Alternative-Splicing Analysis of Exon Array Data using Partek® Genomics Suite™ 6.6

Overview

This tutorial will demonstrate how to:

- Perform exploratory analysis using a PCA scatter plot
- Identify genes that undergo differential expression and alternative splicing
- Visualize exon expression patterns in a gene viewer

> 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.

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*

Exploratory Data Analysis

Start by using Principal Components Analysis (PCA) to explore the probe set summarized exon level data. PCA is a very effective method for exploring very high-dimensional data.

- Under the QA/QC section of the workflow select **Principal Components Analysis (PCA)**

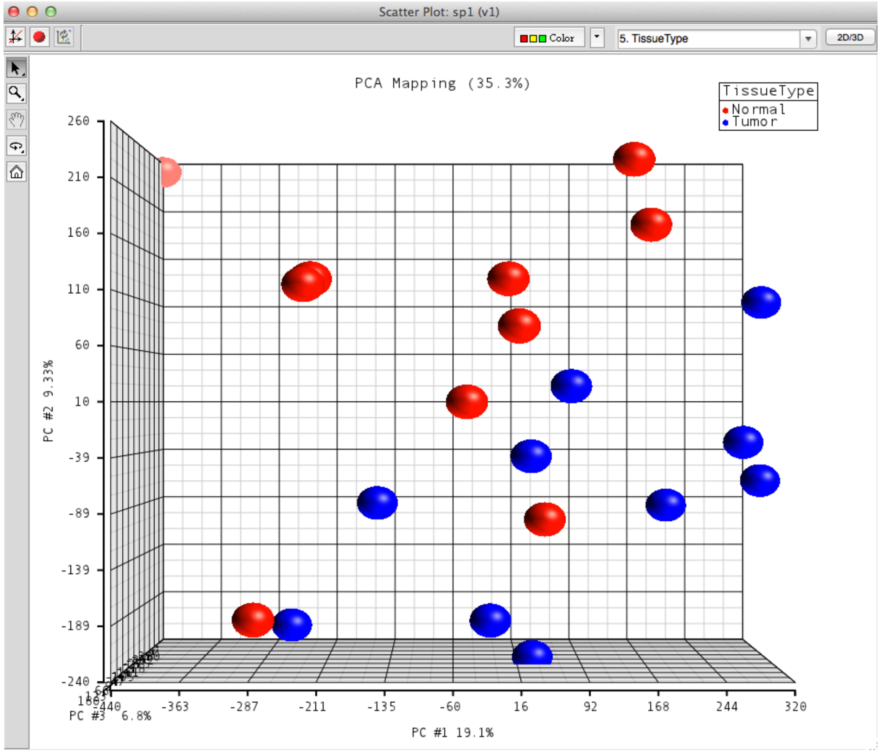
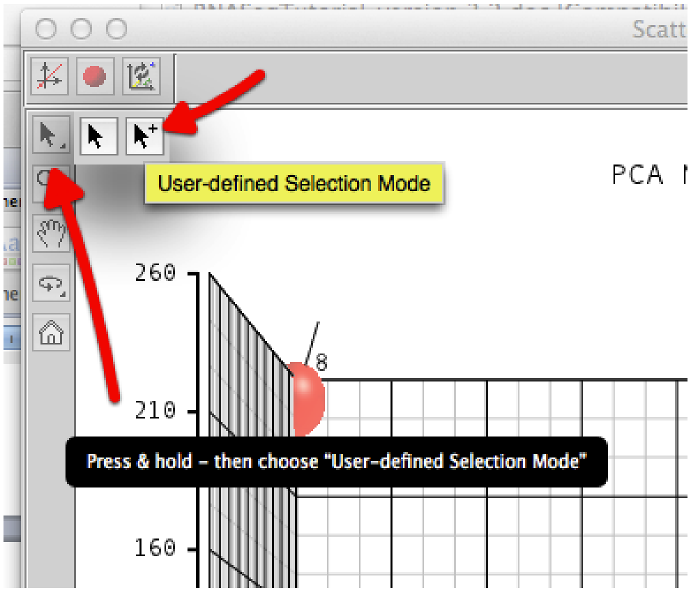


Figure 1: Viewing a PCA scatter plot of the data

In the PCA plot, each sphere represents a single sample (chip), which corresponds to a row in the spreadsheet. Clicking on any sample in the plot will highlight both the sphere on the graph, and the corresponding row in the spreadsheet. Samples that are close together in the PCA plot are similar and samples that are far apart in the plot are dissimilar. In the PCA plot shown in Figure 1, the color of the samples represents the tissue type (normal samples are red; tumor samples are blue).

Notice that there is an outlier in the upper left corner of the plot; you can examine the original chip image of the outlier.

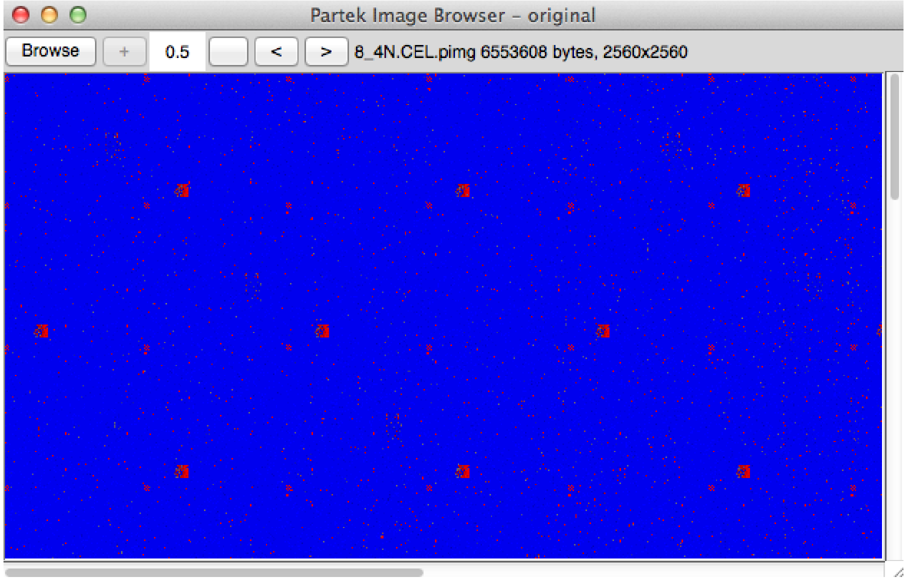
- Hold on the **Selection Mode** button in the viewer's vertical mode bar (Figure 2) and choose the **User-defined Selection Mode**



Press & hold – then choose "User-defined Selection Mode"

Figure 2: User-defined Selection Mode

- Use the default settings to show the *Original Image* in the viewer
- Click on the outlier to select it; the corresponding row in the spreadsheet will be highlighted (sample is 8_4N)
- The chip image will open in the image browser (Figure 3)



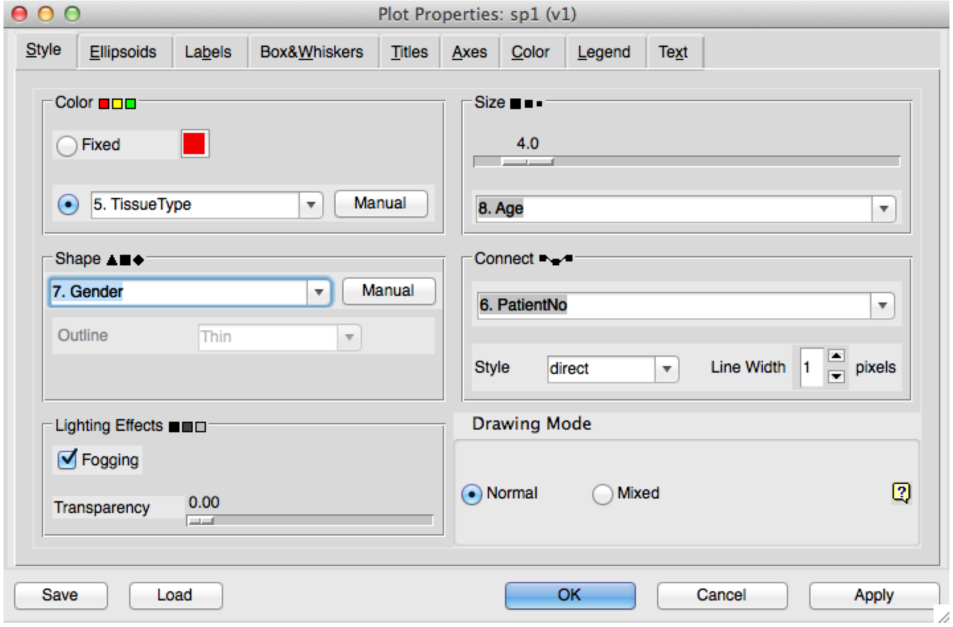
8_4N.CEL.pimg 6553608 bytes, 2560x2560

Figure 3: Viewing an individual chip image

- Close the chip viewer before proceeding

Returning to the PCA plot, on the top of the scatter plot viewer, click on the *Plot Properties* icon (●) and configure the plot as follows (Figure 4a):

- *Size* the points by **Age**
- *Shape* by **Gender**
- Whilst still in the Shape section, click on the **Manual** button and select the gender symbols from the drop-down list (Figure 2b)
- *Connect* by **PatientNo**
- Click **Apply**



Color: ○ Fixed 🟥; ● 5. TissueType

Size: 4.0; 8. Age

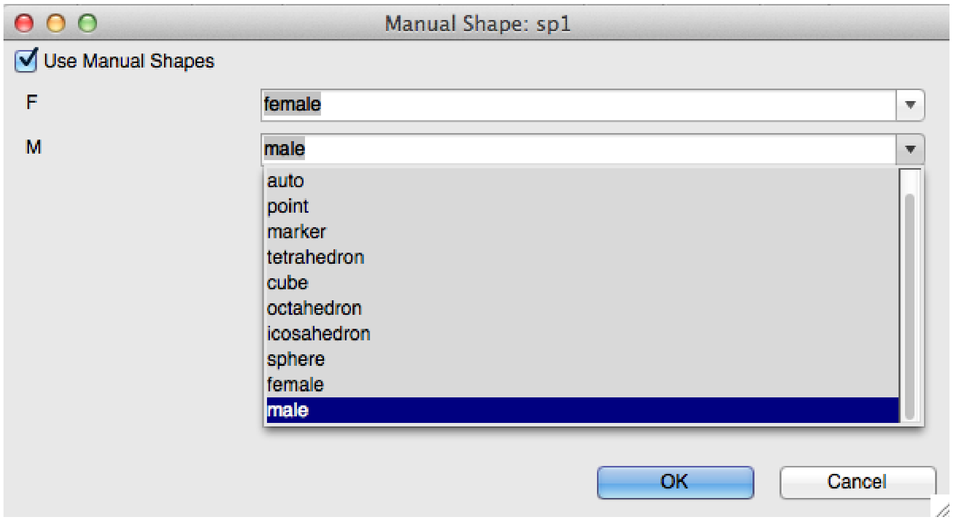
Shape: 7. Gender; Outline: Thin

Connect: 6. PatientNo; Style: direct; Line Width: 1 pixels

Lighting Effects: ☑ Fogging; Transparency: 0.00

Drawing Mode: ● Normal ○ Mixed

Figure 4a: *Configuring the scatter plot's Plot Properties dialog*



☑ Use Manual Shapes

F: female

M: male

Figure 2b: Manually shaping the samples using gender symbols

- The PCA plot is now configured to visually display the sample attributes. (Figure 3)

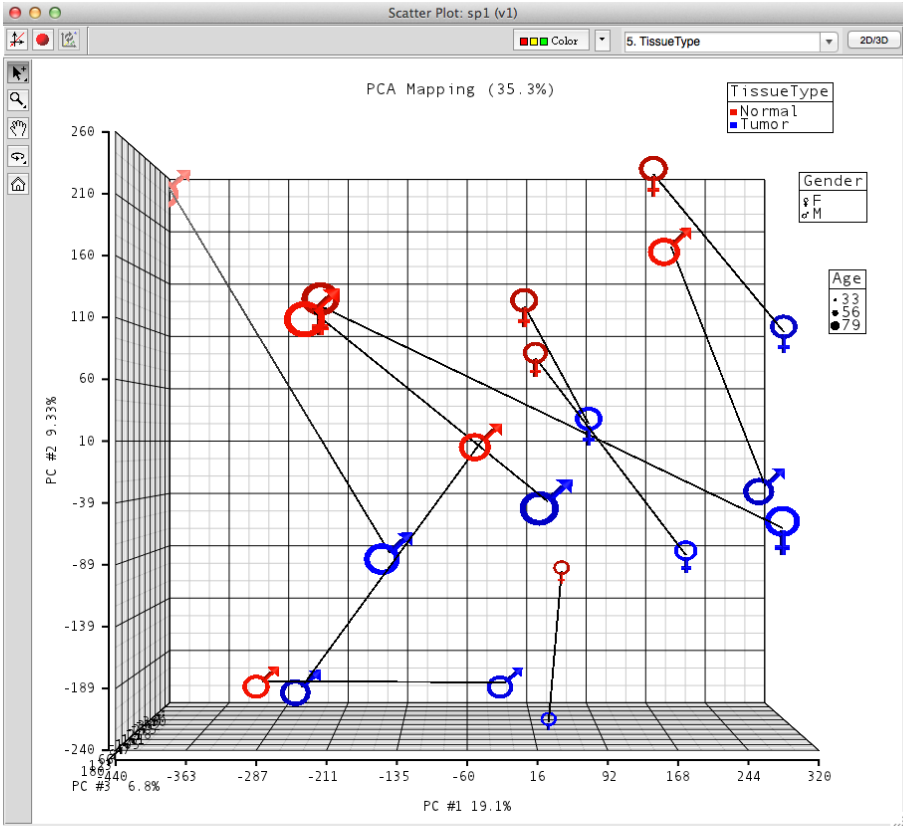


Figure 3: Viewing a scatter plot of the data colored by Tissue, sized by Type, and grouped by Tissue

PCA is an example of exploratory data analysis and is useful for identifying outliers and major effects in the data. Notice that, in this graph, most of the lines that connected the two tissues from each subject are almost parallel to PC1 (with the exception of the subject that has an outlier normal tissue). This means a lot of exons are differently expressed between the two tissue types.

To more clearly see the separation between normal and tumor tissue press the mouse wheel and drag it to rotate the plot or choose the *Rotate Mode* option ( ) and press and drag the left mouse button. Rotate the plot so to bring the Z-axis (PC #3) labels to the front (Figure 4).

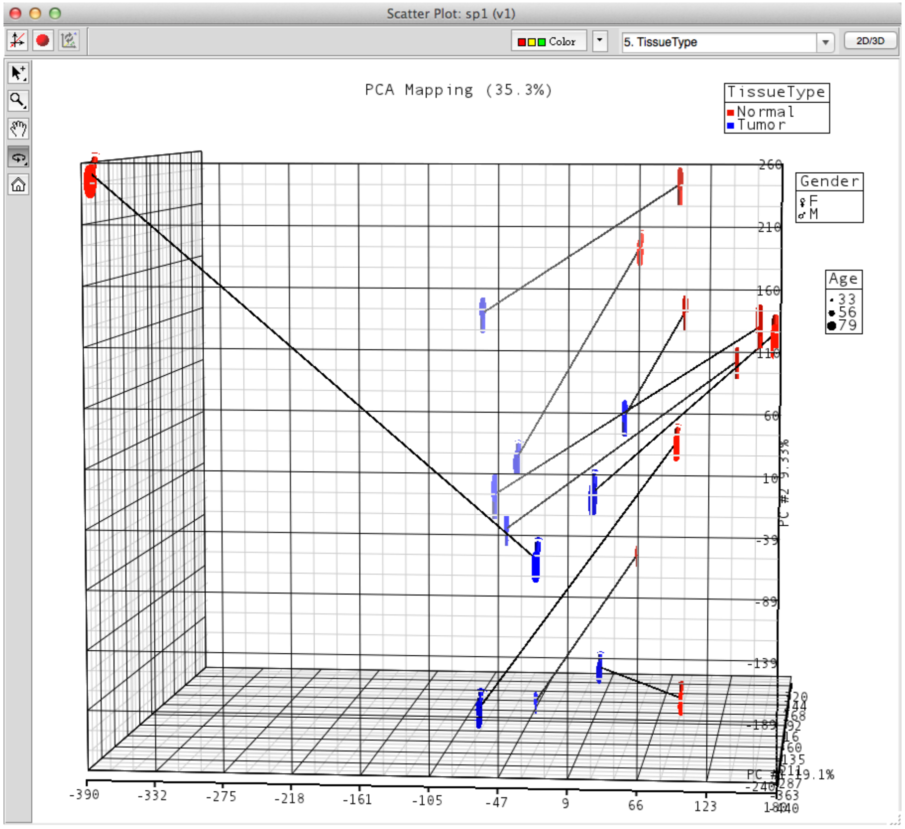


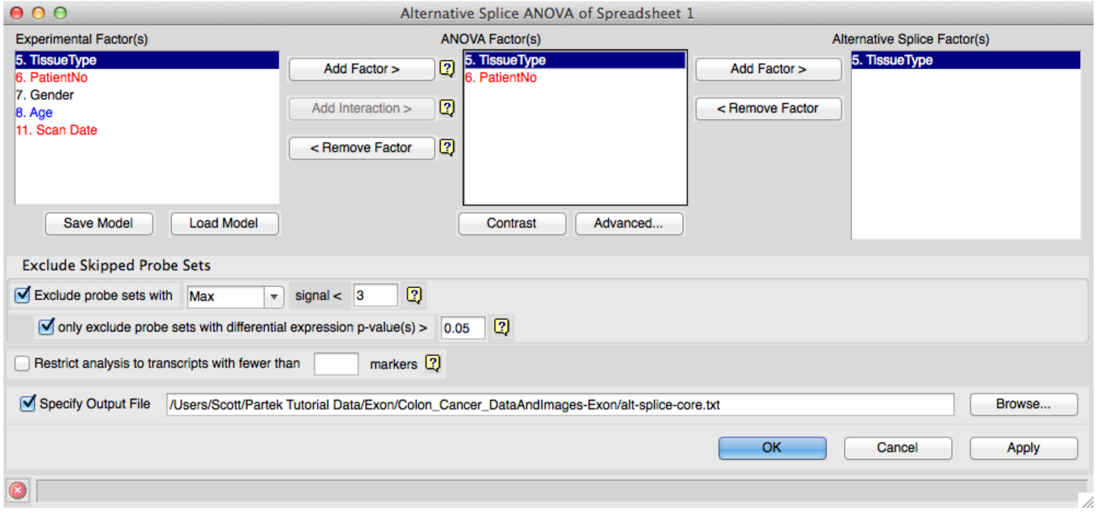
Figure 4: Viewing the rotated scatter plot to see tissue separation

Notice that with the exception of the selected outlier, the tumor tissue is on the left and the normal tissue is on the right.

- Close the scatter plot before continuing

Alternative Splicing Analysis

- Select **Exon** from the *Workflows* drop down
- Select the Alternative Splicing Analysis button
- Select **Detect Alternatively Spliced Genes** to invoke the *Alternative Splice ANOVA* dialog
- Choose **TissueType** and **PatientNo** from the *Experimental Factor(s)* panel, and move them to the *ANOVA Factor(s)* panel by clicking the Add Factor -> button
- Select **TissueType** from the *ANOVA Factor(s)* panel and click the Add Factor -> button to move it to the *Alternative Splice Factor(s)* panel (Figure 5)



Alternative Splice ANOVA of Spreadsheet 1

Experimental Factor(s): 5. TissueType, 6. PatientNo, 7. Gender, 8. Age, 11. Scan Date

ANOVA Factor(s): 5. TissueType, 6. PatientNo

Alternative Splice Factor(s): 5. TissueType

Exclude Skipped Probe Sets

☑ Exclude probe sets with Max signal < 3

☑ only exclude probe sets with differential expression p-value(s) > 0.05

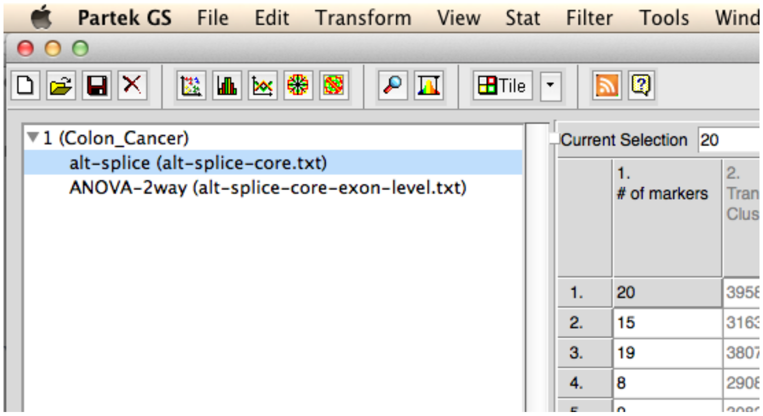
□ Restrict analysis to transcripts with fewer than markers

☑ Specify Output File /Users/Scott/Partek Tutorial Data/Exon/Colon_Cancer_DataAndImages-Exon/alt-splice-core.txt

Figure 5: Configuring the alternative splicing analysis dialog. Specify ANOVA model, the maximum number of probe sets to be detected, and the result file that will be generated

- Check **Exclude probe sets** and **differential expression p-value(s) > to** filter out probe sets which do not express in any of the transcripts of these samples (these should be checked by default)
- The **Exclude Probe Sets** option will remove any probe set (exon) that meets the specified limit. Using the default options, this will remove low expression (non-responsive) exons. In subsequent visualizations, these exons will be displayed as translucent, rather than black
- The sub-checkbox, **differential expression p-values,** provides an override to the low expression limit. Here, an exon will be included in the analysis despite a low expression value ***if*** the exon displays a p-value below the specified limit, suggesting that the exon is differentially expressed
- Leave the **Restrict analysis** box unchecked. This option would limit the results to only transcript clusters (genes) that have fewer than the specified number of exons. Sometimes this helps to ease interpretation, as transcript clusters with many exons can be difficult to interpret. The resulting ANOVA table has a column with number of probe sets, thus enabling this same level of filtering to occur at a later point using the *Partek Interactive Filter*
- Use the default file name for the *Result file*, which will be generated in the same folder and the data file. Alternatively, the output file can be renamed as desired

The ANOVA results will be displayed in two spreadsheets***: alt-splice*** and ***ANOVA*** (Figure 6). The ***ANOVA*** table displays each exon as a separate row and displays p-values showing differential exon expression according to the specified category. In the ***alt-splice*** table each row represents a transcript cluster (gene). The first column is the number of exon probesets in that transcript cluster; the second column is derived from the meta-probeset file and is equal to the transcript_cluster_id. The values in this column act as keys for the transcript annotation file. The remaining columns are the ANOVA results such as p-values of the factors and interactions in the subsequent columns.



1 (Colon_Cancer)
alt-splice (alt-splice-core.txt)
ANOVA-2way (alt-splice-core-exon-level.txt)

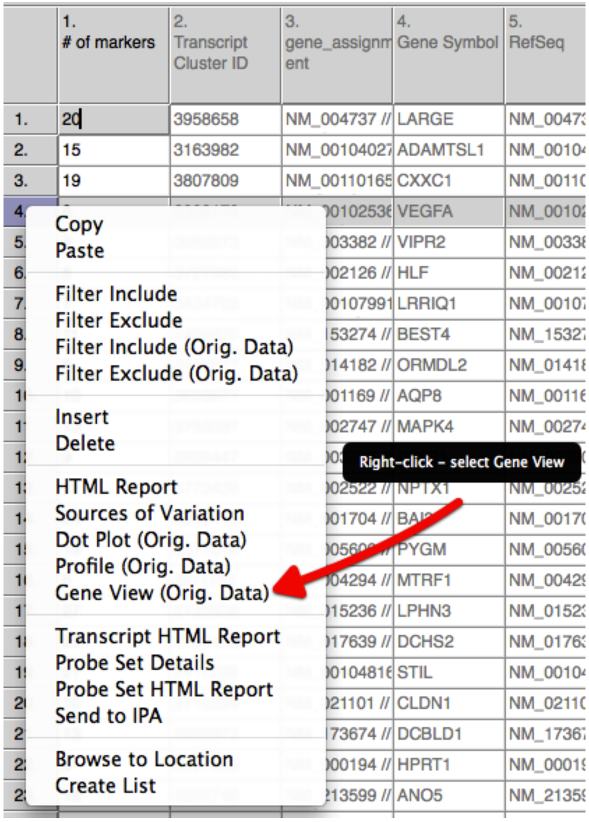
Current Selection 20

	1. # of markers	2. Tran Clus
1.	20	3958
2.	15	3163
3.	19	3807
4.	8	2908

Figure 6: Reviewing the Alt-splice result spreadsheet, genes are sorted by the p-value of TissueType

By default, the results are sorted by the first p-value, which, in this case, is TissueType. The p-value TissueType shows genes that are differentially expressed between normal and tumor tissues are at the top with small p-values. Despite the label of "alt-splice" for this table, gene-level results are displayed and the p-values in this column are consistent with the p-values of tissue type at gene level analysis (not shown in this tutorial).

- Right click on the row header that corresponds to the VEGF gene (**4th** row) of the alternative splicing result spreadsheet
- Select **Gene View (Orig. Data)** from the pop-up menu (
  Figure 7)



	1. # of markers	2. Transcript Cluster ID	3. gene_assignment	4. Gene Symbol	5. RefSeq
1.	20	3958658	NM_004737 //	LARGE	NM_00473
2.	15	3163982	NM_00104027	ADAMTSL1	NM_00104
3.	19	3807809	NM_00110165	CXXC1	NM_00110
4.			0102536	VEGFA	NM_00102
5.			03382 //	VIPR2	NM_00338
6.			02126 //	HLF	NM_00212
7.			0107991	LRRIQ1	NM_00107
8.			53274 //	BEST4	NM_15327
9.			14182 //	ORMDL2	NM_01418
10.			01169 //	AQP8	NM_00116
11.			02747 //	MAPK4	NM_00274
12.			03		
13.			02522 //	NPTX1	NM_00252
14.			01704 //	BAI	NM_00170
15.			05600 //	PYGM	NM_00560
16.			04294 //	MTRF1	NM_00429
17.			15236 //	LPHN3	NM_01523
18.			17639 //	DCHS2	NM_01763
19.			0104816	STIL	NM_00104
20.			21101 //	CLDN1	NM_02110
21.			73674 //	DCBLD1	NM_17367
22.			00194 //	HPRT1	NM_00019
23.			213599 //	ANO5	NM_21359

Right-click – select Gene View

Figure 7: Selecting a Gene View to visualize VEGF gene that is differentially expressed between two tissue types

This gene is up-regulated in the tumor samples (Figure 8).

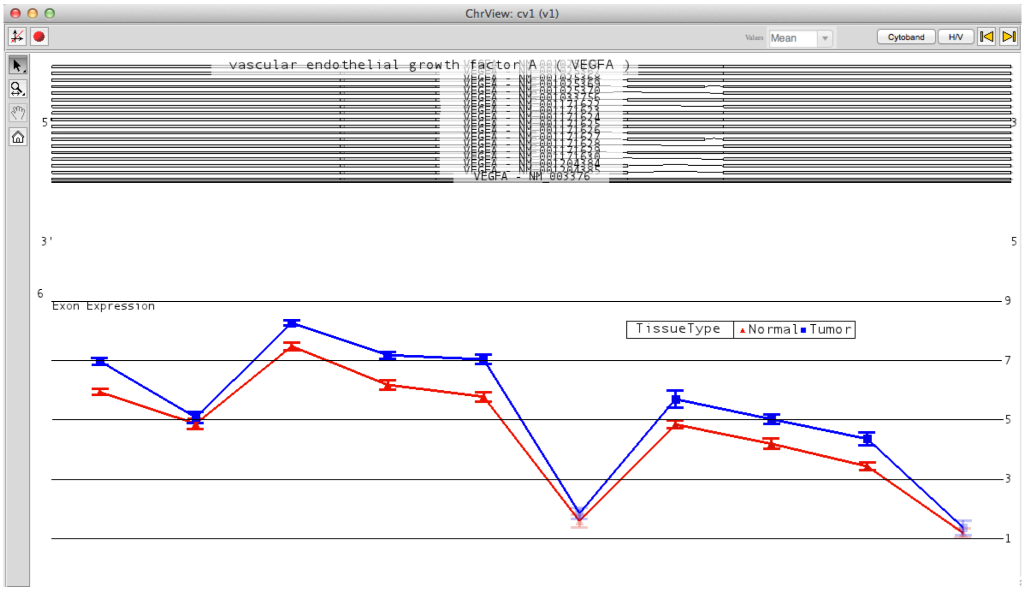


Figure 8: Viewing a gene view of the VEGF gene showing it is up-regulated in tumor samples

FAQ: Why is there so much variability in the intensities of individual probe sets within a transcript cluster (gene)?

Answer: An exon intensity value is a product of (typically) only 4 probes and has a constrained probe selection region. There is less genomic area to "average" the intensities, so this variability is due to differences in both probe quality/performance and physical sequence properties. But remember, within a given exon across two samples, those properties are the same. To look for differential expression across a gene, monitor the parallel nature of the lines between two sample groups.

Two of the exon probe sets are drawn translucent; they were filtered out of the analysis by opting to remove low intensity exons within the Alt-splice ANOVA dialog. These exons are drawn as transparent in the gene view to aid in isoform mapping. The low expression can also be visualized by adding clustering and a heatmap to the plot (Figure 12).

- Seelct the *Plot Properties* icon (●) in the view tool bar
- Select the *Clustering* tab and check **Show clustering**
- Color *dendrogram* by **TissueType**

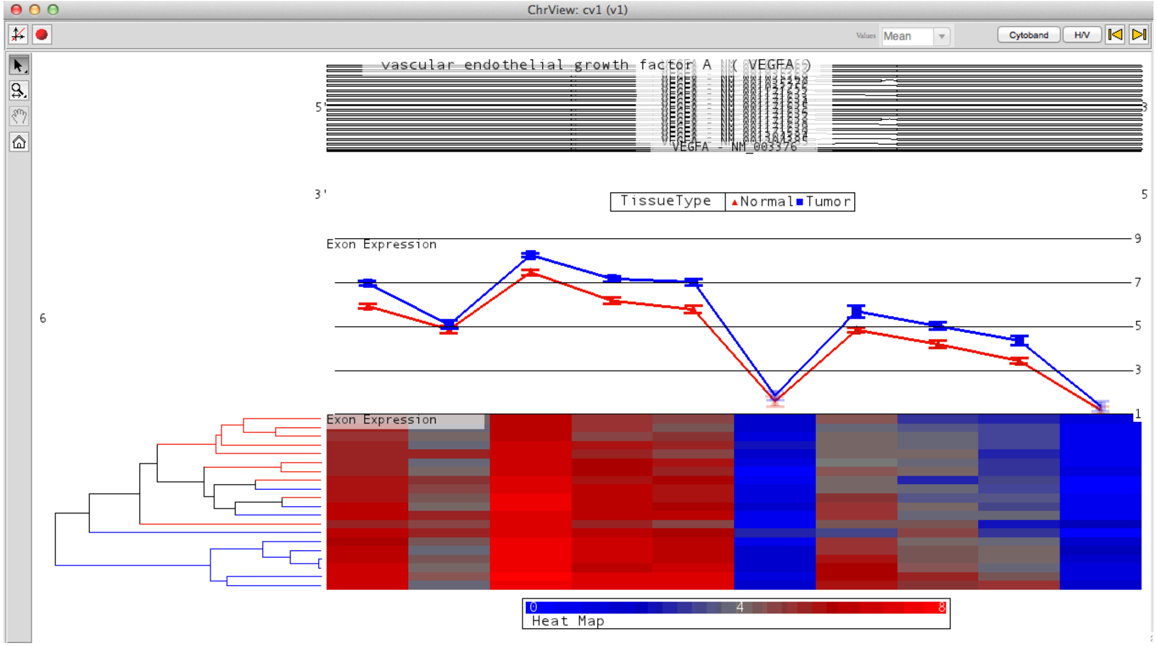


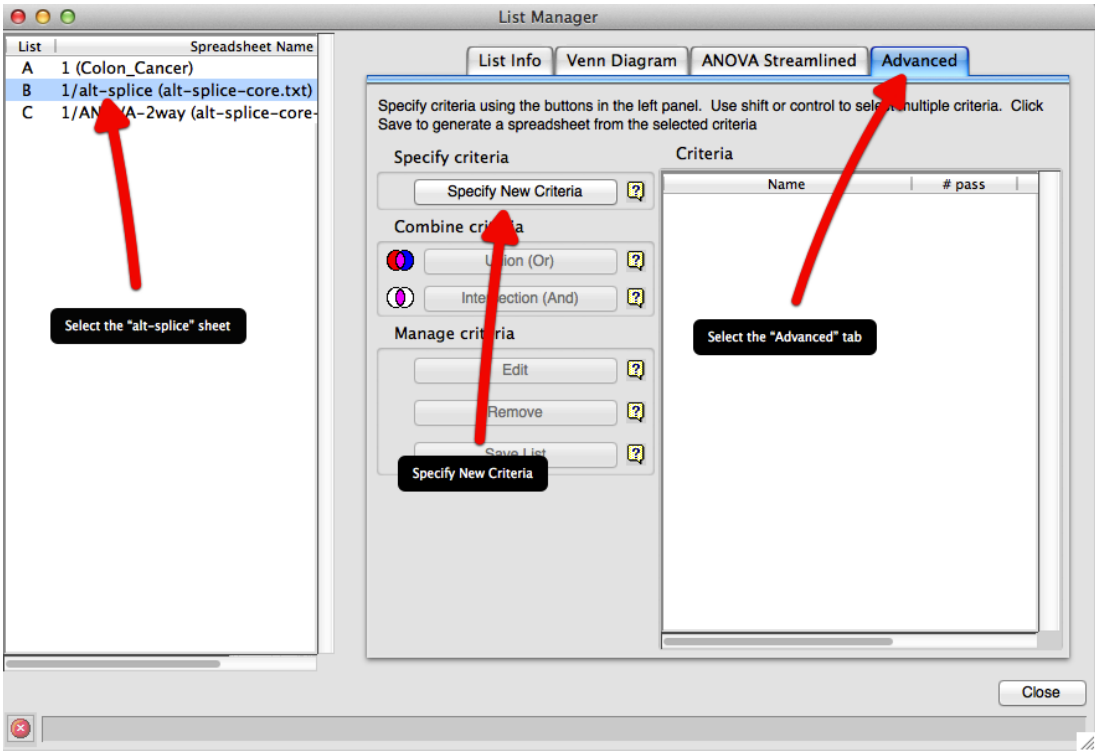
Figure 9: Viewing a gene view with clustering by tissue type

Common Analysis Scenarios

In Partek, the list manager can be used to specify numerous conditions to use in the generation of a list of genes or exons of interest. The following scenarios will illustrate how to use the list manager to create commonly sought lists

Case #1 – Finding genes that are differentially expressed, but NOT alternatively spliced

- Open the ***alt-splice*** sheet and invoke the List Manager (Figure 13a) dialogue by pressing **Create Gene List** under the *Alternative Splicing Analysis* section of the Exon workflow
- The active sheet will be used as the **default** source; you can manually choose which sheet to create the list from in the list manager
- Click on the "Advanced" tab



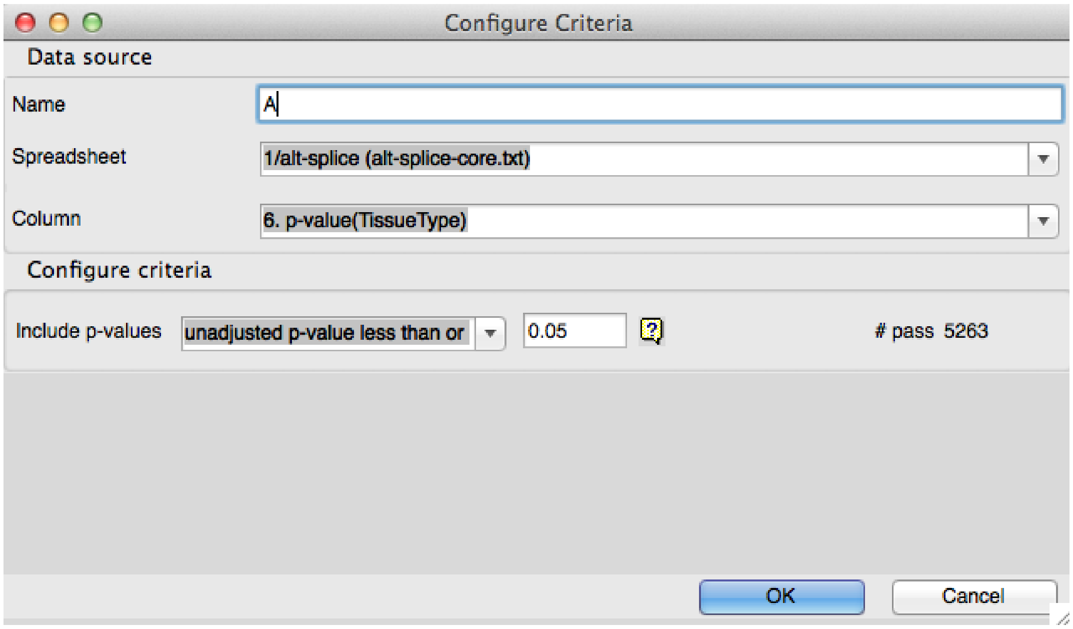
List	Spreadsheet Name
A	1 (Colon_Cancer)
B	1/alt-splice (alt-splice-core.txt)
C	1/ANOVA-2way (alt-splice-core-

Specify criteria using the buttons in the left panel. Use shift or control to select multiple criteria. Click Save to generate a spreadsheet from the selected criteria

Name	# pass

Figure 13a: The List Manager

- Now select **Specify New Criteria** – we will create a list of differentially expressed genes



Name: A
Spreadsheet: 1/alt-splice (alt-splice-core.txt)
Column: 6. p-value(TissueType)
Include p-values: unadjusted p-value less than or 0.05
# pass 5263

Figure 13: Create List dialog

- Name this Criteria “A”
- Ensure the ***alt-splice*** sheet is selected and choose column **6. p-value (TissueType)**
- Set the default criteria to look for differentially expressed genes. Do this by choosing **unadjusted p-value less than or equal to 0.05** from the *Include p-values* drop down (this value can be changed, but we will use the default settings for the tutorial)

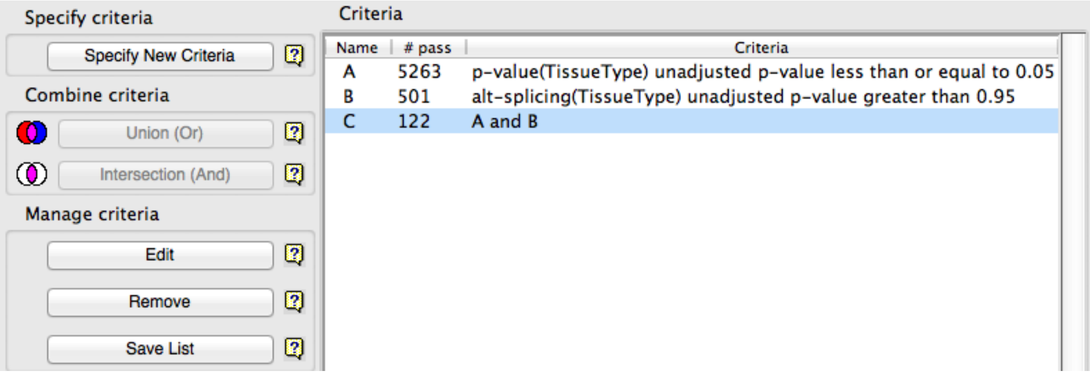
These criteria will generate a list of 5263 genes showing differential expression, 5% of which are likely false positives.

- Now create a second list to eliminate alternatively spliced genes
- Select **Specify New Criteria** again and use the same sheet
- Name this Criteria “B”
- Choose *Column* **8. alt splicing(TissueType)** representing our alternative splicing p-values
- To eliminate alternatively spliced genes change *Include p-values* to unadjusted p-value greater than and *Value* to **0.95**

This second list now includes 501 genes showing NO alternative splicing, 5% of which are likely false negatives.

- To find the genes that meet both sets of criteria we now combine the lists
- Select both lists (A and B) and use the **Intersection (And)** option under the *Combine Criteria* header – call this list “C”
-

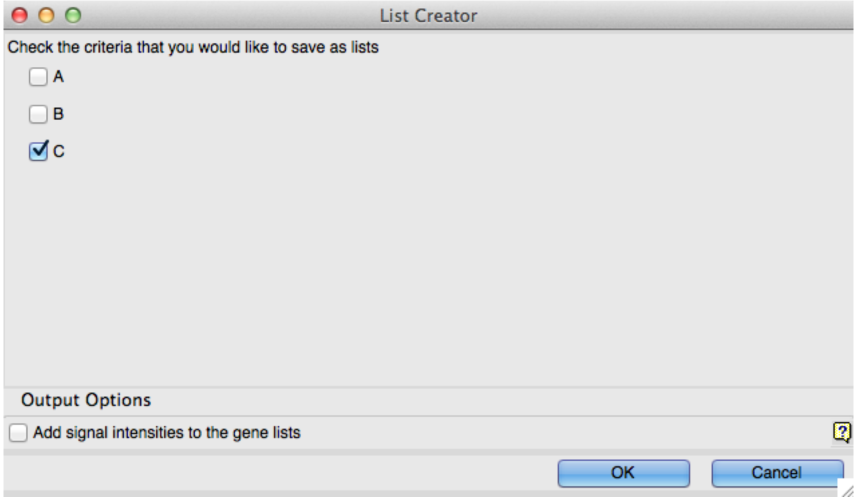
- Beside each criterion the number of genes that pass is displayed, along with the actual criteria used (Figure 14)



Name	# pass	Criteria
A	5263	p-value(TissueType) unadjusted p-value less than or equal to 0.05
B	501	alt-splicing(TissueType) unadjusted p-value greater than 0.95
C	122	A and B

Figure 14: Viewing the criteria and results for each list

- Select Save List
- Check the lists you wish to save, in this case select "C" (Figure 15)



Check the criteria that you would like to save as lists

☐ A

☐ B

☑ C

☐ Add signal intensities to the gene lists

Figure 15: Create List saving window

- Select **OK**

And example of a gene in this list can be seen in Figure 16 – as before (refer to Figure 10), find a gene of interest in the spreadsheet, right-click and select Gene View (GALR1 in this example).

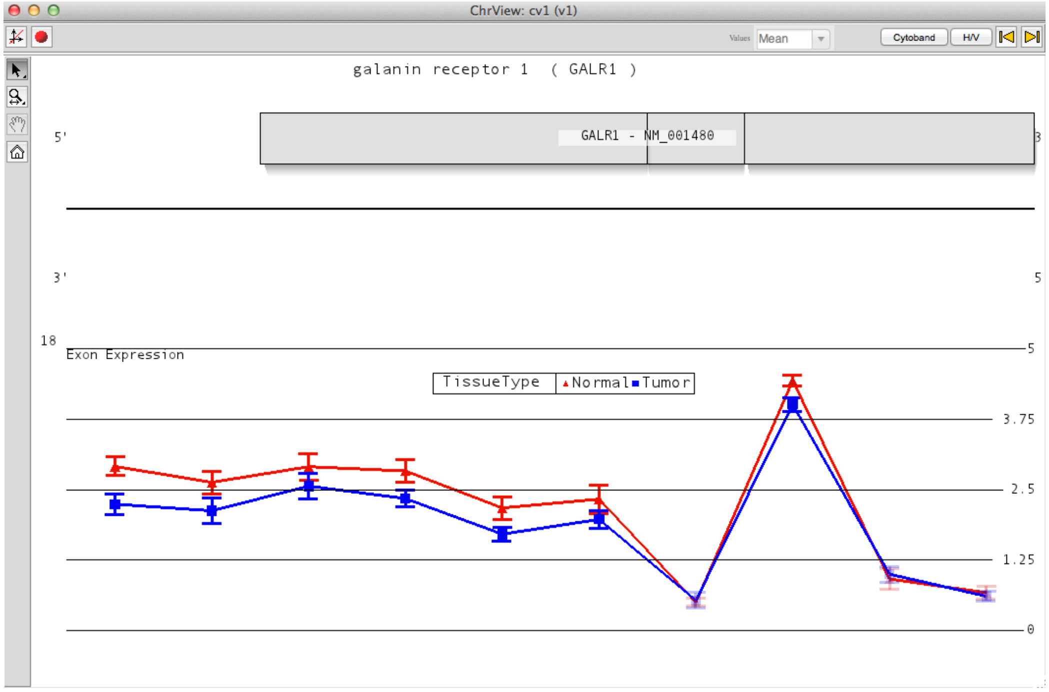
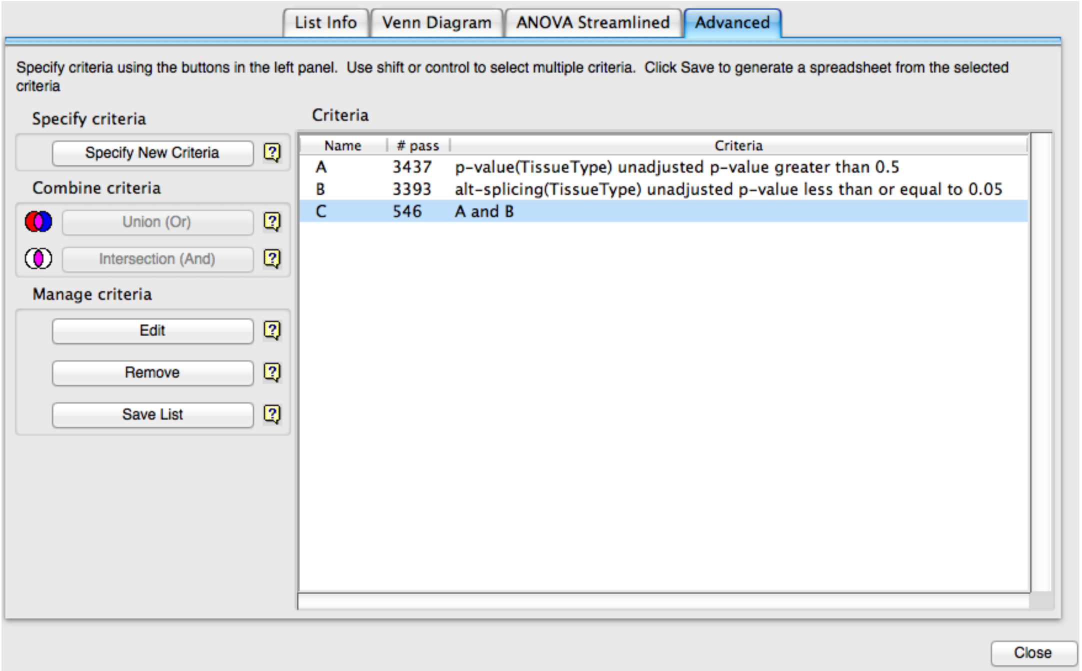


Figure 16: Example of differentially expressed but not alternatively spliced genes

The parallel nature of the expression shows the GALR1 gene being expressed higher in Normal tissue than Tumor tissue, but having the same expression profile across all exons.

Case #2 – Finding genes that are alternatively spliced, but NOT differentially expressed. (This is often described as the "cassette exon" use case).

- Follow the same steps as described in Case #1 but choose to use a **p-value(Tissue Type) > 0.5** and an **alt-splicing(TissueType) <= 0.05** (Figure 17)



Specify criteria using the buttons in the left panel. Use shift or control to select multiple criteria. Click Save to generate a spreadsheet from the selected criteria

Name	# pass	Criteria
A	3437	p-value(TissueType) unadjusted p-value greater than 0.5
B	3393	alt-splicing(TissueType) unadjusted p-value less than or equal to 0.05
C	546	A and B

Figure 17: List creation for Case #2, genes with alternative splicing but without differential expression

And example of a gene in this list can be seen in Figure 18.

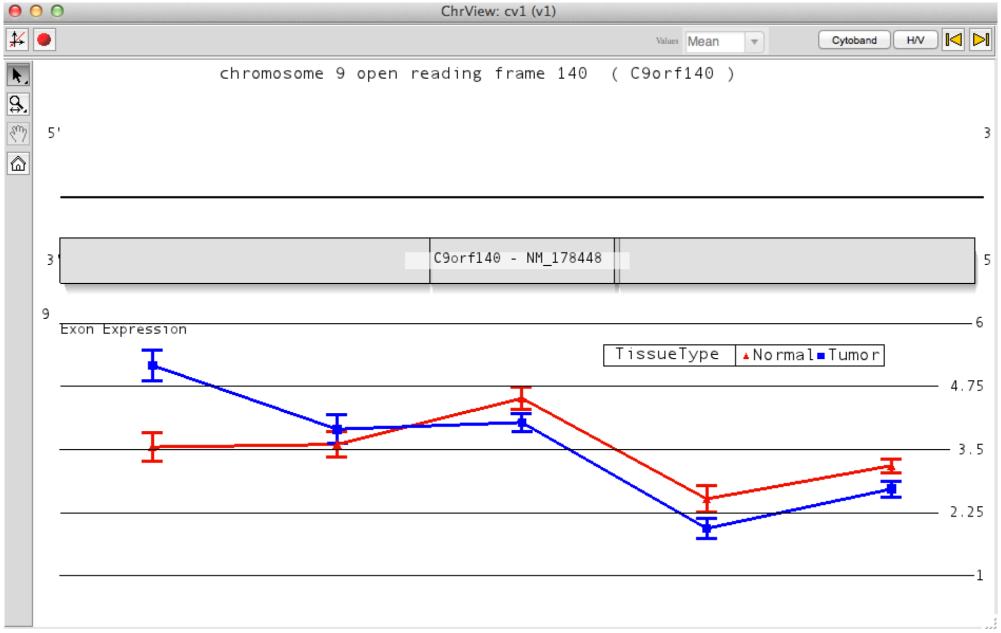


Figure 18: An example of a gene, which is not differentially expressed but is alternatively spliced

Here the gene follows a very different pattern between the two tissue types; the most obvious being that the three exons at the 5’ end of the gene are expressed in greater amounts in Normal tissue, while the exon at the 3’ end is expressed at a much greater level in the Tumor tissue. It should also be noted that this is consistent across both probes in the exon.

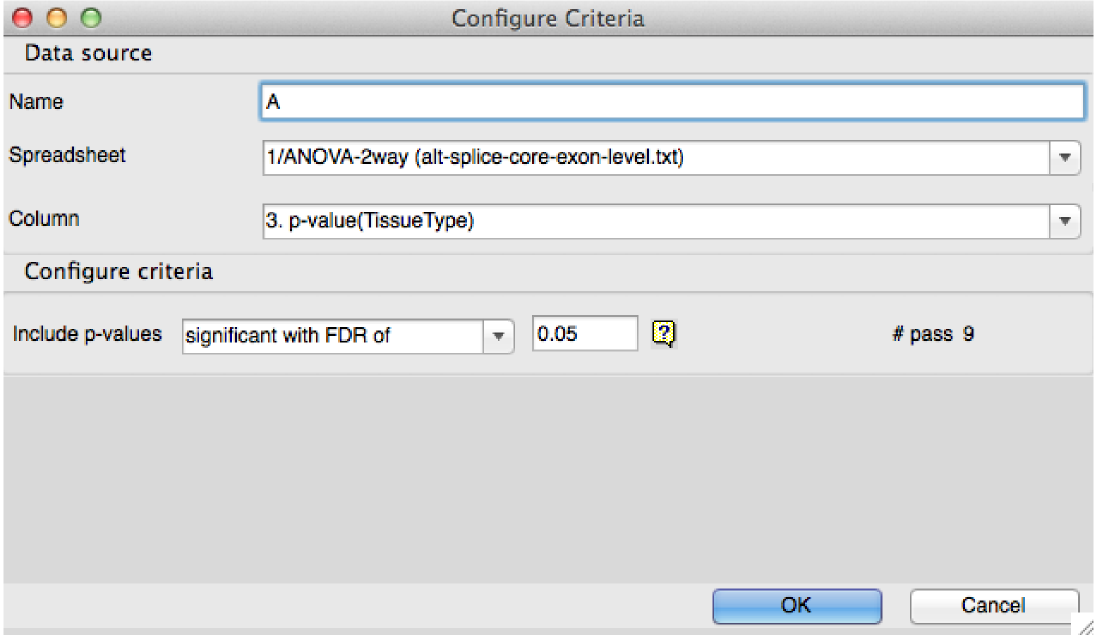
Advantage of exon analysis

- A 3' based expression strategy would detect this gene as increased in tumor relative to normal, due to missing the vital differences at the 5’end.
- A whole transcript approach might miss this due to the compensating increase in normal relative to tumor in the 5' end.
- Only a detailed exon analysis properly enlightens the changes in exon-specific gene expression within this transcript cluster.

Case #3 – Finding differentially expressed exons between tumor and normal

The alternative splicing ANOVA outputs exon level information in addition to gene level information. For this reason, Partek is not limited to detecting differentially expressed genes and can also detect differentially expressed exons which can potentially be used as biomarkers.

- Open the **ANOVA 2-way** sheet and open the list creation dialogue by pressing **Create Gene List** under the *Alternative Splicing Analysis* section of the *Exon* workflow (Figure 19)
- Change the *Include p-values* to **Significant with FDR of** and leave the value at **0.05** and select **Create** (Figure 19)



Data source

Name: A

Spreadsheet: 1/ANOVA-2way (alt-splice-core-exon-level.txt)

Column: 3. p-value(TissueType)

Configure criteria

Include p-values: significant with FDR of 0.05

pass 9

Figure 19: List creation for exons to be used as bio-markers

- Visualize the output by right pressing on a row header and choosing **dot plot (Orig. Data)** from the contextual menu (Figure 20)

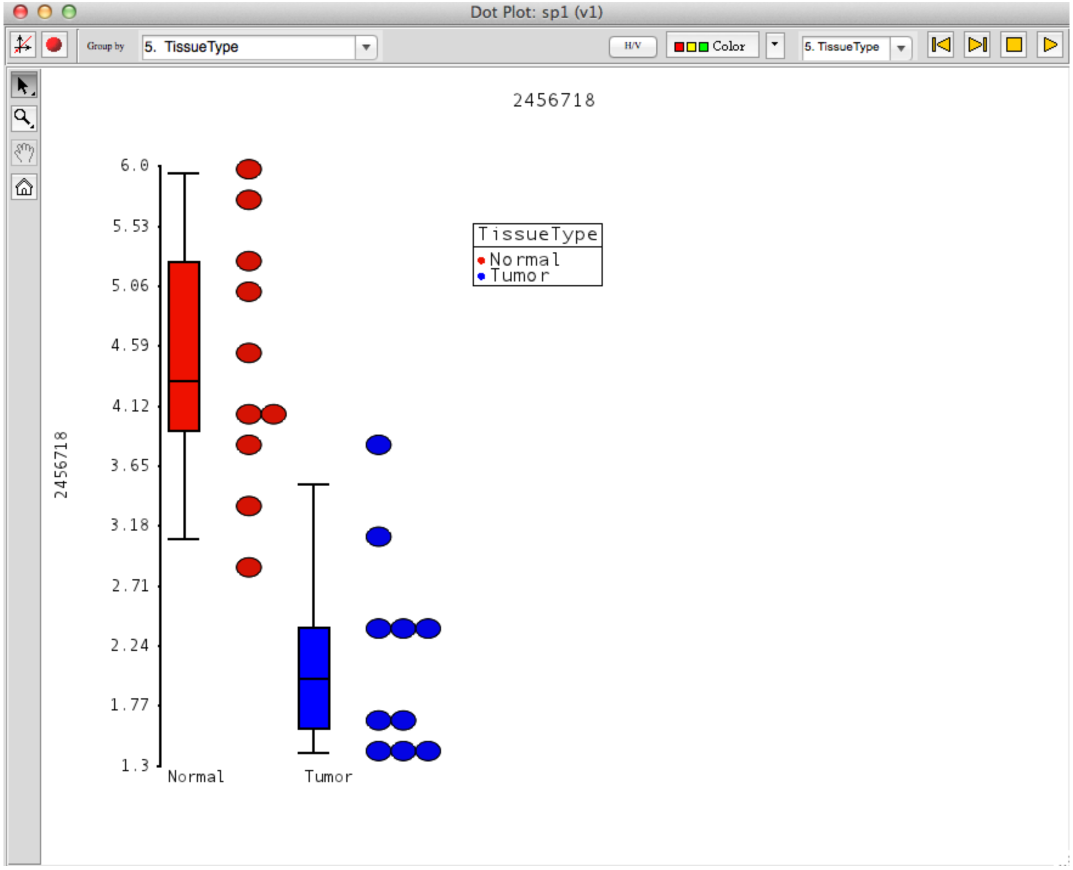


Figure 20: Differentially expressed exon between tissue types

Case #4 – Finding genes with the highest probability of alternative splicing between tumor and normal.

You can also find genes that have the highest probability of alternative splicing between tumor and normal, irrespective of differential expression.

- Rather than create a list, sort the alt-splice result sheet by the *alt-splicing(TissueType)* column by right-clicking on the column header and selecting **Sort Ascending**
- The 13th ranked gene is shown here as it is easy to see the individual exons in genes with fewer probes (Figure 21).
- In contrast, invoke Gene View on the 3rd ranked (COL12A1).

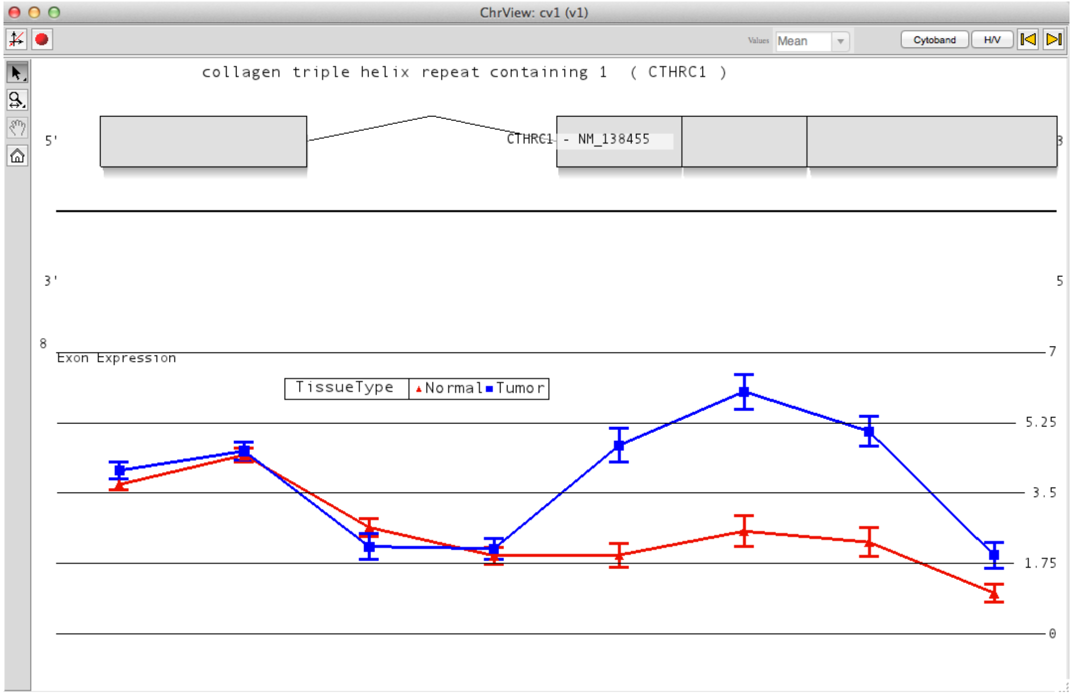


Figure 21: Example of a gene with a high likelihood of alternative splicing

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.