Gene-Level Analysis of Exon Array Data using Partek® Genomics Suite™ 6.6

Overview

This tutorial will demonstrate how to:

- Summarize core exon-level data to produce gene-level data
- Perform exploratory analysis using a PCA scatter plot
- Identify genes that are differentially expressed

Note: the workflow described below is enabled in Partek version 6.6. Please contact the Partek Licensing Team at licensing@partek.com to request this version. The screenshots shown below may vary across platforms and across different versions of Partek.

Description of the Data Set

This experiment was performed using the Affymetrix GeneChip® Human Exon 1.0 ST Array. It includes 20 paired (normal and colon cancer) samples taken from 10 subjects.

Data and associated files for this tutorial can be downloaded by going to **Help > On-line Tutorials** from the Partek main menu. The data can also be downloaded directly from:
http://www.partek.com/Tutorials/microarray/Exon/Colon_Cancer/Colon_Cancer_DataAndImages-Exon.zip

Note: it is recommended that you read **Chapter 6 Pattern Visualization System®** chapter in the *Partek User's Manual* before going though this tutorial.

Open the Data File

For instructions on how to import CEL files, follow the **Importing Exon Array Data into Partek Genomics Suite** (Import Tutorial) tutorial from the *Partek Tutorial and Data Repository* (Help > On-Line Tutorials).

To proceed with tutorial data, open the Partek pre-imported tutorial data that already exists in a Partek format (FMT) file:

- Download Colon_Cancer_DataAndImages-Exon.zip
- Extract the files to C:/Partek Example Data/Colon Cancer (Exon)

- Select **File > Open** to invoke the *File Browser* and open the file *Colon Cancer.fmt*

Gene Summary

The first step in this will show you how to generate gene-level estimates based on the core exons from the data file.

Click on the parent (expression) spreadsheet. Select **Exon** from the workflow combo box. Click the **Gene-level analysis** button. Select **Summarize exons to genes**, and configure the *Gene Summary* dialog (Figure 1) as follows:

- Choose **Mean,** which will estimate the gene expression by averaging all the exons of that gene
- Use the default output file name
- Click **OK**

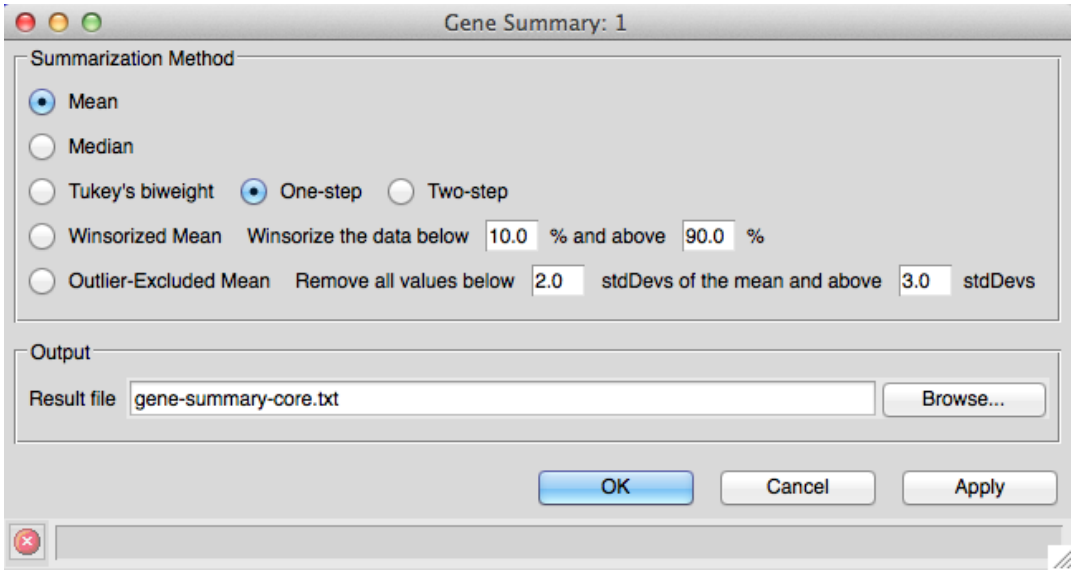


Figure 1: Configuring the Gene Summary dialog

The gene level summary will be generated in a result spreadsheet. The sample information in this spreadsheet will be the same as the parent spreadsheet. The 20 samples are on 20 rows; the columns represent genes summarized from the core exons. The column labels of this spreadsheet are transcript cluster ids, which are keys in the transcript annotation file (Figure 2).

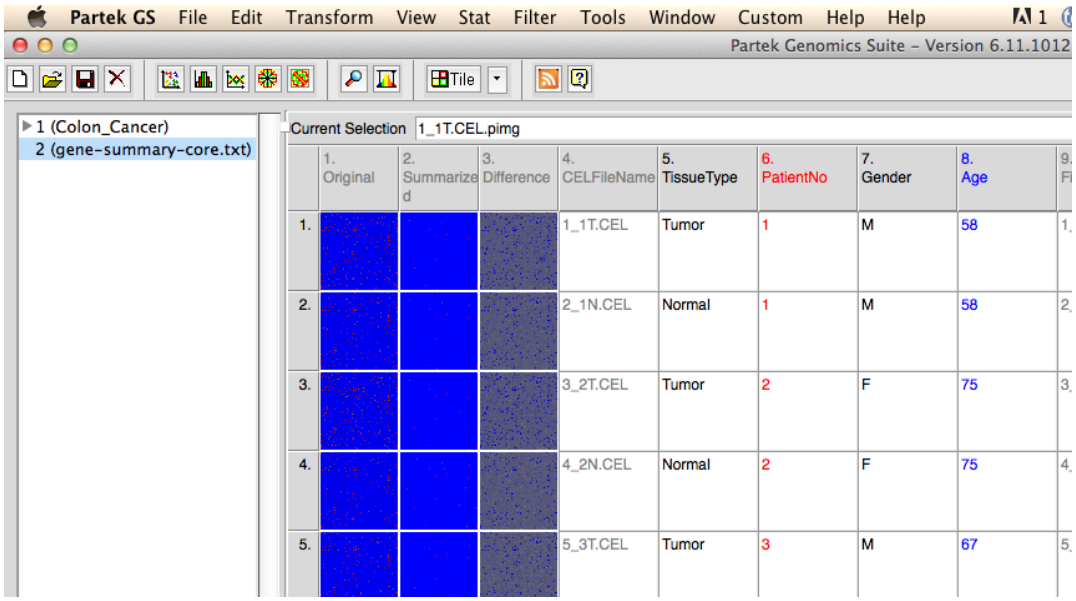


Figure 2: Viewing the gene summary spreadsheet

Identifying Differentially Expressed Genes using the Paired t-test

In this data, two tissue samples were taken in pairs from every patient; therefore, the paired sample t-test will be used to find the genes that are significantly different between the tumor and normal samples.

- Make sure the *gene-summary* spreadsheet is the active spreadsheet
- Invoke the *Paired Sample t-Test* dialog by selecting **Stat > Parametric Tests > Paired Sample t-test** from the Partek main menu
- In the *Candidate Variable(s)* panel, select **PatientNo** and move it to the *Subject ID* panel using the -> button
- Select **TissueType** and move it to the *Factor* panel (Figure 3)
- The data was log-transformed during import, so leave **Yes** checked in response to "Data is already log transformed?"

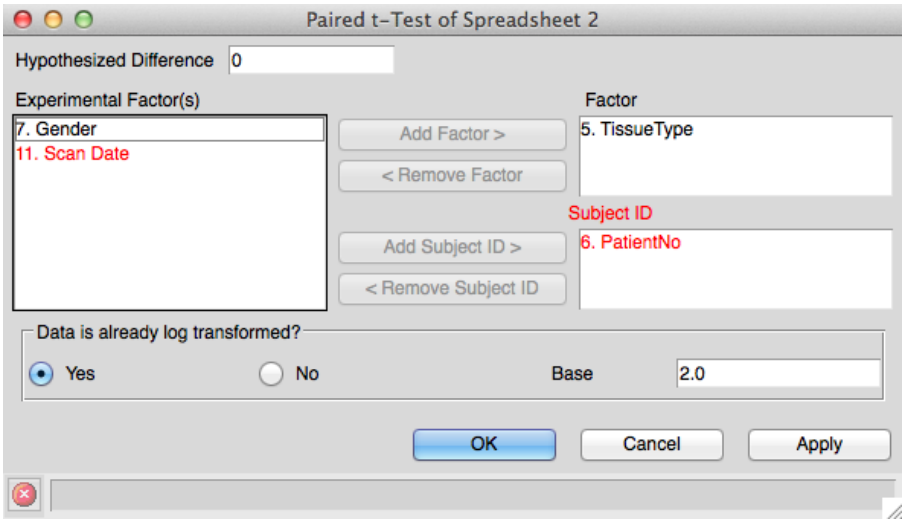
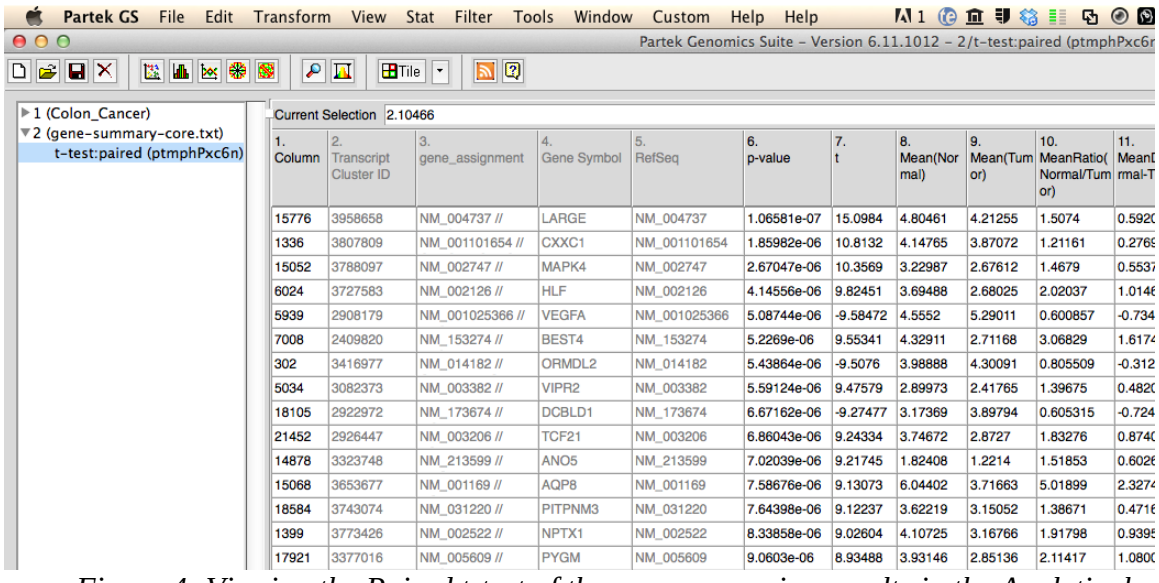


Figure 3: Configuring the Paired t-Test dialog

- Click **OK** in the dialog to compute the t-test on ~20,000 genes

The results will be displayed in a child spreadsheet, where each row represents a gene, and the columns (from Column 6 onwards) represent the statistical results for that gene (Figure 4). By default, the genes are sorted in ascending order by p-value, which means the most significant differently expressed exon between the disease tissue and normal tissue is at the top of the spreadsheet. The p-value of tissue type in this computation is the same as the p-value of tissue type in the Alt-Splicing ANOVA result for the same gene.

The second column (Transcript ID) holds the column labels from the *gene-summary* spreadsheet.



1. Column	2. Transcript Cluster ID	3. gene_assignment	4. Gene Symbol	5. RefSeq	6. p-value	7. t	8. Mean(Normal)	9. Mean(Tumor)	10. MeanRatio(Normal/Tumor)	11. MeanD rmal-T
15776	3958658	NM_004737 //	LARGE	NM_004737	1.06581e-07	15.0984	4.80461	4.21255	1.5074	0.592(
1336	3807809	NM_001101654 //	CXXC1	NM_001101654	1.85982e-06	10.8132	4.14765	3.87072	1.21161	0.276
15052	3788097	NM_002747 //	MAPK4	NM_002747	2.67047e-06	10.3569	3.22987	2.67612	1.4679	0.553
6024	3727583	NM_002126 //	HLF	NM_002126	4.14556e-06	9.82451	3.69488	2.68025	2.02037	1.014
5939	2908179	NM_001025366 //	VEGFA	NM_001025366	5.08744e-06	-9.58472	4.5552	5.29011	0.600857	-0.734
7008	2409820	NM_153274 //	BEST4	NM_153274	5.2269e-06	9.55341	4.32911	2.71168	3.06829	1.617
302	3416977	NM_014182 //	ORMDL2	NM_014182	5.43864e-06	-9.5076	3.98888	4.30091	0.805509	-0.312
5034	3082373	NM_003382 //	VIPR2	NM_003382	5.59124e-06	9.47579	2.89973	2.41765	1.39675	0.482(
18105	2922972	NM_173674 //	DCBLD1	NM_173674	6.67162e-06	-9.27477	3.17369	3.89794	0.605315	-0.724
21452	2926447	NM_003206 //	TCF21	NM_003206	6.86043e-06	9.24334	3.74672	2.8727	1.83276	0.874(
14878	3323748	NM_213599 //	ANO5	NM_213599	7.02039e-06	9.21745	1.82408	1.2214	1.51853	0.602
15068	3653677	NM_001169 //	AQP8	NM_001169	7.58676e-06	9.13073	6.04402	3.71663	5.01899	2.327
18584	3743074	NM_031220 //	PITPNM3	NM_031220	7.64398e-06	9.12237	3.62219	3.15052	1.38671	0.471
1399	3773426	NM_002522 //	NPTX1	NM_002522	8.33858e-06	9.02604	4.10725	3.16766	1.91798	0.939
17921	3377016	NM_005609 //	PYGM	NM_005609	9.0603e-06	8.93488	3.93146	2.85136	2.11417	1.080(

Figure 4: Viewing the Paired t-test of the gene expression results in the Analytical Spreadsheet

Visualizing the results of the t-Test

Right click on the **1st** row header in the results spreadsheet (the LARGE gene), and select **Dot Plot (Orig. Data)** to look at the sample distribution of the gene that is the differentially expressed between the two tissue types (Figure 5).

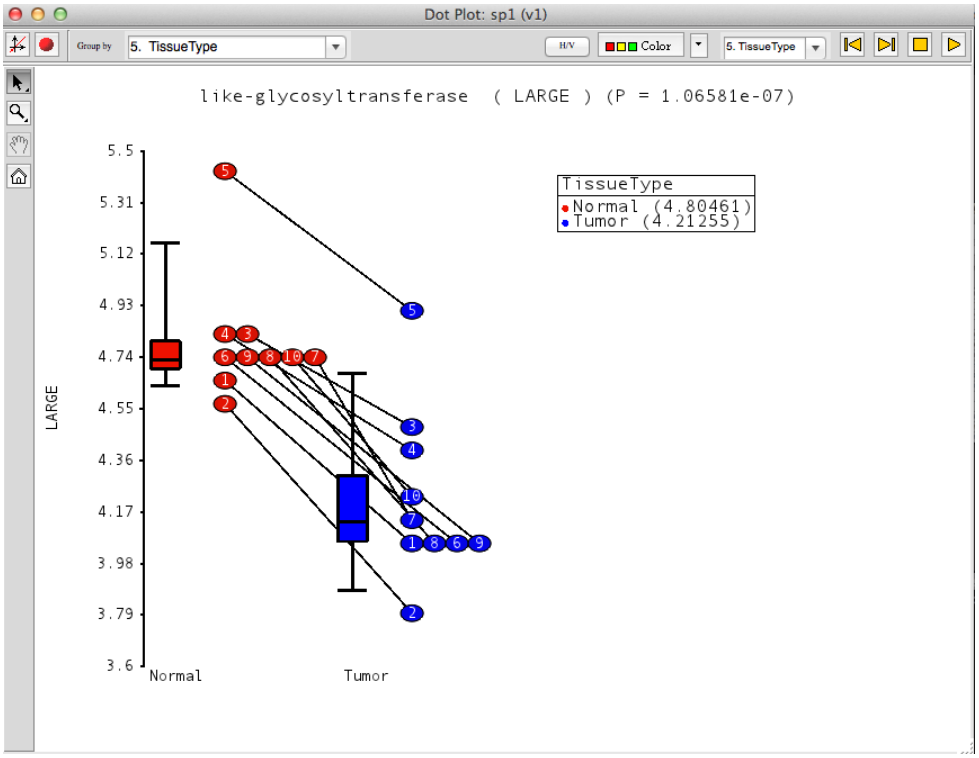


Figure 5: Viewing a dot plot of the gene whose expression is significantly different between the normal and tumor tissues

By default, the paired samples (i.e. taken from the SAME patient) are connected and labeled by patient number. The box and whiskers show the median as well as the 10th, 25th, 75th, and 90th percentiles. In this example, the LARGE gene is down-regulated in the tumour compared to the normal tissue.

In the next step we will filter the paired t-test results down to the top 100 differentially-expressed genes, and the cluster the results.

- Select the *gene-summary* spreadsheet
- Select **Filter > Filter Columns > Filter on Test Results**, the *Filter on Test Results* dialog will appear (Figure 6)

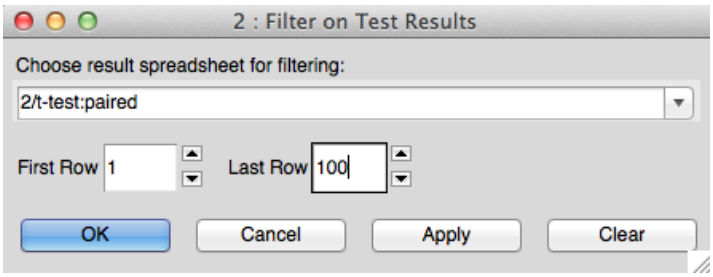


Figure 6: Filtering on Test Results

- Set *Last Row* to **100**. This will filter include in the *gene-summary* spreadsheet the expression values of the top 100 differentially expressed genes
- Click **OK**
- Select **Tools > Discover > Hierarchical Clustering**, which will invoke the *Hierarchical Clustering* dialog (Figure 7)

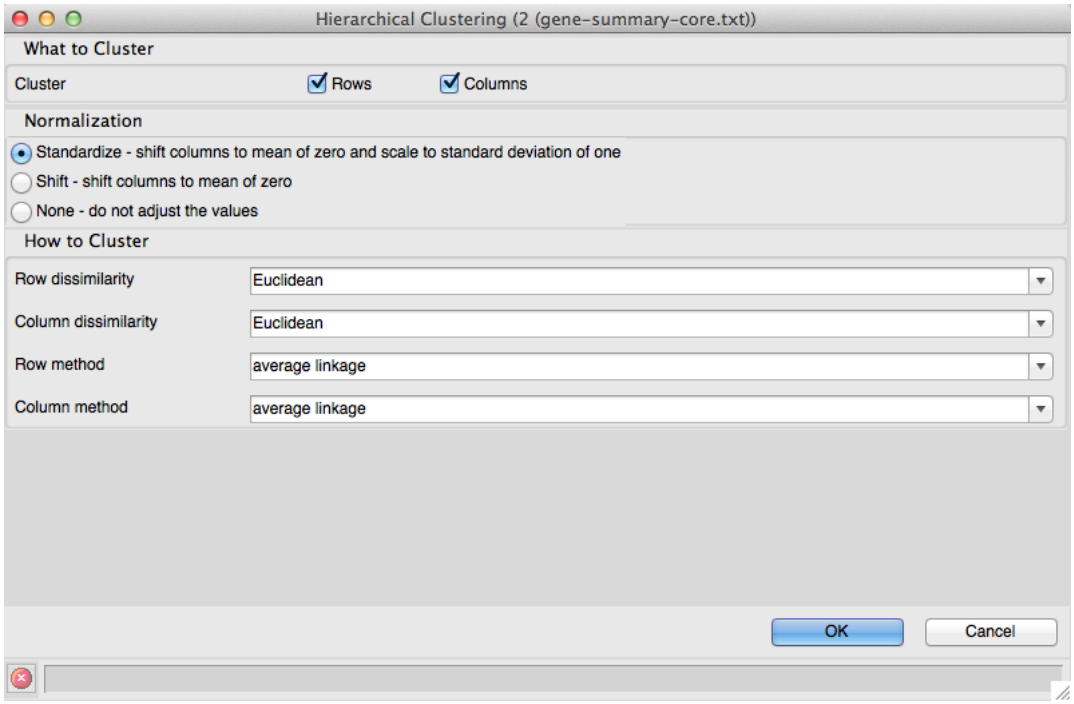


Figure 7: Invoking Hierarchical Clustering

- Use the default settings
- Click **OK**

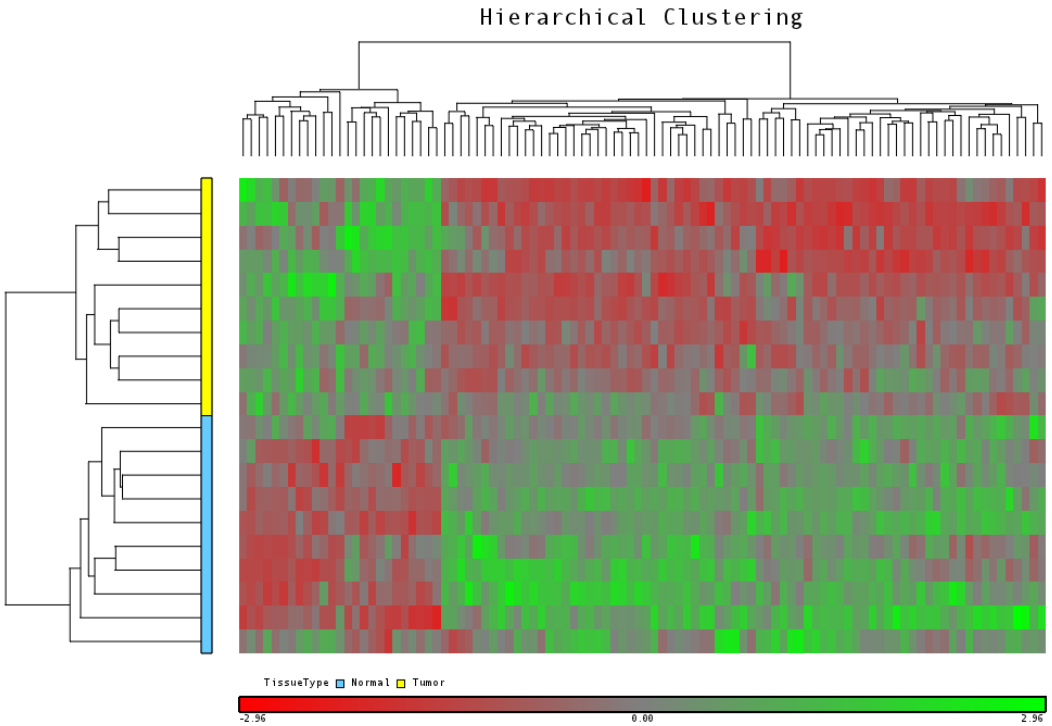


Figure 8: Clustering heat map and dendrograms

The data has been standardized by default, meaning that the data now has a mean of zero and a standard deviation of one. This allows us to more clearly see how each of the genes differentiate the normal and tumor samples.

- Close the plot before continuing

End of Tutorial

This is the end of the Exon data analysis tutorial. If you need additional assistance with this data set, you can call our technical support staff at +1-314-878-2329 or email *support@partek.com*.