Reorder Data
Model Selection Criterion	Correct Rate
Cross-Validation	1-level
Partitions	10

Variable Selection Parameters

Variable Selection Method	ANOVA
Examine	1-Way ANOVA p-value(ALL/AML)
How many groups of variables	One group of top 10 significant variables

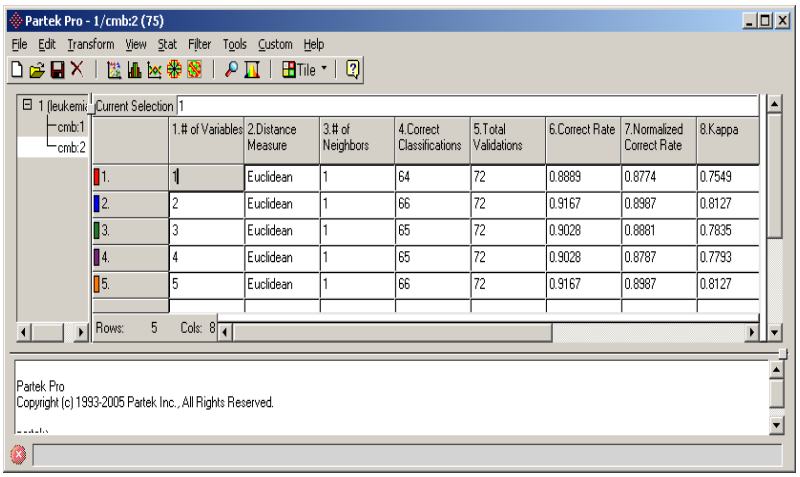
Classification Parameters

Figure 12.45: Viewing the HTML report on a single K-Nearest Neighbor model with one-level cross-validation

Multiple Models with 1-Level Cross-Validation

Spreadsheet Report

Figure 12.46 shows the report on Multiple K-Nearest Neighbor models with single level cross-validation. Each row corresponds to the report of one model with parameters (*# of Variables*, *Distance Measure*, and *# of Neighbors)* in columns 1, 2, and 3, respectively. Column 4 shows the number of correct classifications of that model during the cross-validation. Column 5 shows how many test set validations were performed in the cross-validation. Column 6 is simply the result of column 4 divided by column 5. Column 7 shows the *Normalized Correct Rate*, which averages the correct rates of each category. Column 8 shows the *Kappa* value. Column 6, 7, and 8 can be used as evidences to pick a model and deploy it; however, those values are biased and should not be used as the estimate of prediction accuracy. When evaluating multiple models, the unbiased estimate of prediction accuracy can only be obtained from a two-level nested cross-validation.



	1.# of Variables	2.Distance Measure	3.# of Neighbors	4.Correct Classifications	5.Total Validations	6.Correct Rate	7.Normalized Correct Rate	8.Kappa
1.	1	Euclidean	1	64	72	0.8889	0.8774	0.7549
2.	2	Euclidean	1	66	72	0.9167	0.8987	0.8127
3.	3	Euclidean	1	65	72	0.9028	0.8881	0.7835
4.	4	Euclidean	1	65	72	0.9028	0.8787	0.7793
5.	5	Euclidean	1	66	72	0.9167	0.8987	0.8127

Figure 12.46: Viewing the Multiple K-Nearest Neighbor models with 1-Level Cross-Validation

HTML Report

Fields in the HTML report on Multiple K-Nearest Neighbor models with simple single-level cross-validation are:

- Date
- Time
- Title
- Comments
- Experiment summary
- Variable Selection Parameters
- Classification Parameters
- Model Overall Scores
- Confusion Matrix for Individual Model
- Classification Summary for Individual Model

An example of an HTML report is shown below.

Partek Inc.
www.partek.com

September 24 2007
04:45:29 PM

K-Nearest Neighbor Analysis Test Summary for "leukemia"

Report invoked from model selection

Summary

Variable to Predict	3. ALL/AML
# of Predictor Candidates	7129
# of Samples	72
# of Models	2
Random Seed	10001
Presentation Order	Randomly reorder data
Model Selection Criterion	Normalized Correct Rate
Cross-Validation	1-level
Partitions	2

Variable Selection Parameters

Variable Selection Method	ANOVA
Examine	1-Way ANOVA p-value(ALL/AML)
How many groups of variables	2 groups with manually specified sizes 2 4

Classification Parameters

Figure 12.47: Viewing the Multiple Models 1-Level Cross-Validation HTML report

Multiple Models in a Nested Two-Level Cross-Validation

Inner Cross-Validation Spreadsheet Report

As shown in Figure 12.48, this report is similar to the report in Figure 12.46. Column 1, 2, and 3 identify the model parameters. Column 4 shows the number of correct classifications during the 2-level cross-validation. Column 5 shows how many validations are performed in the 2-level cross-validation. Column 6 is simply the result of column 4 divided by column 5. Column 7 shows the *Normalized Correct Rate*, which averages the correct rates of each category. Column 8 shows the *Kappa* value.

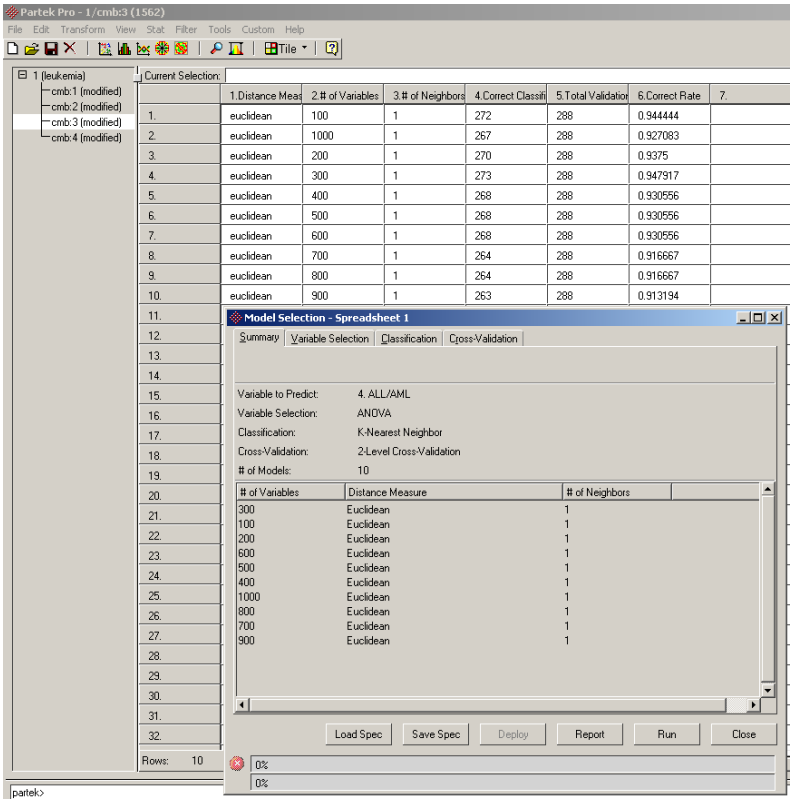


Figure 12.48: Viewing the multiple models 2-Level Cross Validation - all Models Performance report

Outer Cross-Validation Spreadsheet Report

For each pass of the outer cross-validation, there will be one partition of samples held out. All the models will perform an inner cross-validation (single level) upon the remaining samples. After the inner cross-validation, the best model or the tied best models will be tested on the held out samples to get the accuracy rate for that pass of the outer cross-validation. This process will be performed for all the outer cross-validation passes.

Figure 12.49 shows the result of the accuracy rate for each pass. Column 1 shows the pass number. In cases of tied models, there will be multiple rows in that pass. Column 2 shows how many samples were held out during that outer cross-validation pass. Column 3 shows the best validation score in the inner cross-validation. Column 4 shows how many models were tied best models. Column 5, 6, and 7 identify the model. Column 9 shows the number of correct classifications of that model when it is applied to the held-out samples. Column 10 is the same as Column 2. It is the number of samples that were held out in the outer cross-validation. Column 10 is simply the result of column 8 divided by column 9. Column 11 shows the *Normalized Correct Rate*, which averages the correct rates of each category. Column 12 shows the Kappa value.

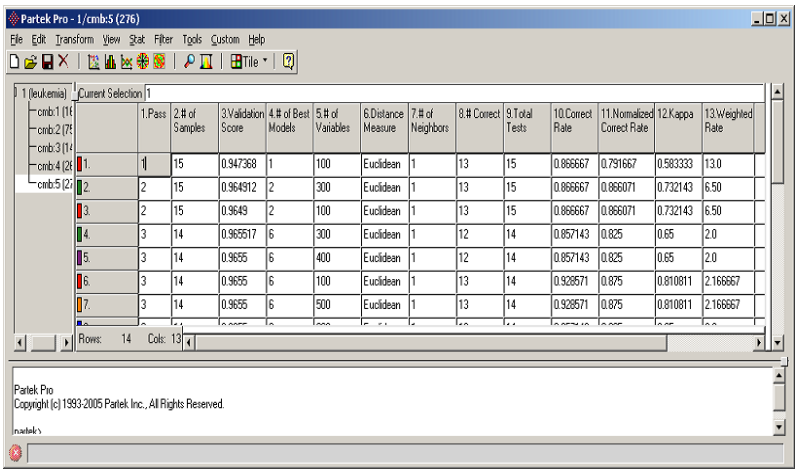


Figure 12.49: Viewing the Multiple Models 2-Level Cross Validation Selected Models report

HTML report on the performance of selected models in each pass of the Outer Cross-Validation

Fields in the HTML report are:

- Date
- Time
- Title
- Comments
- Experiment summary
- Variable Selection Parameters
- Classification Parameters
- Weighted Confusion Matrix during Outer Cross-validation
- Weighted Classification Summary during Outer Cross-validation
- Estimated Scores during Outer Cross-validation

Depending on the Model Selection Criterion, *Confidence Interval*, *Confidence Interval (Normalized)*, or *Confidence Interval (Area Under Curve)* can be found in the *Summary* of the HTML report (Figure 12.7). *Confidence Interval* is calculated by following 3 steps:

1. ARCSINE Transformation

   $p' = \arcsin\sqrt{p}$,where p is the *Correct Rate*, *Normalized Correct Rate* or *Area Under Curve*. See reference Zar, J.H. (1999) for more information on ARCSINE transformation.

2. 95% Confidence Interval

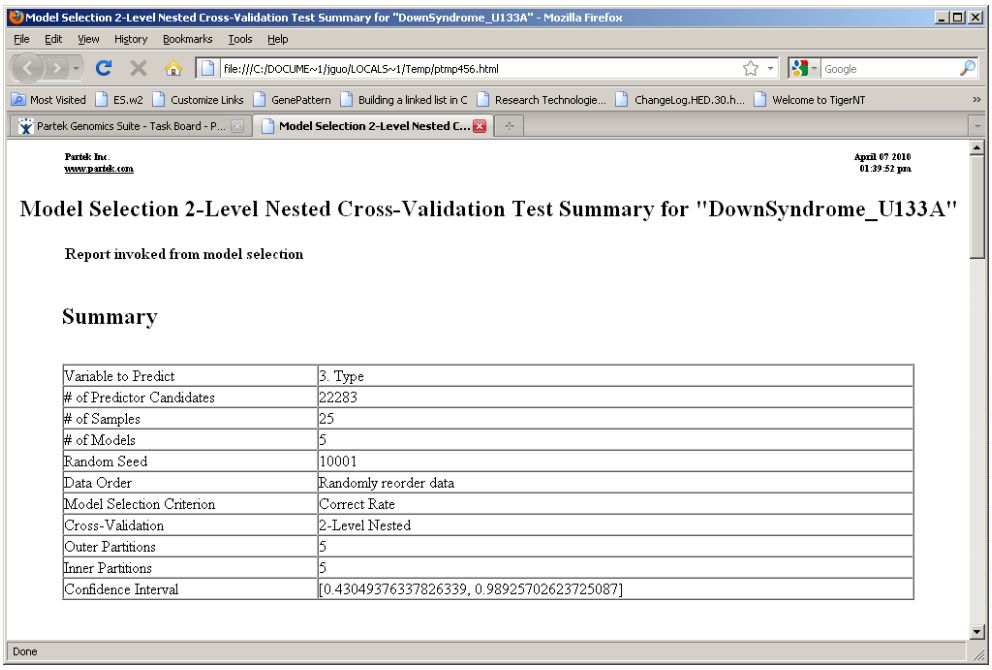
   $\bar{p}' \pm (t_{0.05(2),totalpass-1})S'_{\bar{x}}$, where $\bar{p}'$ is the mean, $t_{0.05(2),totalpass-1}$ is the two-tail t distribution critical value with degree freedom equal to total passes -1 and $S'_{\bar{x}}$ is the standard error

3. Transform backwards

   $p = (\sin p')^2$, where p' is the confidence interval from step 2.

Note: *Correct Rate*, *Normalized Correct Rate* and *Area Under Curve* are all proportions which should range between 0 and 1.0. After being transformed backwards, the confidence interval might go beyond or below the range. In this case, confidence interval will be set to 1.0 if it is beyond the range and set to 0 if it is below the range.

Note: The "Weighted Confusion Matrix of Selected Best Models during Outer Cross-validation" is adjusted by the number of tied best models in the inner cross-validation. The confusion matrix may have non-integer values, but the sum will be equal to the total number of samples.

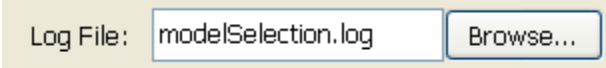


Variable to Predict	3. Type
# of Predictor Candidates	22283
# of Samples	25
# of Models	5
Random Seed	10001
Data Order	Randomly reorder data
Model Selection Criterion	Correct Rate
Cross-Validation	2-Level Nested
Outer Partitions	5
Inner Partitions	5
Confidence Interval	[0.43049376337826339, 0.98925702623725087]

Figure 12.50: Viewing the Multiple Models 2-Level Cross Validation HTML report

Log File Text Report

At the bottom right of the **Model Selection** dialog, there is a **Log File** field that allows users to log model selection steps and results.



Click **Browse…** to give a new name to or select an existing file for the log file. The **Model Selection** tool will append new log to an existing file. The log file will contain the following information:

- Input File
- Variable to Predict
- # of Predictor Candidates
- # of Samples
- # of Models
- Random Seed
- Data Order

- Model Selection Criterion
- Cross-Validation Partitions
- Variable Selection Method
- Variable Selection Examine
- How many groups of variables
- Classification Method
- Classification Parameters
- Experiment Started At
- Experiment Ran By User
- Positive Outcome
- Experiment Stopped At

During the Cross-Validation, the log file will record the following information:

- Selected Variable List
- Cross-Validation Pass
- Variable Selection Method
- Variable Selection Parameter
- Variable Selection Criteria
- Classification Method
- Classification Parameter
- Mean Square Error [1]
- # of Correct Predictions, # of Tests, Correct Rate, Normalized Correct Rate, Kappa, Matthews Correlation Coefficient [2]
- True Positive, False Negative, False Positive, True Negative, Correct Rate, Normalized Correct Rate, Kappa, Sensitivity, Specificity, Positive Predictive Value , Negative Predictive Value, Matthews Correlation Coefficient, and Area Under Curve [3]

[1] When the *Variable to Predict* is a numeric variable
[2] When the *Variable to Predict* is a categorical variable, and *Positive Outcome* was not specified
[3] When the *Variable to Predict* is a categorical variable, and *Positive Outcome* was specified

Deploying the Model

The steps for deploying the model are as follows:

- Perform a nested two-level cross-validation with multiple models to get the unbiased estimate of prediction accuracy (Score A)
- Do a 1-level cross-validation with the same model configurations to pick a model with the best accuracy. If more than one model tied for the best score, you may choose one of the tied models to deploy, or deploy all the

tied models and let use some type of voting scheme from the multiple predictions

- As shown in the Figure 12. 51, highlight the model to deploy, and click the *Deploy* button. In the pop-up dialog *Save Variable Selection and Classification Model*, give the model a file name, and click *Save* to save the model as a *Partek black box (.pbb)* file

Important! Report the score of the *nested* cross-validation (Score A) as the accuracy estimate of the deployed model.

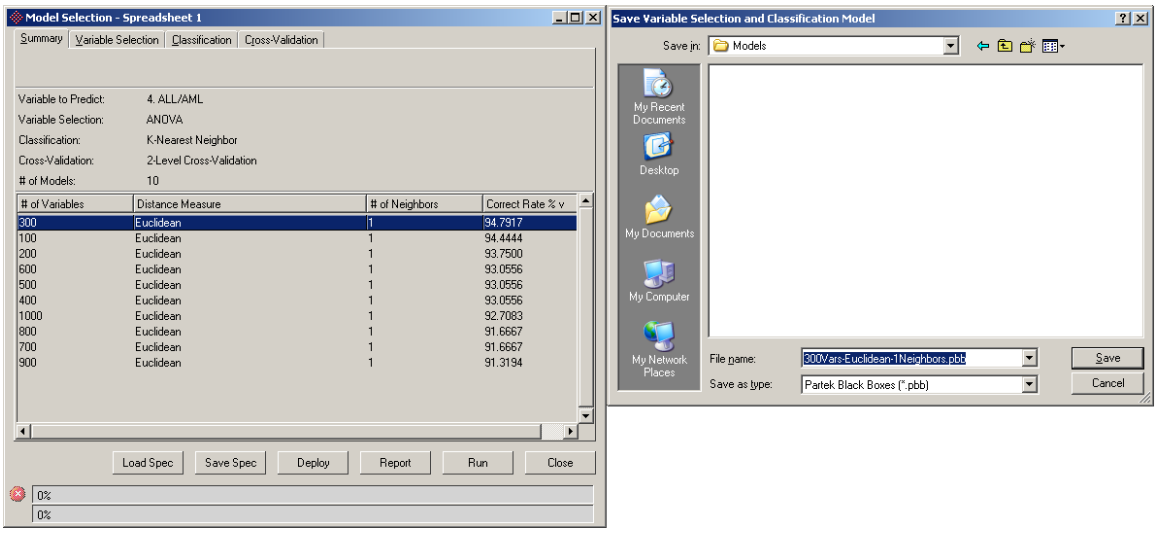


Figure 12. 51: Deploying a model in the Summary tab

Running the Deployed Model

Important! Deploying a model writes out selected variable names (previously column numbers), so the testing set does not have to include the same number of columns as the training set as long as the testing set has all of the selected variable names included. The restrictions are 1) variable names must be unique in the testing set; .2) the variables in the testing set must be in the same order as in the training set, and 3) variable names can not be empty.

- To test the deployed model with a new data set, select **Tools > Predict > Run Deployed Model…** from the Partek main window
- In the file browser dialog *Load Model File (Previously Deployed During Model Selection)* (Figure 12. 52)
- Select a saved model and click **Open**

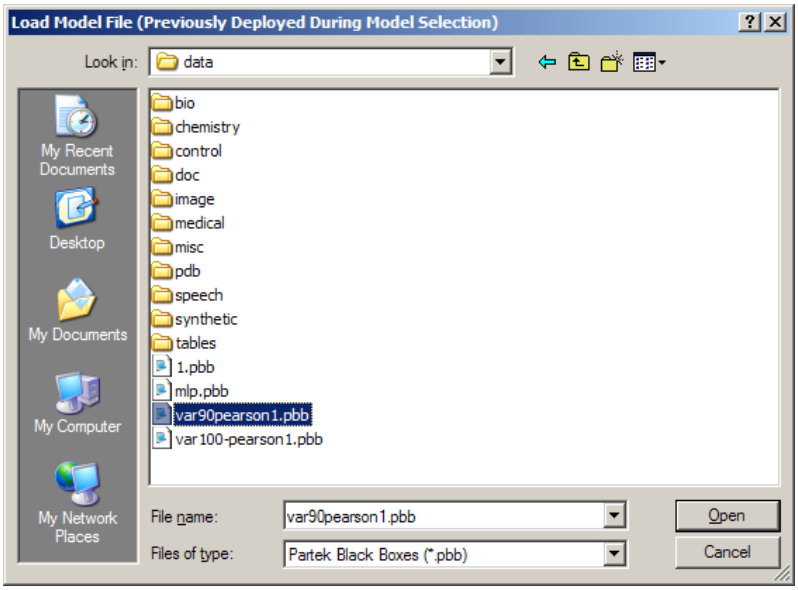


Figure 12. 52: Loading and Testing a Model

Saving the Selected Variables

The steps for saving the selected variables are as follows:

- Perform a nested two-level cross-validation with multiple models to get the unbiased estimate of prediction accuracy (Score A)
- Do a 1-level cross-validation with the same model configurations to pick a model with the best accuracy. If more than one model tied for the best score, you may choose one of the tied models to deploy, or deploy all the tied models and use some type of voting scheme from the multiple predictions
- As shown in Figure 12.53,, highlight the model, and click the **Save Var** button. In the pop-up, *Save Variable Selected* dialog, give the selected variables a file name, and click **Save** to save the file as a *Partek Format (.fmt)* file

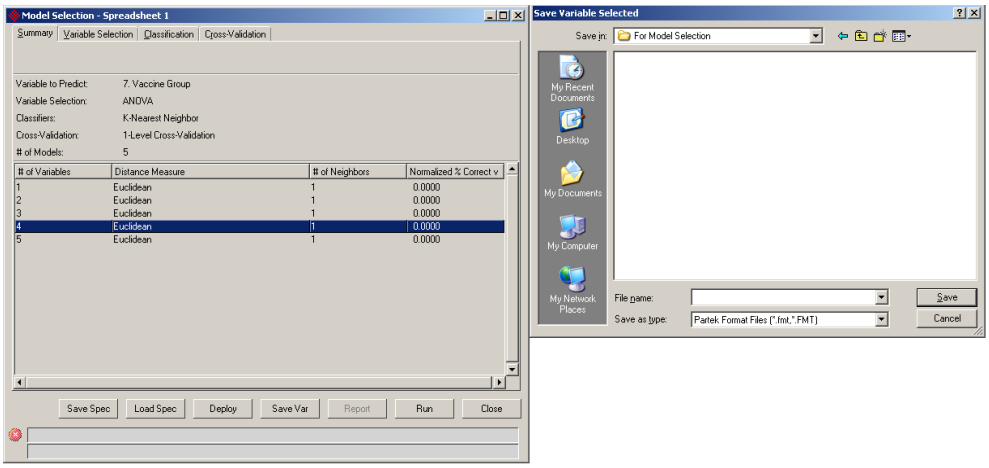


Figure 12.53: Loading and Testing a Model

Test Results

The correct rate and the prediction of each row will be shown in the *Model Test Result* pop-up dialog (Figure 12.54). Those misclassified samples are highlighted in the original data spreadsheet. Click on the *Report* button to get an HTML report (Figure 12.55). The HTML report has the following fields:

- Date
- Time
- Title
- Classification Results, the misclassified samples are also highlighted here
- Confusion Matrix
- Classification Summary
- K-Nearest Neighbor Parameters
- Variables used in the classification

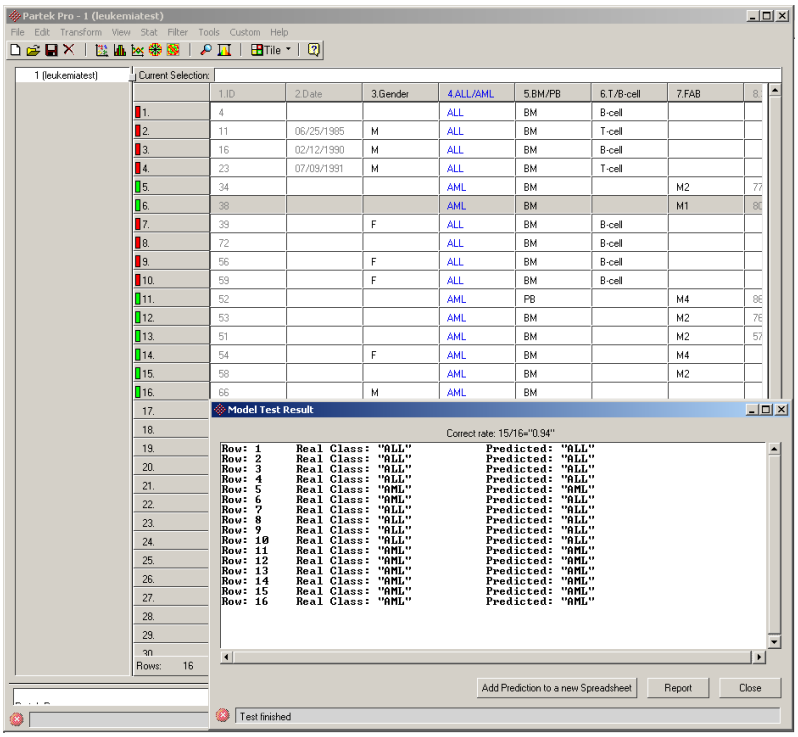
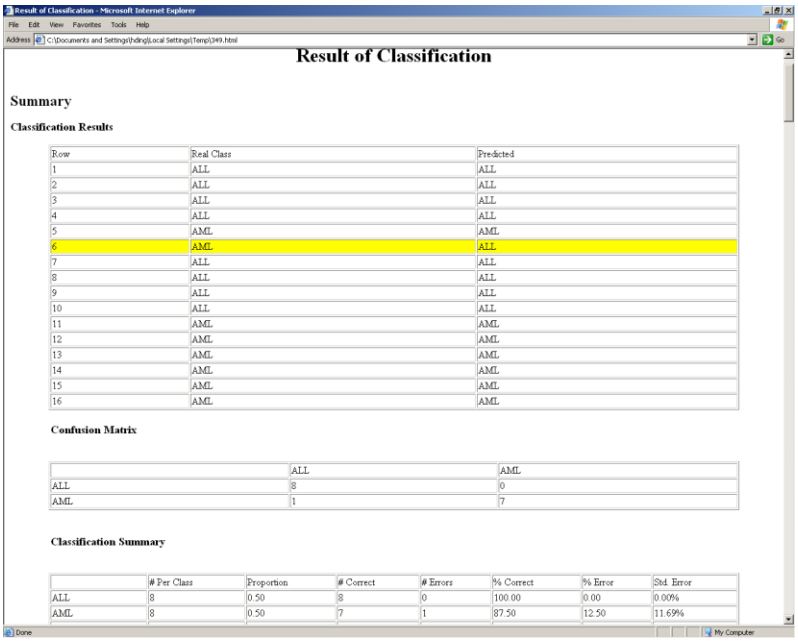


Figure 12.54: Viewing the model test results



Result of Classification

Summary

Classification Results

Row	Real Class	Predicted
1	ALL	ALL
2	ALL	ALL
3	ALL	ALL
4	ALL	ALL
5	AML	AML
6	AML	ALL
7	ALL	ALL
8	ALL	ALL
9	ALL	ALL
10	ALL	ALL
11	AML	AML
12	AML	AML
13	AML	AML
14	AML	AML
15	AML	AML
16	AML	AML

Confusion Matrix

	ALL	AML
ALL	8	0
AML	1	7

Classification Summary

	# Per Class	Proportion	# Correct	# Errors	% Correct	% Error	Std. Error
ALL	8	0.50	8	0	100.00	0.00	0.00%
AML	8	0.50	7	1	87.50	12.50	11.69%

Figure 12.55: Viewing the HTML report on the Result of Classification

Saving Specifications

Save the current *Model Selection* dialog settings and click the **Save Spec** button. In the *Save Specification* pop-up dialog (Figure 12.56) create a file name and click **Save** to save the current *Model Selection* dialog settings as a *Partek Classification Model Specification (.pcms)* file.

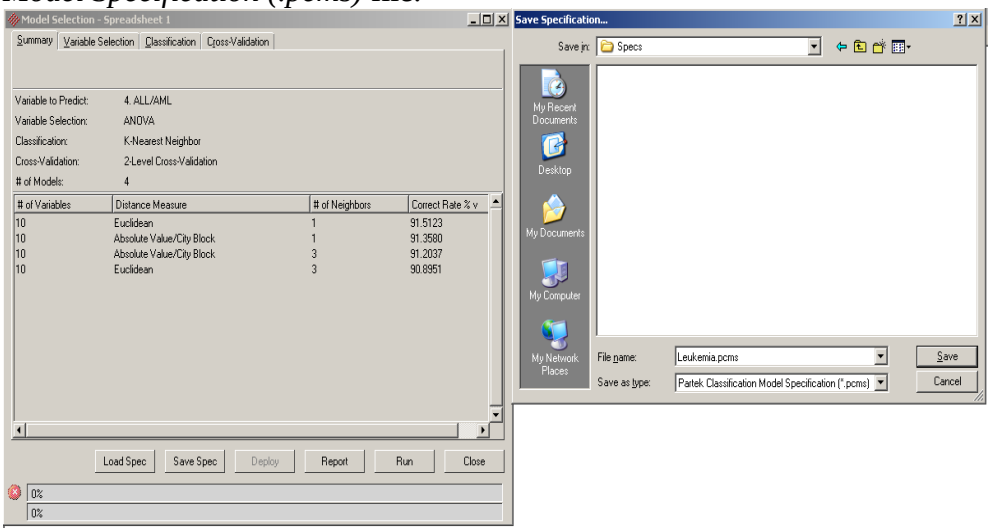


Figure 12.56: Saving the Specification

Loading Specifications

To load the saved *Model Selection* Specification, click the **Load Spec** button. In the *Load Specification* pop-up dialog (Figure 12.57). Select the saved file and click **Open** to restore the *Model Selection* dialog settings.

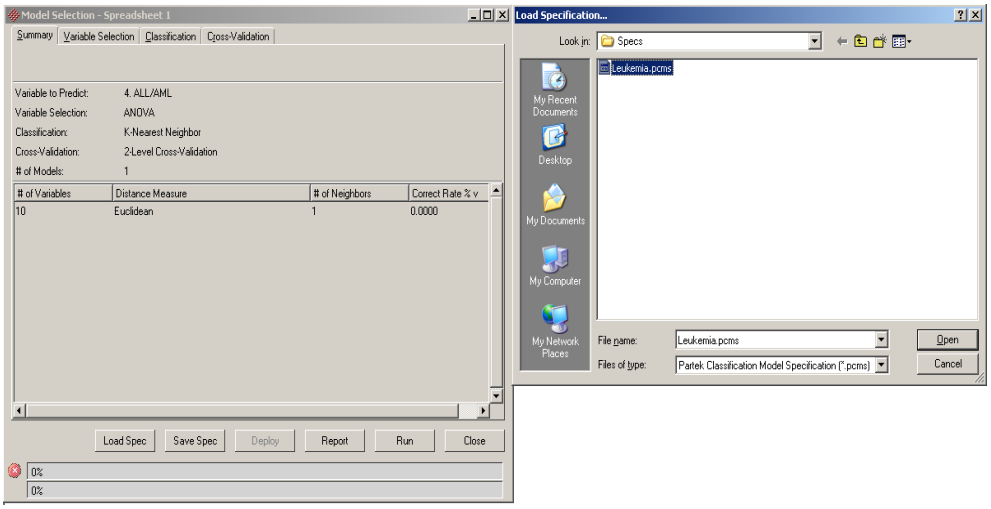


Figure 12.57: Loading the Specification

References

Anderberg, M.R., Cluster Analysis for Applications, Academic Press, 1973.

Bishop, C.M. (1995) Neural Networks for Pattern Recognition. Oxford: Claren Don Press.

Bridle, J.S. (1990), “Probabilistic Interpretation of Feedforward Classification Network Outputs, with Relationships to Statistical Pattern Recognition,” Neuro-Computing: Algorithms, Architectures and Applications, eds F. Fogleman Soulie and J. Herault, pp. 227-236. Springer-Verlag, Berlin.

Chih-Chung Chang and Chih-Jen Lin, LIBSVM : a library for support vector machines, 2001.

C.-W. Hsu and C.-J. Lin. "A comparison of methods for multi-class support vector machines" IEEE Transactions on Neural Networks, 13(2002), 415-425.

Cohen, Jacob (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement.

Downey, T.J. and Meyer, D.J., “A Genetic Algorithm for Feature Selection”, Intelligent Engineering Systems Through Artificial Neural Networks, Dagli, C.H., et al. editors, Vol. 4, 1994: 363-368, ASME Press, New York.

Duda, R.O. and Hart, P.E., Pattern Classification and Scene Analysis, John Wiley, New York, 1973.

Efron, B. and Tibshirani, R.J. (1993), An Introduction to the Bootstrap, London, Chapman & Hall.

Efron, B., (1982) The Jackknife, the Bootstrap and Other Resampling Plans. SIAM, Philadelphia.

Efron, B. and Tibshirani, R.J., (1993) An Introduction to the Bootstrap, Chapman & Hall, New York.

Fahlman, S.E. (1988), "An Empirical Study of Learning Speed in Back-Propagation Networks", Technical Report CMU-CS-88-162, Carnegie-Mellon Univer sity.

Geisser, Seymour. The predictive sample reuse method with applications. Journal of the American Statistical Association, 70:320-328, 1975.

Goldberg, D.E., Genetic Algorithms in Search, Optimization & Machine Learning, Addison-Wesley, Reading, MA (1988).

Hinton, G.E. (1986) "Learning Distributed Representations of Concepts", Proceedings of the Eighth Annual Conference of the Cognitive Science Society (Amherst, 1986), pp. 1-12. Erlbaum Press, Hillsdale.

Holland, J., Adaptation In Natural and Artificial Systems, University of Michigan Press (1975).

Hush, D.R., and Horne, B.G. (January 1993), "Progress in Supervised Neural Networks: What's New Since Lippmann?", IEEE Signal Processing Magazine, 10(1):8-39.

Lippmann, R.P. (April 1987), "An Introduction to Computing with Neural Nets," IEEE Acoustics, Speech and Signal Processing Magazine, 4(2):4-22.

Moller, M., (1993) A scaled conjugate gradient algorithm for fast supervised learning. Neural Networks 6, 525-533.

Munkres, J.R., Topology: A First Course, Prentice Hall, Englewood Cliffs, New Jersey, 1975.

Ripley, B.D. (1996) Pattern Recognition and Neural Networks, Cambridge: Cam-Bridge University Press.

Romesburg, H.C., Cluster Analysis for Researchers, Lifetime Learning Publications, Belmont, California, 1984.

Royden, H.L., Real Analysis, Macmillan Publishing Company, New York, 1988.

Rumelhart, D.E., Hinton, G.E., and Williams, R.J. (1986), "Learning Internal

Representations by Error Propagation" in D.E. Rumelhart and J.L. McClelland (Eds.), Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Vol. 1: Foundations. MIT Press.

Rumelhart, D.E. and McClelland, J.L. (Eds.) (1986), Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Vol. 1: Foundations. MIT Press.

Spath, H., Cluster Analysis Algorithms for Data Reduction and Classification of Objects, Halsted Press, New York, 1980.

Spath, H., Cluster Dissection and Analysis: Theory, FORTRAN Programs, Examples, Halsted Press, New York, 1985.

Stone, M. Cross-validatory choice and assessment of statistical predictions (with discussion). Journal of the Royal Statistical Society B, 36:111-147, 1974.

Tibshirani, R., Hastie, T., Narasimham, B., and Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc. Natl. Acad. Sci. U.S.A. 99 6567–6572.

Tibshirani, R., Hastie, T., Narasimham, B., and Chu, G. (2003). Class Prediction by Nearest Shrunken Centroids, with Applications to DNA Microarrays. Statist. Sci. Volume 18, Issue 1 (2003), 104-117Tou, J.T., and Gonzalez, R.C., Pattern Recognition Principals, Addison- Wesley,
Reading, Massachusetts, 1974.

Zar, J.H., Biostatistical Analysis (4th ed.), Prentice-Hall, New Jersey 07458.

Genomic Specific Analyses

Introduction

This chapter provides insight into allele intensity import and genomic specific analyses of allele intensity for copy number, allele specific copy number, loss of heterozygosity, genomic segmentation, and the MAT algorithm for detecting ChIP-chip regions of significance.

Allele Intensity Import

Marker-level estimates of DNA intensity is dependent on the sequence of the marker (or "probe"), and for DNA that is fragmented prior to amplification (i.e. Affymetrix), which is also dependent on fragment length. In order to reduce bias and noise in the resulting data prior to analysis, Partek provides options to estimate and remove these effects.

When importing Affymetrix data from CEL files, the intensity of individual probes are first to adjust for bias due to fragment length (smaller fragments are sometimes amplified more than larger fragments). This bias is modeled using a nonlinear function and removed on a sample-by-sample basis, leaving the unbiased residual intensities.

After optional fragment length correction, the bias due to the probe sequence are estimated and removed. GC content is known to cause bias in hybridization intensity and is a simple, fast, and effective way to remove this bias. Partek's GC correction is described in the Partek GS User manual (chapter 4). Optionally, a full sequence-based correction ("sequence correction") may be applied, which includes removal of GC bias and other sequence-based hybridization bias. Sequence correction may give slightly better results than GC correction alone, but it takes substantially longer to perform the calculation and therefore leads to slower import. Since sequence correction also adjusts for GC content, only one of the two algorithms is applied. Interrogating probes and control probes are both used and treated identically during fitting and adjustment.

For best results, it is important to apply the same settings for bias correction to both the reference ("baseline") samples and the study samples.

Each allele is summarized using the geometric mean of the multiple probes for that allele. The output intensity spreadsheet contains the log intensities for each allele or CNV probe.

The total probeset intensities are defined as:

- $I_a + I_b$ for SNPs where I_i is the intensity for allele *i*
- CNV probes are defined as just the intensity, I_{CNV}.

After calculating total probeset intensities, samples are normalized by scaling the samples' geometric mean intensity to one (0 in log space).

Creating Copy Number from Allele Intensities

Creating copy number from the summarized intensities is accomplished by normalizing each sample to the reference - either paired references or a pooled reference depending on paired or unpaired workflow.

Unpaired Baseline Creation

Inputs:

- Gender for each sample (optional)
- Chromosome variation file (optional – it includes the probes that are uniquely on the non-autosomal regions of the X and Y chromosome – which used together with gender information, can further remove bias – see below.)
- Data spreadsheet containing log base 2 normalized intensities for alleles and CNVs

Outputs:

- A file with the reference copy number intensities

Each probeset is summarized to its total intensity as $I_a + I_b$ or I_{CNV} for SNPs and CNVs respectively. Each sample's overall ($\log_2$) intensity is adjusted such that the geometric mean intensity is 1 on the non-gender influenced chromosomes (as defined by the chromosome variation file).

The pooled reference intensity is calculated as the mean of all samples for the respective probe set.

Probes within each group specified in the chromosome variation file are considered as their own normalization group and their geometric mean intensity is adjusted to 1 (or to the geometric mean of females if available on chromosome X).

If there are multiple expected copy number levels on the X chromosome, these are remembered to be used as a bias adjustment during copy number creation.

Unpaired Copy Number Creation

Inputs:

- Spreadsheet containing log base 2 normalized intensities for alleles and CNVs.
- Reference intensity file created during baseline creation.

Each probeset is summarized to its total intensity. A log ratio calculation is carried out to find the intensity relative to the reference intensity.

A bias adjustment uses the X geometric mean intensities found during baseline creation (or a predefined coefficient based on the built-in 270 HapMap reference) to adjust the log ratio values to be unbiased (i.e. females should have 2 copies of X and 0 copies of Y, while should have 1 copy of X and 1 copy of Y). This operation is simply finding the proper coefficient that will remove copy number estimation bias.

Log ratios are turned into copy number space using the formula:

$$Copy\ Number = 2 \times 2^{(log2\ intensity)}$$

As a final step, outliers are removed in a method suggested Olshen et al. An estimate of the standard deviation is found for each chromosome using the average squared lag one distance to estimate 2*var. Each local neighborhood is defined by the current probe and the two next probes and two previous probes in genomic ordering for a total of 5 probes. All probes greater than four standard deviations from the neighborhood median are adjusted to two standard deviations.

Allele Specific Copy Number

Allele specific copy number (AsCN) estimates provide potentially important information for the analysis, visualization, and interpretation of copy number changes. AsCN provides an estimated number of copies for each allele rather than an estimated number of total copies of each chromosome. In tumor samples, changes in copy number or loss of heterozygosity (LOH) often occur in only a proportion of the tissue sample due to the heterogeneous nature of biopsy samples, which include some degree of "normal" contamination. When these differences are in only a fraction of the tissue sample, genotyping algorithms may still make heterozygous calls, making it less likely to detect LOH using discrete genotype calls. AsCN allows these changes to be detected as an imbalance in copy number between alleles. In addition, AsCN is useful in the interpretation of total copy number analysis results. A gain of total copy number may be shown by one allele or both alleles. AsCN allows the researcher to resolve these uncertainties.

The calculation of allele specific copy number is as follows:
$AsCN_{ij} = f\left(I_{ij} / R_{ij}\right)$ if informative, missing otherwise

- $AsCN_{ij}$ is the allele specific copy number estimate for allele *i* of SNP *j*
- I_{ij} is the intensity of allele *i* for SNP *j*
- R_{ij} is the reference intensity of allele *i* for SNP *j*. This reference represents the expected intensity for one copy of the allele

- $f(x)$ is a function correcting bias in the intensity measurement for each allele

$$AsCN_{max,j} = max(AsCN_{A,j}, AsCN_{B,j})$$
$$AsCN_{min,j} = min(AsCN_{A,j}, AsCN_{B,j})$$

Not all SNPs are informative in a sample. When a SNP is not informative, it is treated as a missing value. A SNP is considered informative when it would be expected to be heterozygous if the tissue did not contain any copy number variations.

AsCN can be calculated using two different references depending on the experiment design:

- Paired analysis
- Unpaired analysis

These two analyses will be described below.

Paired Analysis

Paired analysis is the most ideal way to generate AsCN. For best results, we recommend the tumor and normal samples be processed and hybridized together, which reduces the chance for a difference due to batch effects. In addition, the informative SNPs will be determined based on the normal sample regardless of copy number changes in the tumor sample.

In paired analysis, Partek requires genotype calls for the normal paired sample and allele intensities for both the normal and study sample.

Using paired analysis, only the heterozygous SNPs in the reference (normal) sample are considered informative and the reference intensity R_{ij} is taken to be the intensity of allele *j* of SNP *i* in the normal sample.

Unpaired Analysis

Unpaired analysis requires genotypes and allele intensities for both the reference (normal) and study (tumor) samples. The reference intensity R_{ij} is taken as the average intensity of all reference (normal) samples that are heterozygous for SNP *i*.

Unpaired analysis considers a SNP in the tumor sample to be informative if it is a heterozygous call. In tumor samples with LOH or gains of homozygosity, the genotype algorithms will call many more homozygous SNPs than heterozygous relative to the normal samples. This limits the usefulness of unpaired analysis to be informative only to mixed tissue samples. Long stretches of homozygous calls will

appear as segments with no informative SNPs (missing values). Consequently, paired analysis is recommended, when possible

Allele Imbalance

The allele imbalance procedure uses the min and max alleles to determine regions that are believed to diverge from a "normal" balance of 1 copy each. Partek defines a proportion score for each informative SNP as:

$$Proportion = \frac{\left(AsCN_{max} - AsCN_{min}\right)}{\left(AsCN_{max} + AsCN_{min}\right)}$$

In idealized data, the following table can illustrate some common scenarios and the expected proportion scores.

$AsCN_{max}$	$AsCN_{min}$	***Proportion***	***Example Description***
1	1	0	Expected balance
2	0	1	Copy neutral LOH
1.5	0.5	0.5	Copy neutral LOH in 50% of a mixed tissue sample
1	0	1	Loss of one allele
1	0.5	0.33	Loss of one allele in 50% of mixed tissue
2	1	0.33	Gain of one allele
2	2	0	Gain of both allele—allele balance does not change.

As seen from the table, the proportion score does not change with changes in total copy number, but rather the relative mixture of each allele. Total copy number analysis is also necessary to find regions of amplification or deletion of both alleles.

To determine regions of similar allele imbalance across many SNPs, Partek transforms the allele specific copy number for each SNP into its proportion score. This score is then segmented to find regions of similar proportion. The proportion reported for each detected region is the mean proportion score of all informative SNPs in the region. The mean proportion score per segment is reported in the imbalance table, which can be sorted on proportion to find segments with the largest degree of allelic imbalance within the sample.

It is very rare to have equal $AsCN_{max}$ and $AsCN_{min}$ (both alleles with identical intensity), which would be required to produce a proportion score of 0. Since the min and max are assigned after AsCN is estimated, a region's $AsCN_{min}$ will always be lower than or equal to the $AsCN_{max}$. For this reason, we recommended considering any regions of small allele proportion as normal. When analyzing good

performing samples, we have found proportions less than 0.15 to be common in normal regions. This value may increase in noisier data and may be specific to each sample.

Loss of Heterozygosity (LOH)

Loss of Heterozygosity (LOH) in Partek uses a Hidden Markov Model (HMM) to find regions that are most likely to be loss events based on the genotype error and the expected heterozygous frequency at each SNP. Both paired and unpaired analysis are available, however paired analysis is preferred when possible as it is more accurate in its expected genotype frequencies and does not report regions of LOH caused by common haplotype blocks within the study population.

HMM Emission Probability

The HMM will uses the expected probability of observing a given genotype call for every informative SNP. The expected probability used will depend on the type of analysis being used.

Unpaired

For unpaired analysis, the probability of observing a heterozygous SNP in a region of LOH is the genotype error rate. In a region without LOH, the probability of observing a heterozygous SNP is estimated using the observed frequency from the baseline samples.

The probability of each state emitting each observed genotype is described as follows:

P(AB | LOH) = e
P(~AB | LOH) = 1 - e
P(AB | ~LOH) = O(AB)
P(~AB | ~LOH) = 1 – O(AB)

AB represents a heterozygous genotype call. The parameter e is the expected genotype error rate specified in the LOH dialog. O(AB) is the observed frequency of heterozygous calls for each SNP. If a genotype baseline is not available to estimate O(AB) for a given SNP, the default heterozygous frequency parameter value will be used.

Paired Analysis

The paired LOH analysis also uses a similar HMM model for each pair of samples. Homozygous SNPs in the paired normal do not provide any information of LOH in the study sample, and are excluded from paired analysis.

P(AB | LOH) = e
P(~AB | LOH) = 1 – e

P(AB | ~LOH) = 1 – e
P(~AB | ~LOH) = e

HMM Transition Probability

The HMM uses the expected probability of being in a current state given the previous state to find the most likely regions of LOH.

The probability of being in a state given the previous state is calculated as

$$a = e^{-d/decay}$$
$$P(S_t = S_{t-1}) = a P_{max} + (1-a) P_{initial}$$
$$P(S_t != S_{t-1}) = 1 - P(S_t = S_{t-1})$$

Where d is the number of base pairs between neighboring observations, decay is a parameter specified in base pairs, and S is the hidden state. P_{max} is specified within the dialog and represents the maximum probability of retaining the same hidden LOH state as the previous SNP. Setting the decay parameter to 0 will disable the genomic decay using P_{max} for every transition probability. This is the recommended and default setting for LOH analysis within Partek.

Genomic Segmentation

The genomic segmentation algorithm finds a segmentation according to the following criteria:

1. Neighboring regions have statistically significantly different average intensities (as defined by user-specified p-value)
2. Breakpoints (region boundaries) are chosen to give optimal statistical significance (smallest p-value)
3. Detected regions must contain a user-specified minimum number of data points

In addition to specifying a p-value threshold, the user also specifies a the minimum magnitude of change to be detected relative to the noise estimate for each chromosome. The signal to noise parameter allows the one parameter to represent the desired magnitude of change for all samples without regard to the samples' noise. Partek estimates the amount of noise for each sample using the difference between neighboring probes. This provides a good estimate of local variance with very minor influence of true biological changes.

While segmentation does not strictly produce a unique solution, it does produce a locally optimal solution, which we have found to out-perform HMM in sensitivity and specificity while requiring very little time and comparable results when compared to other algorithms such as CBS, etc. The algorithm has been developed

to handle large amounts of genomic data efficiently while dealing with many artifacts found in microarray data.

After determining the segmentation result, two one-sided t-tests are performed on the probes in each region—one test above a given threshold, and one below a threshold. The minimum p-value of these two tests will be used to determine if the region is a significant deviation from the expected normal. The specified report p-value threshold will determine if a region is reported in the detected region result. For example, if the goal is to detect regions of copy # gain or loss, one may set the upper threshold to 2.1 and the lower threshold to 1.9 so that the p-value reflects the probability of being > 2.1 or < 1.9.

Parameters

The genomic segmentation procedure is a two step process.

1. Find a segmentation that produces significantly different neighboring regions
2. Filter these regions to only report those that are of interest

Segmentation parameters:

- The minimum number of probe sets specified will search for regions that contain at least a number of probe sets
- The p-value specifies the level of significance that the regions are different
- The signal to noise parameter describes the magnitude of significant region differences relative to the noise level in each sample. Increasing this parameter will report fewer breakpoints caused by small differences between neighboring regions

Report parameters:

- Below specifies the lower test filter. Any regions with means significantly below this value will be reported
- Above specifies the upper test filter. Any regions with means significantly above this value will be reported
- The p-value specifies the level of significance required in the above two tests

Detect Regions of Significance (MAT)

The MAT algorithm (Johnson et al., 2006) is used to find regions of binding in ChIP-chip experiments for a single, or multiple samples. This was done by estimating t-statistics for each probe by using a linear model fit on a subset of probe intensities taking into consideration multiple factors such as GC content and the number of times a sequence maps to the genome. These probe-level t-statistics are used to generate MAT scores using the trimmed mean of probe-level t-statistics in a window of fixed genomic length. An empirical distribution is used to determine MAT score significance by sampling windows from the original data. After identifying regions of a specified target length as significant, they were the combined with other close regions.

To make the method more flexible and able to handle multiple factors and contrasts, Partek uses ANOVA contrast t-statistics on each probe, then identifies regions of significance using a method based on the methodology above. The main difference from those in MAT is the empirical distribution is estimated by sampling non-overlapping windows of permuted (rather than original) data.

Using the Detect Regions of Significance Dialog

- After creating an ANOVA probe level result with t-statistics added for the contrast of interest, select **Detect Regions of Significance** from the *Tiling* workflow in *Analysis* section. The *Detect Regions of Significance* dialog will appear (Figure 13. 1)

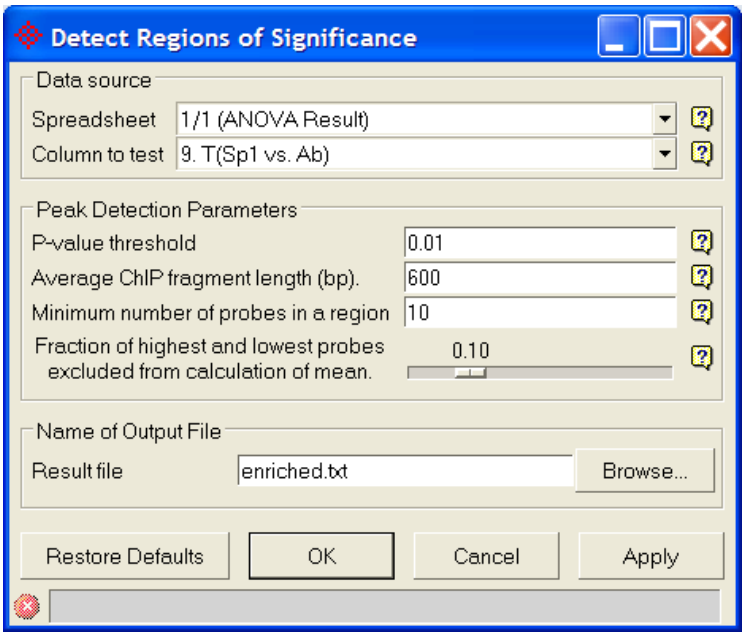


Figure 13. 1: Configuring the Detect Regions of Significance dialog

Descriptions for the options contained in the *Detect Regions of Significance* dialog are shown below:

- *Spreadsheet*: the previously created ANOVA result containing the contrast of interest
- *Column to test*: the column containing the contrast's t-statistics
- *P-value threshold*: (Default 0.001) this p-value cut off is used when determining if a region's MAT score is significant.
- *Average ChIP fragment length (bp)*: (Default 600 bp) the length of an expected ChIP region
- *Minimum number of probes in a region*: (Default 10) excluding windows with very low probe coverage from consideration can improve the specificity of the results
- *Fraction of highest and lowest probes excluded from calculation of the mean*: (Default 0.1) this represents the fraction of extreme t-statistics that will be excluded from each region when calculating the MAT score. For example, using a value of 0.1 will exclude the upper and lower 10% of the data, using the central 80% to calculate the MAT score for the region.

Descriptions of the resulting spreadsheet columns:

- *Chromosome*: the chromosome of the detected enrichment region
- *Start*: the position in base pairs of the first base of the region
- *Stop*: the position of the last base included in the detected region
- *length(bps)*: the length of the detected region in base pairs
- *probes in region*: the number of probes included in this region
- *p-value(region)*: the empirical p-value of the most significant window contained in the region
- *Fraction of negatively enriched:* represents the proportion of false positive probes included in this region. This is calculated as the # probes not significant / # probes in reported region. Regions with a high value may be less confident or only caused by a large number of outliers within the data rather than a true discovery
- *MAT-score*: the maximum MAT score for this region. A positive value means the trimmed mean of t-statistics from the specified contrast was positive, negative scores result from a negative trimmed mean of t-statistics

References

Johnson WE, Li W, Meyer CA, Gottardo R, Carroll JS, Brown M and Liu XS. Model-based analysis of tiling-arrays for ChIP-chip. *Proc. Natl. Acad. Sci. USA* 103 (2006) 12457-12462.

Olshen, A.B., Venkatraman, E.S., Lucito, R., & Wigler, M. Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics, 2004, 5(4):557-72.