Methylation Array Analysis

NCI Workshop

December 13th, 2017

turning data into discovery®

Eric Seiser

Field Application Scientist

Partek® Genomic Suite™ Main Dialog

Analytical spreadsheet: Central repository of data

- No limitation on number of rows or columns
- Rows represent observations of interest (experiments, samples, chips)
- Columns represent measures of the observations (variables, features, genes,)

Menu bar: Execute commands from a graphical user interface

- When spreadsheet is empty, most of the menu items are not displayed

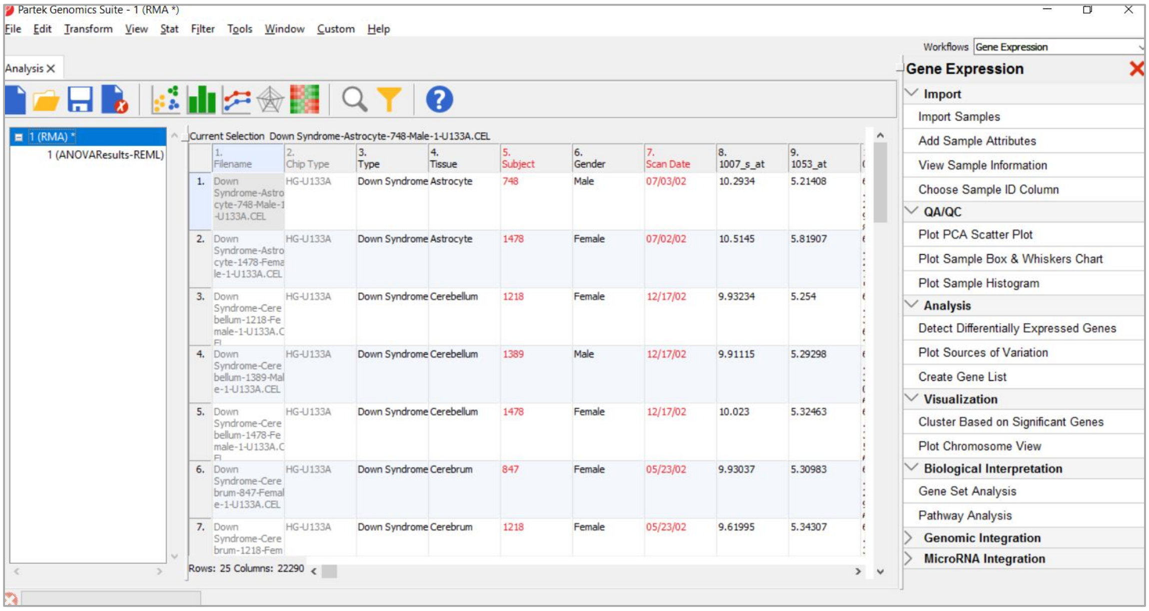
Tool bar: Accelerator buttons allow quick access to commonly used commands

Spreadsheet hierarchy: Open multiple datasets and see the hierarchy

- Original spreadsheet: parent
- Result spreadsheet: child

Active spreadsheet: The active spreadsheet is shown highlighted in blue, and the spreadsheet name and associated file name are shown at the top of the dialog

Workflow: Used to guide you through a typical analysis of a specific assay



Current Selection Down Syndrome-Astrocyte-748-Male-1-U133A.CEL

	1. Filename	2. Chip Type	3. Type	4. Tissue	5. Subject	6. Gender	7. Scan Date	8. 1007_s_at	9. 1053_at
1.	Down Syndrome-Astrocyte-748-Male-1-U133A.CEL	HG-U133A	Down Syndrome	Astrocyte	748	Male	07/03/02	10.2934	5.21408
2.	Down Syndrome-Astrocyte-1478-Female-1-U133A.CEL	HG-U133A	Down Syndrome	Astrocyte	1478	Female	07/02/02	10.5145	5.81907
3.	Down Syndrome-Cerebellum-1218-Female-1-U133A.CEL	HG-U133A	Down Syndrome	Cerebellum	1218	Female	12/17/02	9.93234	5.254
4.	Down Syndrome-Cerebellum-1389-Male-1-U133A.CEL	HG-U133A	Down Syndrome	Cerebellum	1389	Male	12/17/02	9.91115	5.29298
5.	Down Syndrome-Cerebellum-1478-Female-1-U133A.CEL	HG-U133A	Down Syndrome	Cerebellum	1478	Female	12/17/02	10.023	5.32463
6.	Down Syndrome-Cerebrum-847-Female-1-U133A.CEL	HG-U133A	Down Syndrome	Cerebrum	847	Female	05/23/02	9.93037	5.30983
7.	Down Syndrome-Cerebrum-1218-Fem	HG-U133A	Down Syndrome	Cerebrum	1218	Female	05/23/02	9.61995	5.34307

Rows: 25 Columns: 22290

Notes:____

____

____

____

____

____

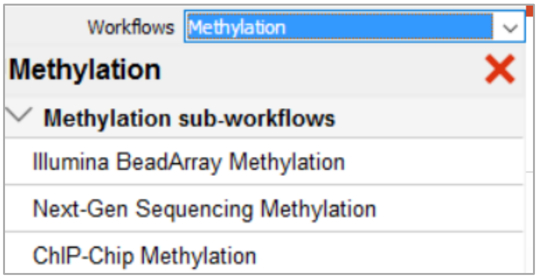
Tutorial Data Set

- The data set for the exercise is based on Gene Expression Omnibus GSE38240
- Download data from:
  - https://customer.partek.com/Methylation_training.zip
- Aryee *et al.* DNA methylation alterations exhibit intra-individual stability and inter-individual heterogeneity in prostate cancer metastases. Sci Transl Med 2013 Jan 23;5(169):169ra10.
- Prostate samples from
  - Normal individuals
  - Those diagnosed with prostate cancer
- Profiled using Illumina HumanMethylation450 BeadChip
- The goal of the exercise is to come up with a list of genes that show evidence of hyper- or hypo-methylation in tumor comparing to normal in promoter regions

Notes:____

Importing .iDAT Files

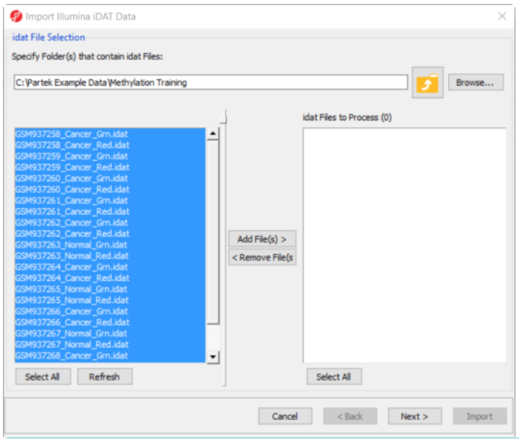
- Set the *Workflows* selector to **Methylation**
- Select **Illumina BeadArray Methylation**
- Select **Import Illumina Methylation Data**
- In the pop-up window, select **Import human methylation 450/850 .idat files**
- Browse to the folder, add the files to the right panel by clicking **Add Files,** there are 24 files to process
- Click **Next>**
- **Use Available Manifest File** option, name the output file as *GSE38240 data,* click **Import**
- The needed library file will be automatically downloaded
- The default is using functional normalization to generate β-values, which correspond to the percentage of methylation at each site
  - Ratio of methylated probe intensity over the overall intensity at each site.
- Each row of the spreadsheet corresponds to a single sample with the methylation probes on columns


Workflows: Methylation

Methylation

Methylation sub-workflows

- Illumina BeadArray Methylation
- Next-Gen Sequencing Methylation
- ChIP-Chip Methylation

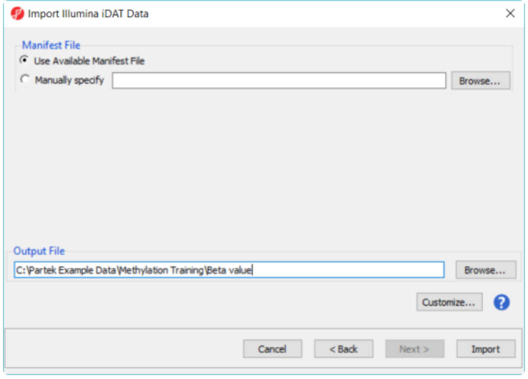

Import Illumina iDAT Data

idat File Selection

Specify Folder(s) that contain idat Files: C:\Partek Example Data\Methylation Training

```
GSM937258_Cancer_Grn.idat
GSM937258_Cancer_Red.idat
GSM937259_Cancer_Grn.idat
GSM937259_Cancer_Red.idat
GSM937260_Cancer_Grn.idat
GSM937260_Cancer_Red.idat
GSM937261_Cancer_Grn.idat
GSM937261_Cancer_Red.idat
GSM937262_Cancer_Grn.idat
GSM937262_Cancer_Red.idat
GSM937263_Normal_Grn.idat
GSM937263_Normal_Red.idat
GSM937264_Cancer_Grn.idat
GSM937264_Cancer_Red.idat
GSM937265_Normal_Grn.idat
GSM937265_Normal_Red.idat
GSM937266_Cancer_Grn.idat
GSM937266_Cancer_Red.idat
GSM937267_Normal_Grn.idat
GSM937267_Normal_Red.idat
GSM937268_Cancer_Grn.idat
```

idat Files to Process (0)


Import Illumina iDAT Data

Manifest File

● Use Available Manifest File

○ Manually specify

Output File: C:\Partek Example Data\Methylation Training\Beta value

Notes:____

____

____

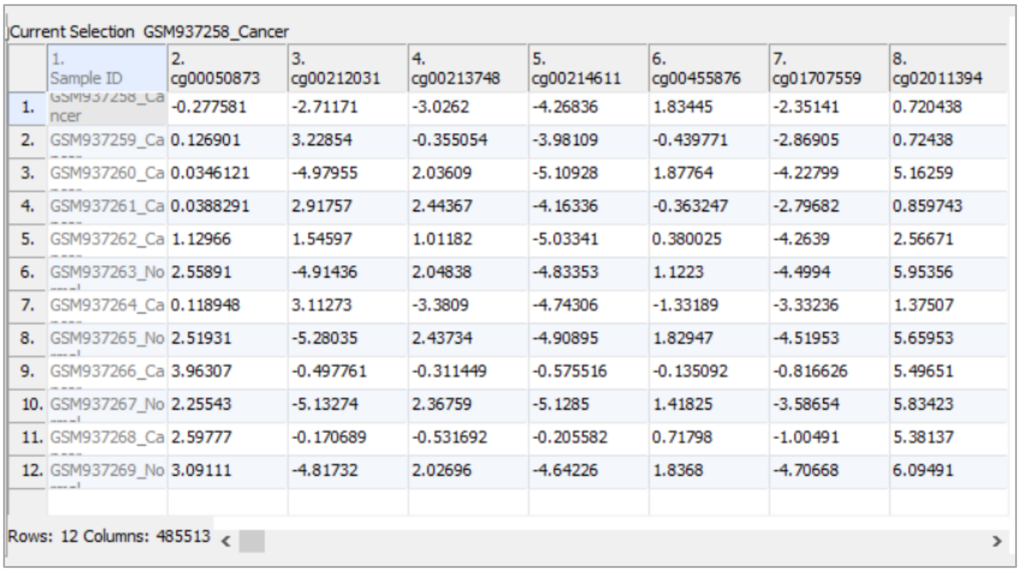
____

____

____

Convert Beta value to M Value

- Click **Convert Beta Value to M Values**
  - M-value = log2(β / (1 - β))
- The spreadsheet is overwritten with M value, click **Save**
- M value interpretation:
  - a M-value close to 0 indicates a similar intensity between the methylated and unmethylated probes, which means the CpG site is about half-methylated. Positive M-values mean that more molecules are methylated than unmethylated, while negative M-values mean that more molecules are unmethylated than methylated.
  - the M-value is more statistically valid for the differential analysis of methylation levels.



Current Selection GSM937258_Cancer

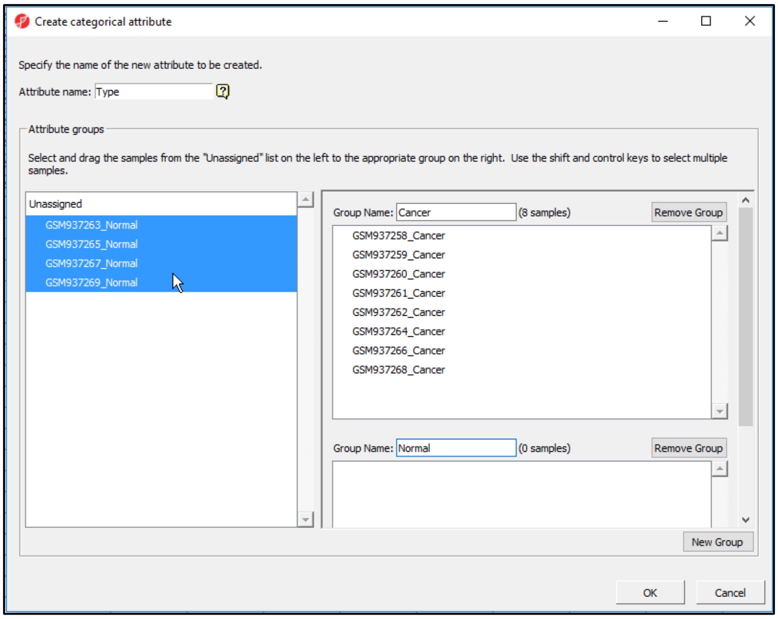
	1. Sample ID	2. cg00050873	3. cg00212031	4. cg00213748	5. cg00214611	6. cg00455876	7. cg01707559	8. cg02011394
1.	GSM937258_Cancer	-0.277581	-2.71171	-3.0262	-4.26836	1.83445	-2.35141	0.720438
2.	GSM937259_Ca	0.126901	3.22854	-0.355054	-3.98109	-0.439771	-2.86905	0.72438
3.	GSM937260_Ca	0.0346121	-4.97955	2.03609	-5.10928	1.87764	-4.22799	5.16259
4.	GSM937261_Ca	0.0388291	2.91757	2.44367	-4.16336	-0.363247	-2.79682	0.859743
5.	GSM937262_Ca	1.12966	1.54597	1.01182	-5.03341	0.380025	-4.2639	2.56671
6.	GSM937263_No	2.55891	-4.91436	2.04838	-4.83353	1.1223	-4.4994	5.95356
7.	GSM937264_Ca	0.118948	3.11273	-3.3809	-4.74306	-1.33189	-3.33236	1.37507
8.	GSM937265_No	2.51931	-5.28035	2.43734	-4.90895	1.82947	-4.51953	5.65953
9.	GSM937266_Ca	3.96307	-0.497761	-0.311449	-0.575516	-0.135092	-0.816626	5.49651
10.	GSM937267_No	2.25543	-5.13274	2.36759	-5.1285	1.41825	-3.58654	5.83423
11.	GSM937268_Ca	2.59777	-0.170689	-0.531692	-0.205582	0.71798	-1.00491	5.38137
12.	GSM937269_No	3.09111	-4.81732	2.02696	-4.64226	1.8368	-4.70668	6.09491

Rows: 12 Columns: 485513

Notes:

Annotating Samples

- Click **Add sample attributes > Add a categorical attribute > OK**
- In the dialog, set the *Attribute name* to **Type**, change the label *Group 1* to **Cancer** and *Group 2* to **Normal**
- **Ctrl-select** the samples labeled *Cancer* and **drag and drop** to the *Cancer* group
- **Drag and drop** the remaining samples to the *Normal* group. Click **OK**.
- When prompted to *Add another attribute*, click **No**
- Save spreadsheet with the new sample attribute, click **Yes**



Create categorical attribute

Specify the name of the new attribute to be created.

Attribute name: Type

Attribute groups

Select and drag the samples from the "Unassigned" list on the left to the appropriate group on the right. Use the shift and control keys to select multiple samples.

Unassigned
GSM937263_Normal
GSM937265_Normal
GSM937267_Normal
GSM937269_Normal

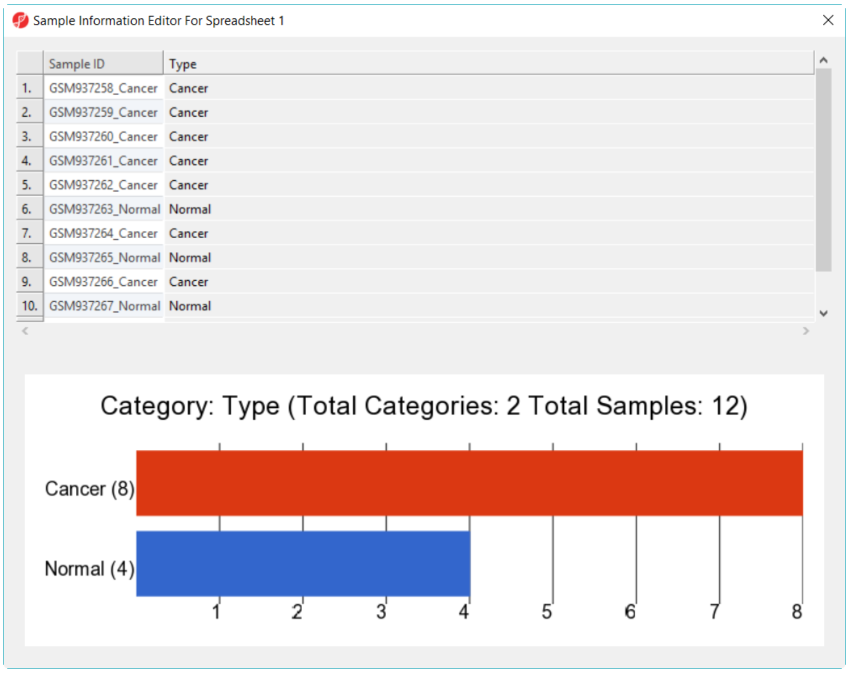
Group Name: Cancer (8 samples)
GSM937258_Cancer
GSM937259_Cancer
GSM937260_Cancer
GSM937261_Cancer
GSM937262_Cancer
GSM937264_Cancer
GSM937266_Cancer
GSM937268_Cancer

Group Name: Normal (0 samples)

Notes:

View Sample Information

- Click **View Sample Information** on the workflow
- There are 8 Cancer samples, 4 Normal samples
- Choose Sample ID column: use the default, it has to be unique ID for each sample


Sample Information Editor For Spreadsheet 1

	Sample ID	Type
1.	GSM937258_Cancer	Cancer
2.	GSM937259_Cancer	Cancer
3.	GSM937260_Cancer	Cancer
4.	GSM937261_Cancer	Cancer
5.	GSM937262_Cancer	Cancer
6.	GSM937263_Normal	Normal
7.	GSM937264_Cancer	Cancer
8.	GSM937265_Normal	Normal
9.	GSM937266_Cancer	Cancer
10.	GSM937267_Normal	Normal

Category: Type (Total Categories: 2 Total Samples: 12)

Notes:

Exploratory Analysis

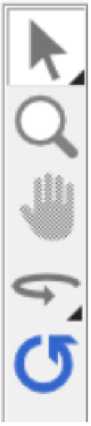
PCA scatter plot is one way to identify clustering patterns and outliers

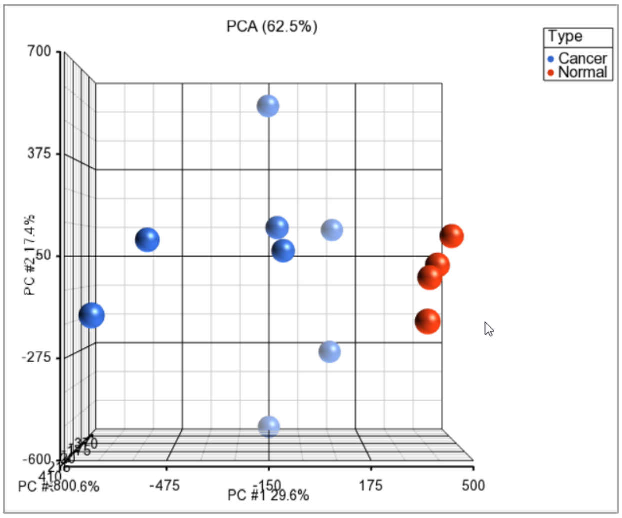
- Go to the *QA/QC* setion of the workflow > **Plot PCA Scatter Plot**

Notes

- Each point in the scatter plot corresponds to a specific row in the spreadsheet
- Points that are close together in the plot are similar in the original high-dimensional space
- Points that are far apart in the plot are dissimilar

- Click on **Plot Properties** to configure color
- Click on **Ellipsoid** to put the ellipsoid on each group
- Select mode:
    - left click to select; scroll mouse wheel to zoom; drag mouse wheel to rotate
    - right click after select a point to filter/clear filter





Notes:____

____

____

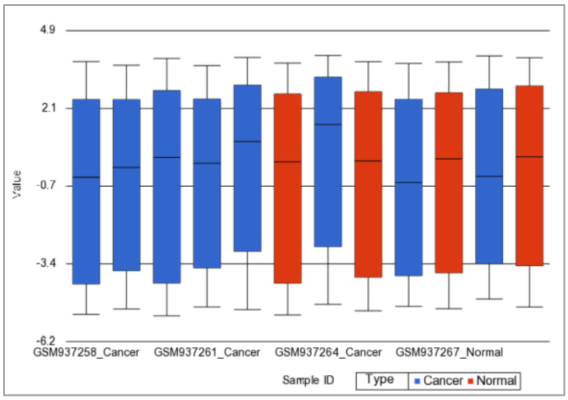
____

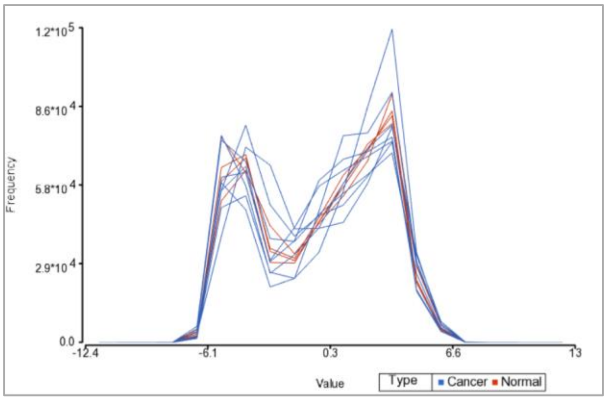
____

____

QA/QC – Histogram and Box plot

- Select **Plot Sample Box & Whiskers Chart**
  - Each box is a sample
  - Line inside the box is the median (2nd quartile)
  - Box represent the first and third quartiles
  - Whiskers represent 10th percentile and 90the percentile by default, can be configured
- Select **Plot sample histogram**
  - Each line is a sample
  - X-axis is the range of the values
  - Default 20 bins on X-axis, can be configured from **Plot Properties**

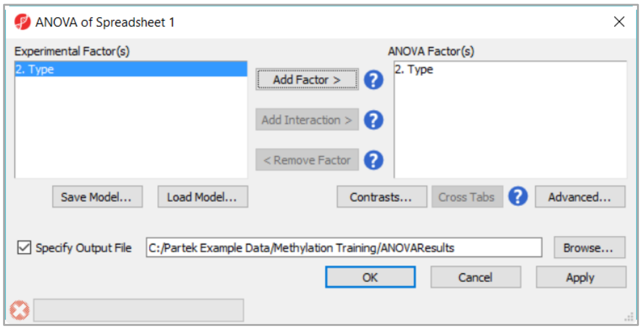




Notes:

Detecting Differentially Methylated Loci

- Go to **Analysis > Detect differential methylation**
- Select **2. Type** under *Experimental factor* and click the **Add Factor>** button
- Click the **Contrast** button



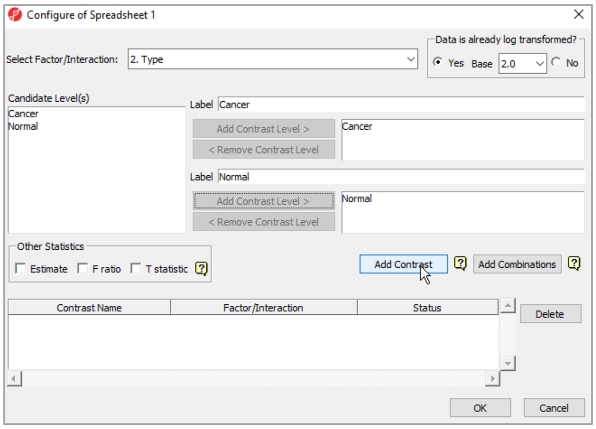
ANOVA of Spreadsheet 1

Experimental Factor(s): 2. Type

ANOVA Factor(s): 2. Type

☑ Specify Output File: C:/Partek Example Data/Methylation Training/ANOVAResults

- Choose **Yes** for *Data is already log transformed*
- Use **Add Contrast Level>** to move **Cancer** to *Group 1* and **Normal** to *Group 2*
- Select **Add Contrast** and then **OK.** In the ANOVA dialog also click **OK.**



Configure of Spreadsheet 1

Select Factor/Interaction: 2. Type

Data is already log transformed? ● Yes Base 2.0 ○ No

Candidate Level(s)
Cancer
Normal

Label: Cancer

Cancer

Label: Normal

Normal

Other Statistics: □ Estimate □ F ratio □ T statistic

Contrast Name	Factor/Interaction	Status

Notes:____

____

____

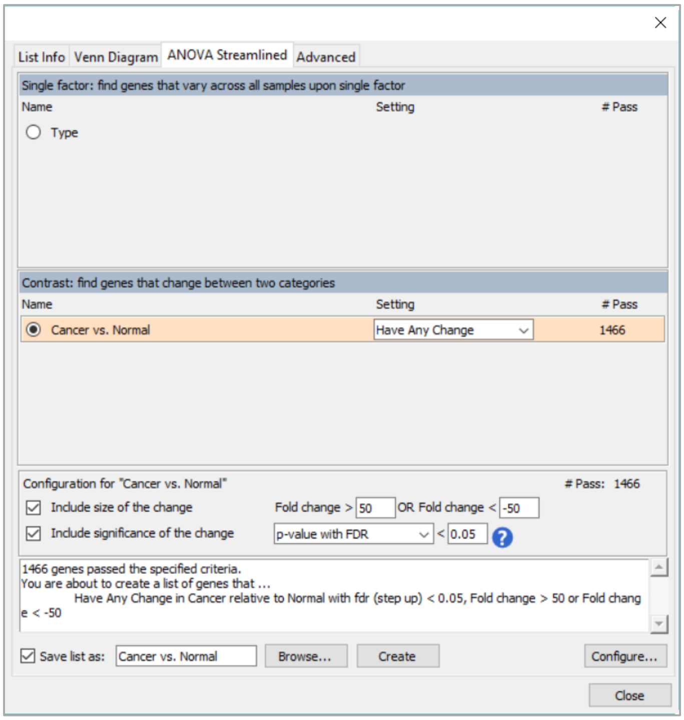
____

____

____

Creating marker list

- Click **Create marker list** from the workflow
- Select the **Cancer vs Normal** radio button
- Set the size of the *Fold Change* filter to *>50* and *<-50*
  - This selects markers that are either hyper- or hypo-methylated in Cancer comparing to normal
- Set the significance threshold to *p-value with FDR <0.05*
- Click the **Create** button to make a new list with these filtered markers



Single factor: find genes that vary across all samples upon single factor

Name	Setting	# Pass
○ Type		

Contrast: find genes that change between two categories

Name	Setting	# Pass
● Cancer vs. Normal	Have Any Change	1466

Configuration for "Cancer vs. Normal" # Pass: 1466

☑ Include size of the change Fold change > 50 OR Fold change < -50

☑ Include significance of the change p-value with FDR < 0.05

```
1466 genes passed the specified criteria.
You are about to create a list of genes that ...
        Have Any Change in Cancer relative to Normal with fdr (step up) < 0.05, Fold change > 50 or Fold chang
e < -50
```

☑ Save list as: Cancer vs. Normal

Notes:____

____

____

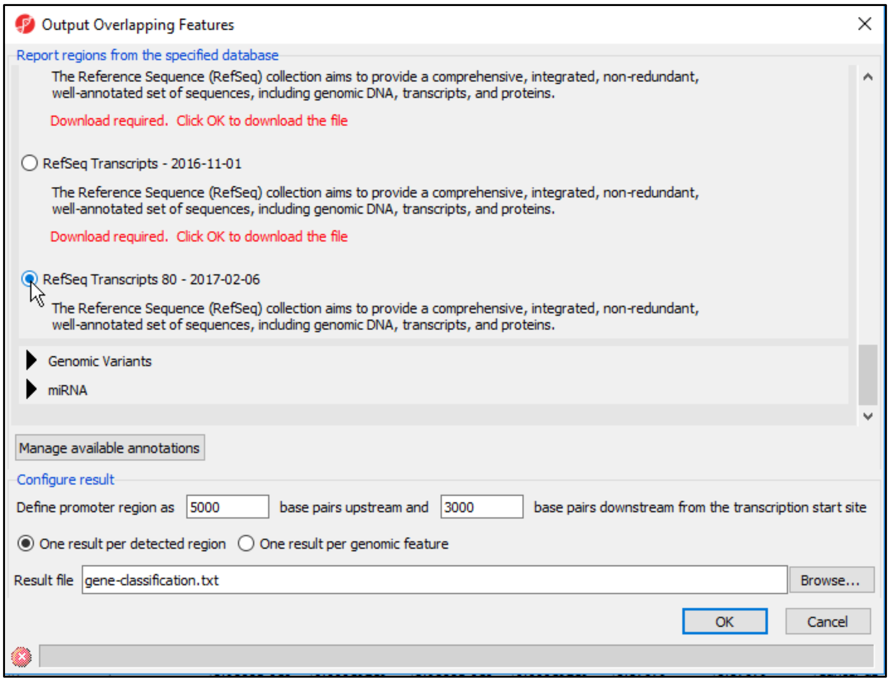
____

____

____

Classify regions by gene section

- With the *Cancer vs Normal* spreadsheet select **Classify regions by gene section** in the workflow
- Select the *RefSeq Transcripts 80 – 2017-02-06* radio button and click **OK**
- Using the default settings, the output spreadsheet contains each row is a probe overlap with a gene section
- One location can overlap with multiple transcripts



Output Overlapping Features

Report regions from the specified database

The Reference Sequence (RefSeq) collection aims to provide a comprehensive, integrated, non-redundant, well-annotated set of sequences, including genomic DNA, transcripts, and proteins.

Download required. Click OK to download the file

○ RefSeq Transcripts - 2016-11-01

The Reference Sequence (RefSeq) collection aims to provide a comprehensive, integrated, non-redundant, well-annotated set of sequences, including genomic DNA, transcripts, and proteins.

Download required. Click OK to download the file

● RefSeq Transcripts 80 - 2017-02-06

The Reference Sequence (RefSeq) collection aims to provide a comprehensive, integrated, non-redundant, well-annotated set of sequences, including genomic DNA, transcripts, and proteins.

Configure result

Define promoter region as 5000 base pairs upstream and 3000 base pairs downstream from the transcription start site

● One result per detected region ○ One result per genomic feature

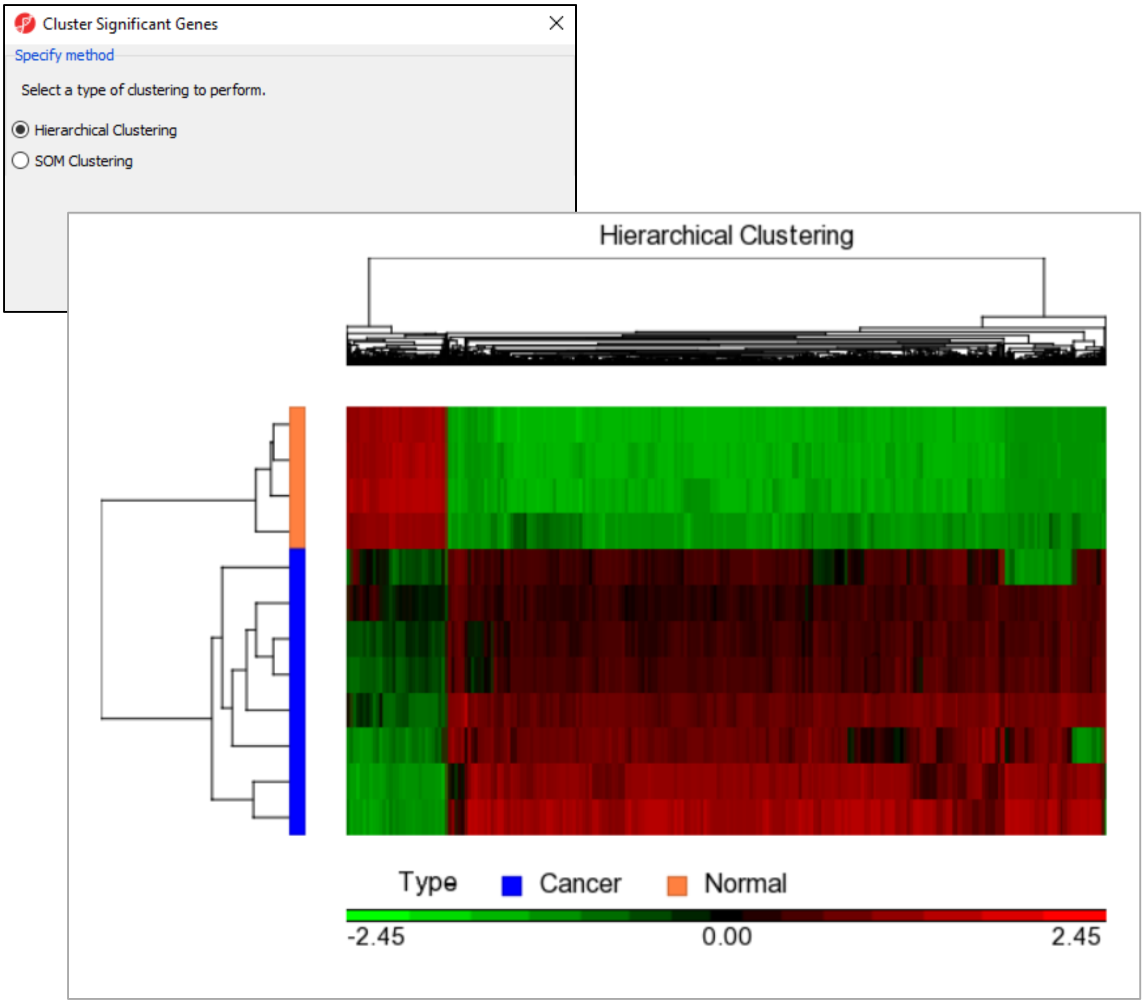
Result file gene-classification.txt

Notes:____

Cluster based on significant genes

- With the *Cancer vs Normal* spreadsheet select **Cluster based on significant genes** in the workflow

- Select **Hierarchical Clustering** and run the default settings



Cluster Significant Genes

Specify method

Select a type of clustering to perform.

● Hierarchical Clustering

○ SOM Clustering

Notes:

Hierarchical Clustering Configuration

Heatmap

- Click on the color square to change the heatmap color
- Change the orientation

Dendrograms

- Change the width/height of the dendrogram
- Color dendrogram

Rows

- Change the width of annotation
- Check show label
- Change color
- Add new annotation

Columns

- Label with column header or gene symbol

Save/Load: save or load configuration settings

Mode: mouse over, select, zoom, and flip

Notes:____

____

____

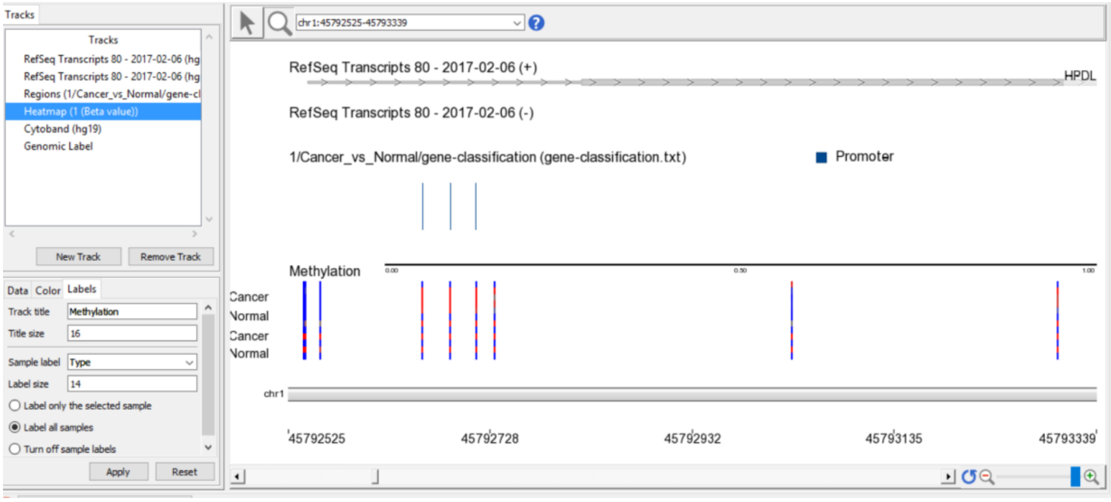
____

____

____

Viewing probes using chromosome view

- In the *gene-classification* spreadsheet, click the **interactive filter** button
- The interactive filter gives a graphical representation of the values within a dataset and makes it easy to select values to filter
- Select **7. Gene Section** from the drop-down menu
- **Right click** on the rightmost bar, representing *Promoter*
- Choose Plot Chromosome View to visualize the result
- Select each track to change the configuration:
  - Remove the track of Cancer vs Normal list
  - Select region track of gene classification, change the Separate bars by to **None**
- In zoom mode, click and drag a region to zoom in



Tracks

- RefSeq Transcripts 80 - 2017-02-06 (hg
- RefSeq Transcripts 80 - 2017-02-06 (hg
- Regions (1/Cancer_vs_Normal/gene-cl
- Heatmap (1 (Beta value))
- Cytoband (hg19)
- Genomic Label

Track title: Methylation

Title size: 16

Sample label: Type

Label size: 14

○ Label only the selected sample

● Label all samples

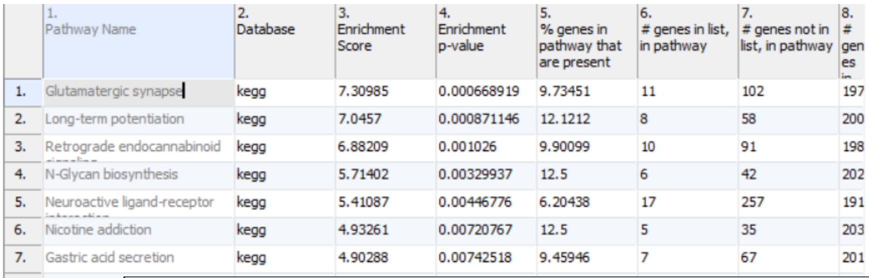
○ Turn off sample labels

chr1:45792525-45793339

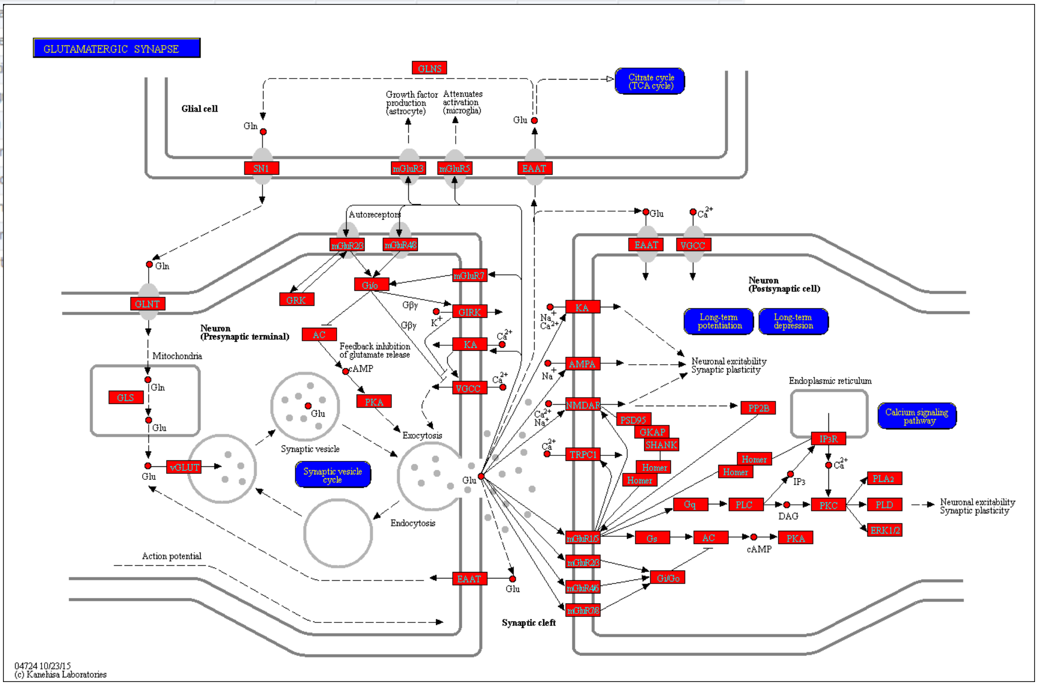
Notes:

Pathway Enrichment Analysis

- With the *gene-classification* spreadsheet select **Pathway analysis** in the workflow.
- Select the **Pathway Enrichment** radio button
- The result spreadsheet contains each row representing a pathway with enrichment score



	1. Pathway Name	2. Database	3. Enrichment Score	4. Enrichment p-value	5. % genes in pathway that are present	6. # genes in list, in pathway	7. # genes not in list, in pathway	8. # gen es in
1.	Glutamatergic synapse	kegg	7.30985	0.000668919	9.73451	11	102	197
2.	Long-term potentiation	kegg	7.0457	0.000871146	12.1212	8	58	200
3.	Retrograde endocannabinoid signaling	kegg	6.88209	0.001026	9.90099	10	91	198
4.	N-Glycan biosynthesis	kegg	5.71402	0.00329937	12.5	6	42	202
5.	Neuroactive ligand-receptor interaction	kegg	5.41087	0.00446776	6.20438	17	257	191
6.	Nicotine addiction	kegg	4.93261	0.00720767	12.5	5	35	203
7.	Gastric acid secretion	kegg	4.90288	0.00742518	9.45946	7	67	201
8.	Calcium signa							
9.	Long-term de							
10.	Renin secreti							
11.	TGF-beta sig							
12.	Gap junction							
13.	GnRH signali							
14.	Morphine add							
15.	Proteoglycar							
16.	MAPK signalir							
17.	Circadian ent							



Notes:____

Further Training

Self-learning

- Check out https://documentation.partek.com/display/PGS for documentation and additional resources
- Recorded webinars available on http://www.partek.com/webinars

Regional Technical Support

- Email: support@partek.com
- Phone: +1-314-878-2329

Notes:____