

# RNA-Seq Analysis in Partek<sup>®</sup> Flow<sup>®</sup>

HANDS-ON TRAINING

---

**National Institutes of Health  
November 2018**

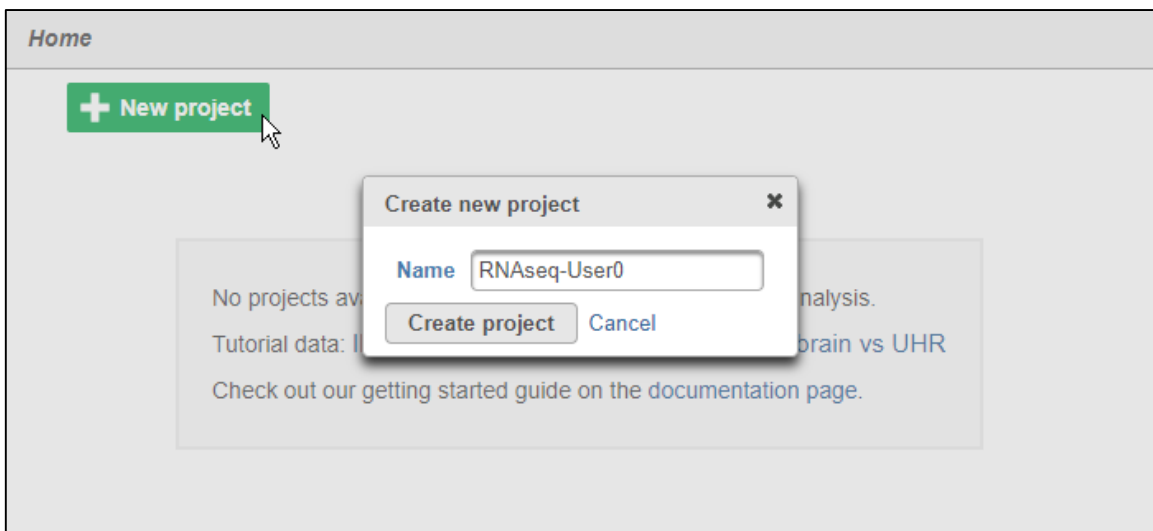


**Xiaowen Wang**  
*Field Application Specialist*  
*Partek Incorporated*  
*support@partek.com*

# Login and Project Set-up

---

- Open Google Chrome and enter: **training-server-url**
- Log in using the username and password given to you
- This will open to the Partek Flow homepage
- Click **New Project** and enter project name: RNAseq-[username]
- This will create a new project

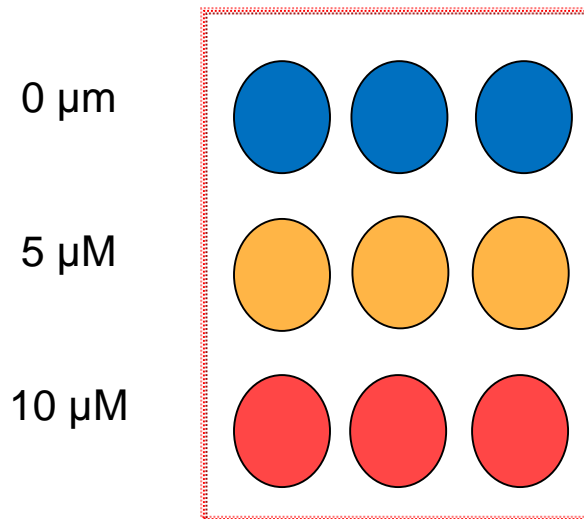


**Notes:** \_\_\_\_\_  
\_\_\_\_\_  
\_\_\_\_\_  
\_\_\_\_\_  
\_\_\_\_\_

# Experiment Description

---

- HT29 colon cancer cells exposed to 5-aza drug with 3 different doses
  - 0  $\mu\text{M}$  (Control)
  - 5  $\mu\text{M}$
  - 10  $\mu\text{M}$
- *Goal: Identify differentially expressed genes between different groups*
- mRNA purified and sequenced using Illumina HiSeq (Paired end reads)
- Xu et al. 2013 BMC Bioinformatics (PMID: 23902433)



**Notes:** \_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

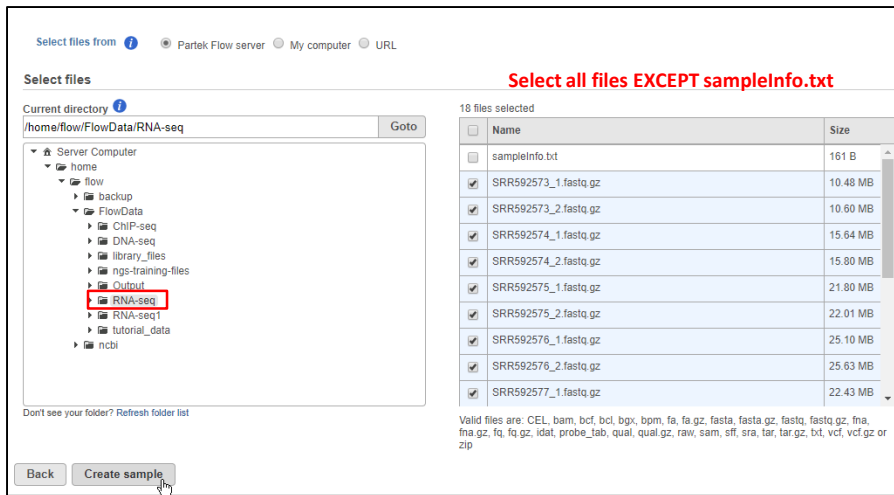
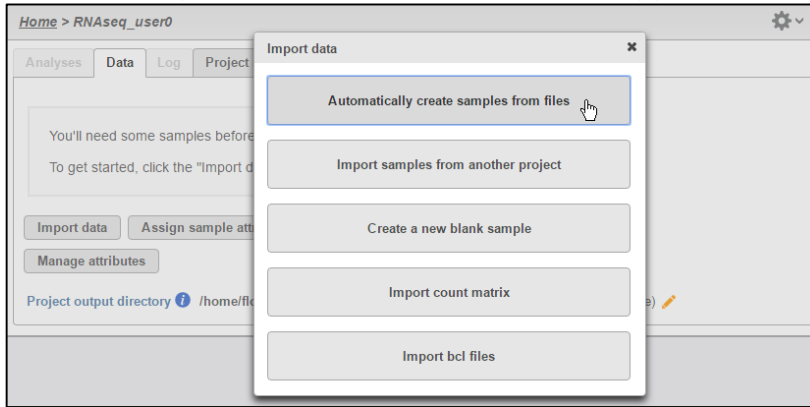
\_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

# Data Upload

- Creating a new project automatically opens up the **Data** tab
- To upload your data, click **Import data>Automatically create samples from files**
- Browse to `/home/flow/FlowData/RNA-seq`
- Select *all 18 fastq.gz files* and click **Create sample**
  - Partek Flow recognizes paired-end read data if tagged with (`_1` or `_R1`)



Notes:

---

---

---

---

---

---

# Sample Attribute Assignment

- Assign sample attributes using a tab-delimited text file
  - Contains table with ID in 1<sup>st</sup> column, followed by corresponding treatment groups
- Click **Assign sample attributes from a file**
- In the same folder, select *sampleInfo.txt*, click **Next**
- Click **Import**
- This will assign treatment groups to all samples

Home > RNAseq-User0 (Project owner)

Analyses | Data | Log | Project settings | Attachments

	Sample name	
1	SRR592573	
2	SRR592574	
3	SRR592575	
4	SRR592576	
5	SRR592577	
6	SRR592578	
7	SRR592579	
8	SRR592580	
9	SRR592581	

Show data files Download

Import data **Assign sample attributes from a file**

Add a system-wide attribute column Manage attributes Apply attributes by importing a file with information about your samples

Project output directory /home/flow/FlowData/project\_output/Project\_RNAseq-User0 (8.37 TB free)

Imported attributes that do not currently exist will create new Project-specific attributes.

Attribute name	Terms	Import	Attribute type
<input type="text" value="sample name"/>	SRR592573, SRR592574, SRR...	<input type="checkbox"/>	Categorical
<input type="text" value="Treatment"/>	0uM, 10uM, 5uM	<input checked="" type="checkbox"/>	Categorical

Show/hide file preview

**Notes:**

---



---



---



---



---

# Analyses Tab Overview

- Go to the **Analyses** tab
- Your first data node, the **Unaligned reads** node appears
  - *All data nodes are circles*
- Select the **Unaligned reads** data node and select **Pre-alignment QA/QC**
- Use the default settings and click **Finish**
- This will create a new task node in the **Analyses** tab
  - *All task nodes are rectangles*
- Clicking any node will bring up a **Context sensitive menu** on the right. Only the tasks that can be performed on that node will appear in this menu

The screenshot displays the 'Analyses' tab in the software interface. At the top, there are navigation tabs: 'Analyses', 'Data', 'Log', 'Project settings', and 'Attachments'. A 'List generator' button is located in the top right corner. The main workspace shows a workflow diagram with two nodes: 'Unaligned reads' (a circular data node) and 'Pre-alignment QA/QC' (a rectangular task node). A tooltip points to the 'Pre-alignment QA/QC' node, stating: 'Assess the quality of raw reads to decide whether trimming or filtering is necessary before alignment.' A context-sensitive menu is open on the right side of the 'Pre-alignment QA/QC' node, listing various actions: 'Unaligned reads', 'QA/QC', 'Pre-alignment QA/QC' (highlighted with a mouse cursor), 'ERCC', 'Filter contaminant (Bowtie 2)', 'Pre-alignment tools', 'Metagenomics', and 'Aligners'. Below the menu, there is a 'Context sensitive menu' section with the option 'Download data (351 MB)' and a download icon. At the bottom of the interface, there are links for 'Make a pipeline' and 'Import a pipeline', and a 'Project disk space' indicator showing a yellow bar.

**Notes:** \_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

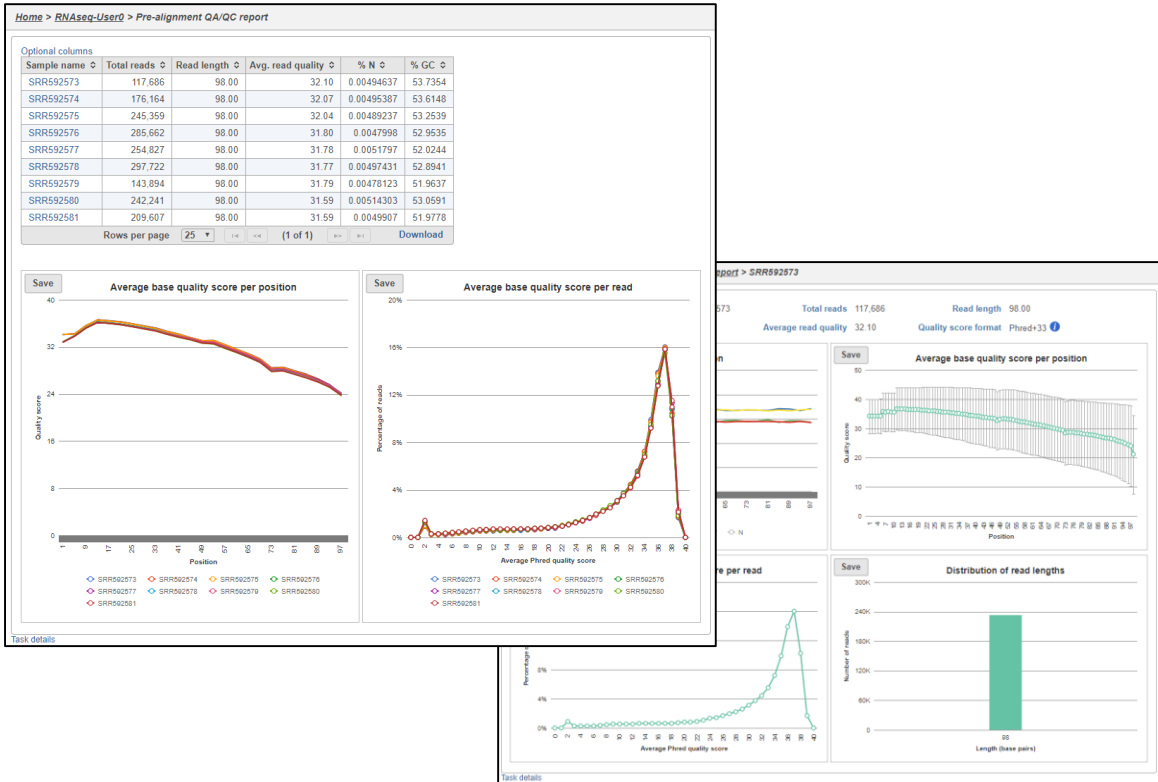
\_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

# Pre-alignment QA/QC

- Double-clicking on the **Pre-alignment QA/QC** node opens the task report
- Double-clicking each sample name also shows QA/QC results per sample



Quality score is  $-10\log_{10}\text{Prob}$

Phred Quality Score	Prob. of error	Base call accuracy
10	1/10	90%
20	1/100	99%
30	1/1000	99.9%
40	1/10000	99.99%

Notes:

# Pre-analysis Tools: Trim Bases

## Base trimming based on quality score

- Select **Unaligned reads** data node
- Click **Trim bases** from the **Pre-analysis tools** section in the toolbox
- Select **Trim based on: Quality score** with default settings and click **Finish**
- This will trim the reads at the 3' end with a Phred quality score less than 20
- This produces your 1<sup>st</sup> new data node, the **Trimmed reads** data node

*Tip:* Hover over any **i** to get additional information about a specific option

The screenshot shows the 'Trim bases' tool interface. The 'Trim based on' section has 'Quality score' selected. The 'Quality trimming' section has 'End min quality level (Phred)' set to 20 and 'Trim from end' set to '3-prime (right end)'. The 'Advanced options' section has 'Min read length' set to 25, 'Max N' set to 1%, and 'Quality encoding' set to 'Auto detect'. A tooltip is displayed over the 'Quality score' option, explaining that this mode scans the read from the 5' or 3' end for the first base at or above the specified Phred quality score. A diagram in the tooltip shows a quality score curve and a sequence 'A C G T T A C C A' with a red dashed line indicating the cutoff at the 7th position.

**Notes:**

---

---

---

---

---

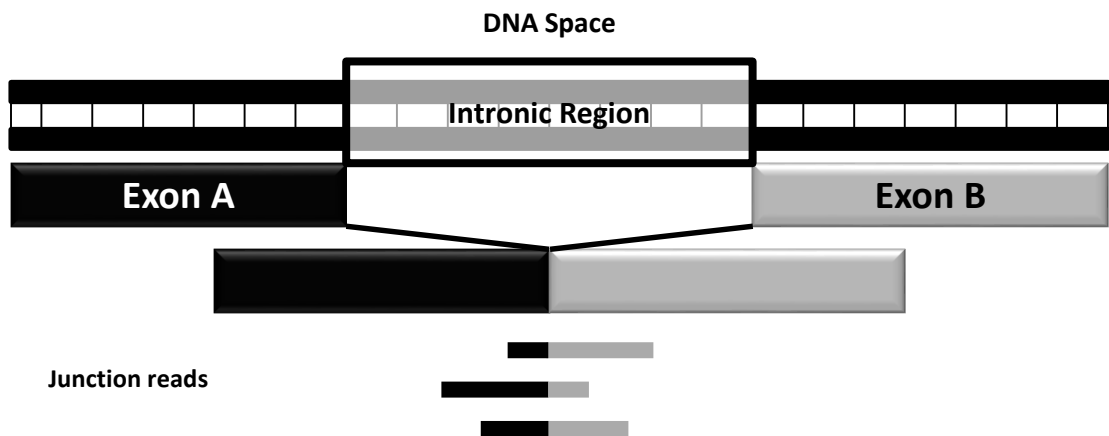
---



# Aligning RNA-Seq Data

---

- RNA-Seq data must be aligned using an aligner that supports junction reads
- A junction read is one that spans two exons
- STAR is one of several aligners in Flow that you can use
  - Others include TopHat and GSNAP



**Notes:** \_\_\_\_\_  
\_\_\_\_\_  
\_\_\_\_\_  
\_\_\_\_\_  
\_\_\_\_\_

# Alignment

---

- Select the **Trimmed reads** data node
- Click **STAR** from the **Aligners** section of the menu
- Select STAR index:
  - Genome build: **Homo sapiens (human) - hg19\_chr22**
  - Index: **Whole genome**
- Use the default options, click **Finish**

*Home > RNAseq-User0 > STAR*

### Select STAR 2.4.1d index

Assembly

Aligner index

### Alignment options

Generate unaligned reads

### Advanced options

Option set  [Configure](#)

**Notes:** \_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

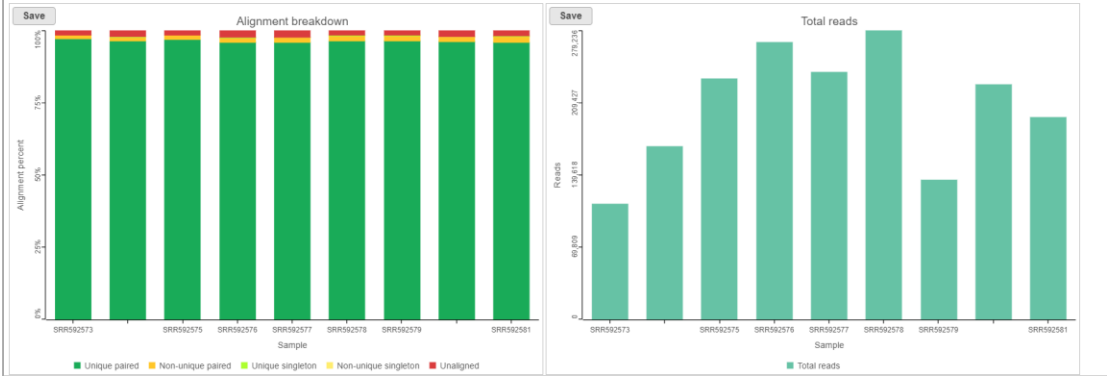
\_\_\_\_\_

# Post-alignment QA/QC

- Perform Post-alignment QA/QC to assess the quality of the alignment task
- Select **Aligned reads** data node
- Click **Post-alignment QA/QC** from the **QA/QC** section of the menu
- Use default settings and click **Finish**
- Click on a sample name to get QA/QC results for that sample name

Sample name	Total reads	Total alignments	Aligned	Unique singleton	Unique paired	Non-unique paired	Non-unique singleton	Coverage	Avg. coverage depth	Avg. length	Avg. quality	%GC
SRR592573	111,727	224,382	98.44%	14%	96.98%	1.31%	0%	8.75%	4.62	92.93	34.40	53.37%
SRR592574	106,846	333,014	97.80%	16%	96.34%	1.30%	0%	10.49%	5.71	92.84	34.40	53.24%
SRR592575	232,251	466,885	98.43%	15%	96.95%	1.32%	0%	14.17%	5.92	92.78	34.43	52.89%
SRR592576	267,646	536,699	97.63%	17%	95.75%	1.71%	0%	17.20%	5.60	92.69	34.35	52.60%
SRR592577	238,647	478,459	97.68%	17%	95.85%	1.66%	0%	15.91%	5.40	92.75	34.39	51.66%
SRR592578	279,236	564,099	98.29%	19%	96.36%	1.73%	0%	17.78%	5.70	92.72	34.35	52.53%
SRR592579	134,913	272,022	98.28%	17%	96.44%	1.67%	0%	12.84%	3.80	92.70	34.38	51.63%
SRR592580	226,772	457,002	97.96%	20%	96.01%	1.75%	0%	14.32%	5.71	92.42	34.30	52.70%
SRR592581	195,532	394,633	98.01%	26%	95.87%	1.87%	0.1%	14.68%	4.80	92.27	34.39	51.65%

These samples were aligned by STAR. See STAR options.  
 Bases were inserted and deleted during alignment.  
 Read quality scores are in the Phred-33 format.



## Notes:

---



---



---



---



---



---

# Quantification to Annotation Model

- Mapping aligned reads to a database of known transcripts
  - This method can be used with any gene or feature annotation
- Select **Aligned reads** data node
- Click **Quantify to annotation model (Partek E/M)** from the **Quantification** section of the menu
- Select **RefSeq** as the Annotation model and click **Finish**
  - By default, features with total number of reads less than 10 will be filtered out

Home > RNAseq-user0 > Quantify to annotation model (Partek E/M)

Select Annotation file

Assembly Homo sapiens (human) - hg19-chr22only

Gene/feature annotation refseq

Quantification options

Strict paired-end compatibility

Require junction reads to match introns

Minimum read overlap with feature

Percent of read length 100

Number of bases 50

Min reads 10

Advanced options

Strand specificity No

Unexplained regions

Report unexplained regions

Min reads for unexplained region 30

Back Finish

Notes:

---

---

---

---

---

---

# Viewing Quantification Results

- Since the RefSeq annotation has both *gene-* and *transcript-level* information, this task will generate 2 data nodes:
  - Gene counts
  - Transcript counts
- To view the results, double-click the **Gene counts** data node
- Data at this level can be downloaded as text file containing count matrix



**Notes:**

---



---



---



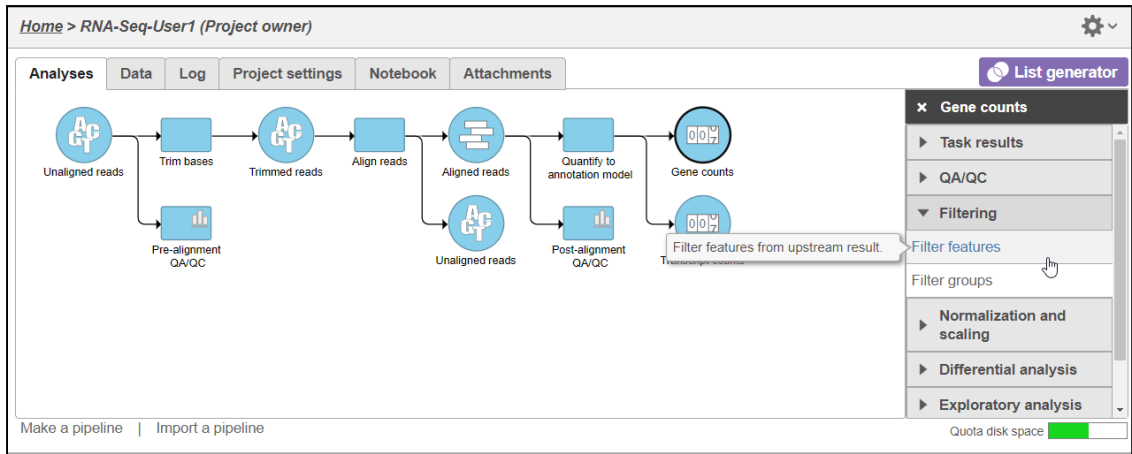
---



---

# Filter Features

- Low expression genes maybe indistinguishable from noise, will decrease the sensitivity of DEG detection
- To filter out low expression at the gene level, select the **Gene counts** data node and click **Filter features** under the **Filtering** portion of the menu
- Choose **Filter exclude features if Maximum<=10**
- Click **Finish**



Noise reduction filter

Exclude features where maximum <= 10

Statistics based filter

Filter features by  Counts  Percentiles

Keep the top 100.0 features with highest variance

Notes:

---

---

---

---

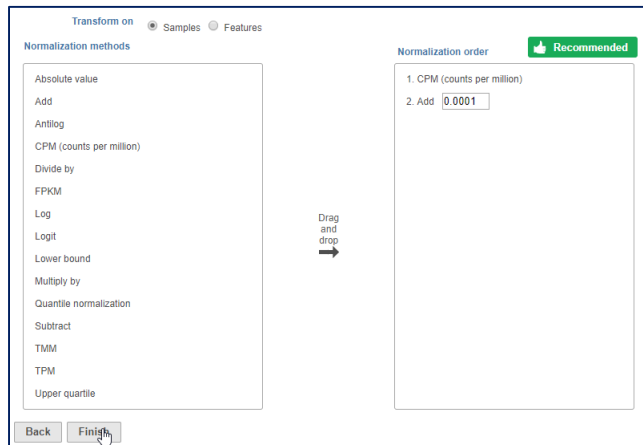
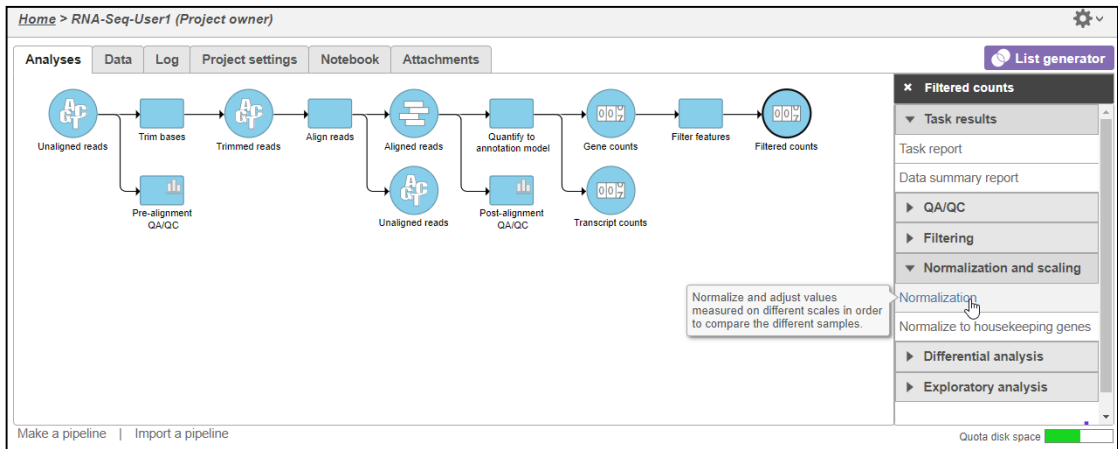
---

---

---

# Normalize Counts

- Data must be normalized before differential expression analysis
- Select the filtered gene count node and click **Normalize counts** under the **Normalization and scaling** portion of the menu
- Click the **Recommended** button and select **Finish**



**Notes:** \_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

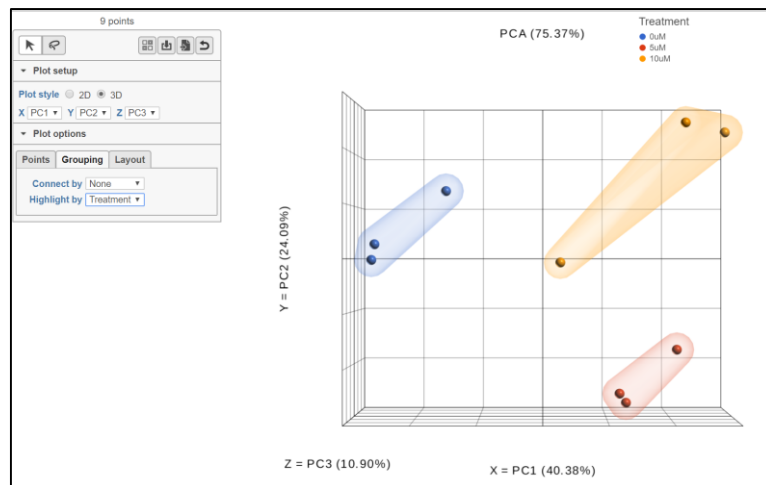
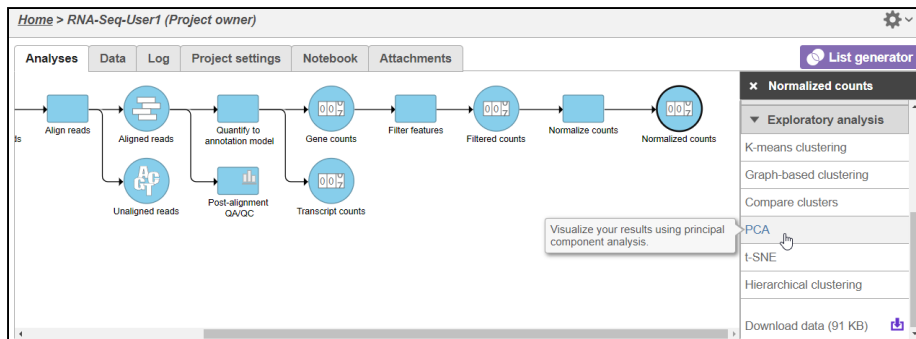
\_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

# Principal Components Analysis

- The principal components analysis (PCA) scatter plot allows you to assess relatedness between samples and identify outliers
- This can only be performed on quantified data
- To create the PCA plot, select the **Normalized counts** data node, click **PCA** under the **Exploratory analysis** portion of the menu, use default settings and select **Finish**



Notes:

---

---

---

---

---

---

---

---



# Differential Expression Analysis

- Select the **Normalized counts** data node
- Click **GSA** from the **Differential analysis** section of the menu
- Select **Treatment** as an attribute to include in statistical test and click **Next**
- Setup the following comparisons and click the *Add comparison* button
  - **5uM vs 0uM**
  - **10uM vs 0uM**
- Click **Finish**

The screenshot shows the RNA-Seq analysis pipeline interface. The main workflow consists of the following steps: Reads → Aligned reads → Quantify to annotation model → Gene counts → Filter features → Filtered counts → Normalize counts → Normalized counts → PCA. A tooltip for the 'GSA' step reads: "Identify differentially expressed features with Partek GSA algorithm by applying multiple statistical models to each individual gene in order to account for each gene's varying response to different experimental factors, and differing data distributions." The right-hand menu is open to the 'Normalized counts' node, showing options for 'Differential analysis' including GSA, ANOVA, and Non-parametric ANOVA. Below the main interface is a 'Comparison selector' dialog box.

**Comparison selector**

**Treatment**

0uM    10uM

5uM

10uM

vs.

**Treatment**

0uM    0uM

5uM

10uM

**Add comparison**

	Treatment		Treatment	
1	5uM	vs.	0uM	✗
2	10uM	vs.	0uM	✗


**Advanced options**

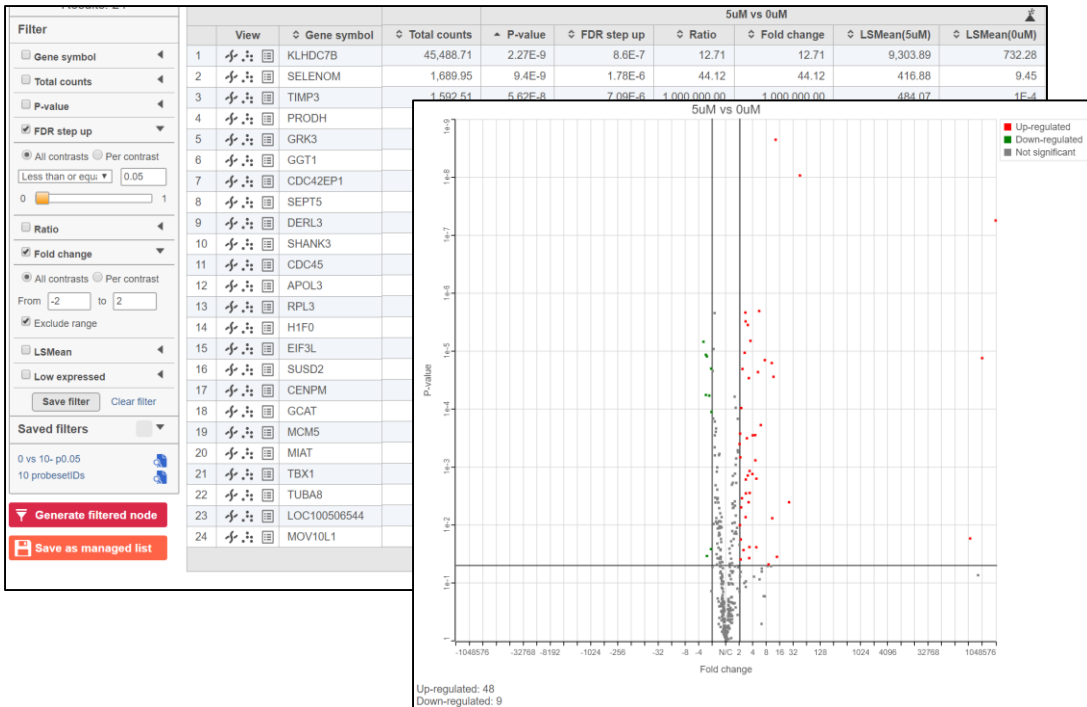
Option set: -- Default --    [Configure](#)

**Back**    **Finish**

**Notes:**

# Creating a Filtered Gene List

- Select **Feature list** data node and then click **Task report** in the toolbox
- To get a sense of how to filter list, view the *Volcano plot* by clicking 
- Under the **Gene list** section, on the **Filter** panel select:
  - **FDR step up**, then select **All contrasts** and set it to Less than or equal to 0.05
  - **Fold-change**, then select **All contrasts** and set it to From **-2 to 2**, with **Exclude range** selected
- At the bottom of the table, click **Generate filtered node**



**Notes:**

---



---



---






---



---

# Viewing Gene/Transcript Level Results

- Select **Feature List** data node and then click **Task report** in the toolbox
- On the table, under the **View** column, select
  -  to view the Dot plot
  -  to see the region in Chromosome View
  -  to see additional information about the statistical results



**Notes:**

---



---



---



---



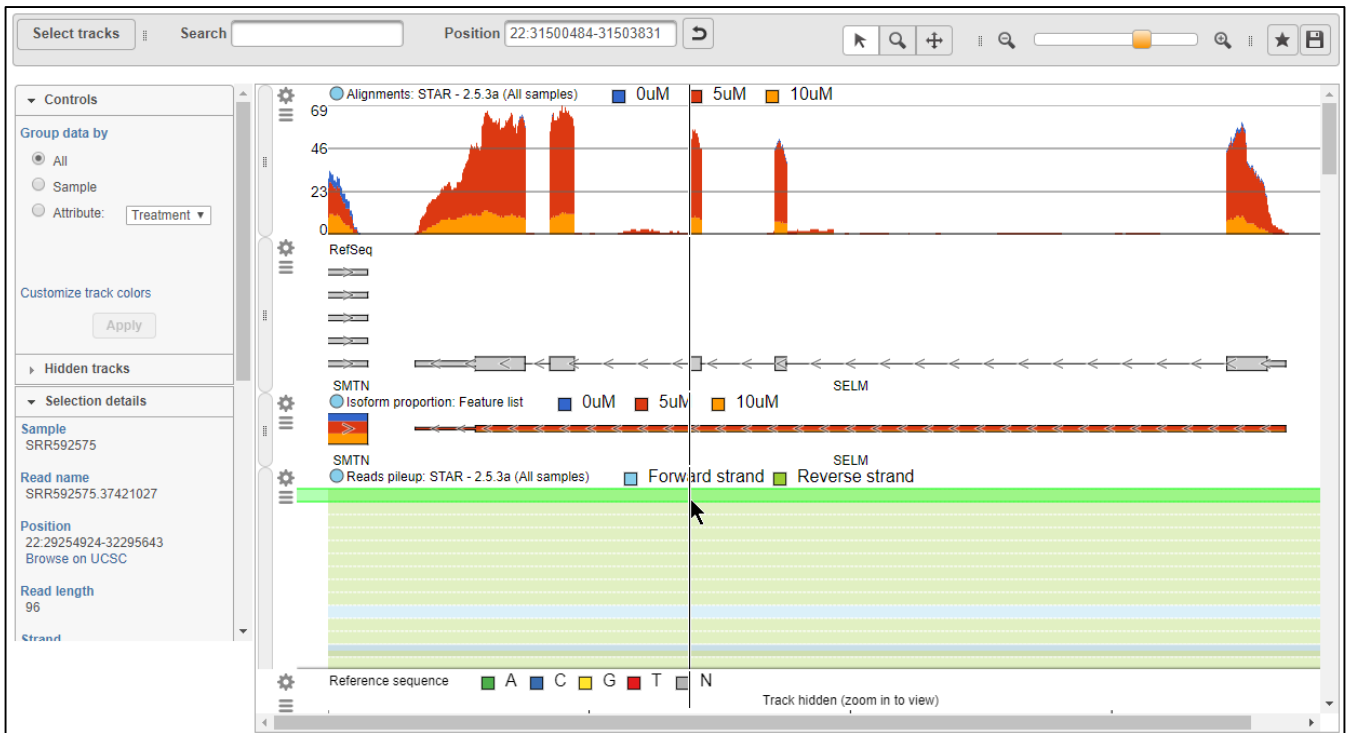
---



---

# Chromosome Viewer

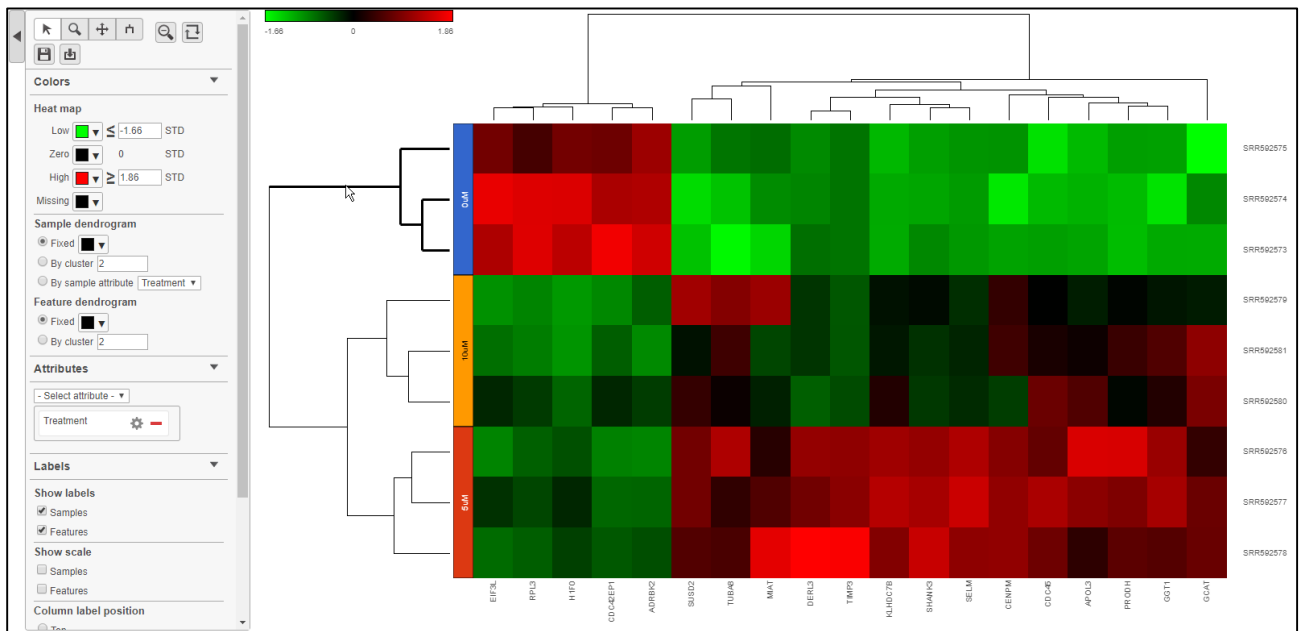
- **Select tracks** allows you to select different annotations or datasets to view together
- Sample grouping, color and transcript labeling can be edited in the **Controls** panel
- Search for any gene using the **Search** box
- Navigate to a genomic coordinate using the **Position** box
- Change and pin any displayed tracks using **Track order**
- Select any read in the reads pileup track to display additional information about the read



**Notes:** \_\_\_\_\_  
\_\_\_\_\_  
\_\_\_\_\_  
\_\_\_\_\_  
\_\_\_\_\_

# Hierarchical Clustering

- Select any **Feature list** data node to perform clustering on that list of genes/transcripts
- For this training, select the **Feature list** produced after filtering
- Click **Hierarchical clustering** from the **Exploratory analysis** section of the menu
- Click **Finish** to run hierarchical clustering with default settings
- Select the **Hierarchical clustering** task node and click on **Task Report**



Notes: \_\_\_\_\_

\_\_\_\_\_


\_\_\_\_\_





\_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

# Enrichment Analysis

- Perform gene set enrichment analysis using filtered list of genes
- Select **Feature List** data node resulting from the Filtered gene analysis task
- Select **Enrichment analysis** from the **Biological interpretation** section of the menu
- Select **GO** (Gene Ontology) as Gene set annotation and then click **Finish**
- Select the **Enrichment** task node and click on **Task Report**
- Select  to get additional information about each specific pathway

Gene set	Description	Enrichment score	P-value	Genes in list	Genes not in list	
GO:1901605	alpha-amino acid metabolic process	9.65	6.42E-5	3	0	
GO:0034622	cellular macromolecular complex assembly	9.04	1.18E-4	5	9	
GO:0065004	protein-DNA complex assembly	7.40	6.09E-4	3	2	
GO:0071824	protein-DNA complex subunit organization	6.74	1.19E-3	3	3	

View extra details

**Gene set** GO:1901605 **Enrichment score** 9.65303

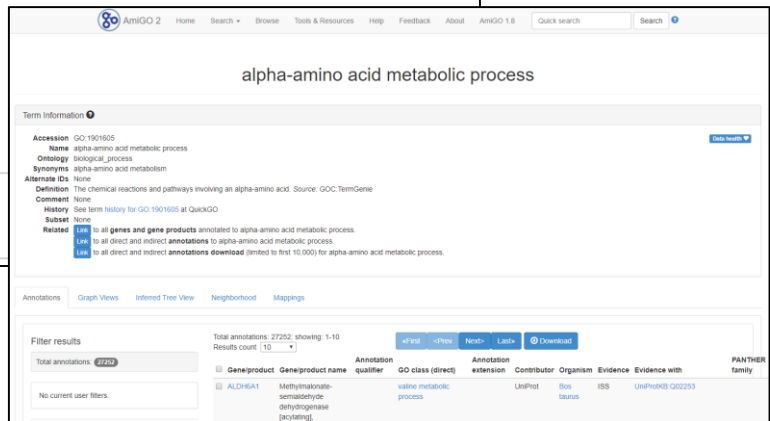
**Description** alpha-amino acid metabolic process **P-value** 6.42309E-5

**Gene breakdown**

	In list	Not in list
<b>In set</b>	3	0
<b>Not in set</b>	14	383

▶ **Genes in list**

▶ **Genes not in list**



The screenshot shows the AmiGO 2 interface for the term 'alpha-amino acid metabolic process'. It includes sections for Term Information (Accession, Name, Ontology, Synonyms, etc.), Annotations, and Filter results. The filter results table shows a total of 2722 annotations, with a list of gene products including ALDH6A1.

Gene/product	Gene/product name	Annotation qualifier	GO class (direct)	Annotation extension	Contributor	Organism	Evidence	Evidence with	PANTHER family
ALDH6A1	Methylmalonate-semialdehyde dehydrogenase (acetylating)		metabolic process		UniProt	Sus laevis	ISS	UniProtKB:Q02293	

**Notes:**

---



---



---



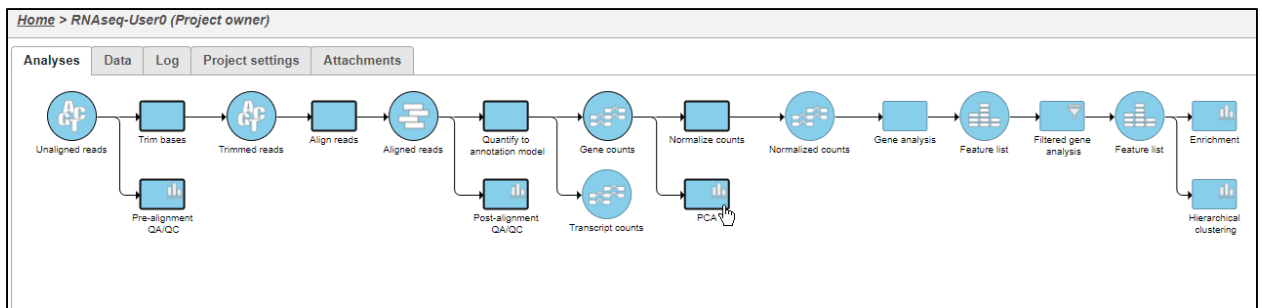
---



---

# Creating Pipelines

- Pipelines allows you to repeat the same set of tasks on different datasets
- On the Analyses tab, click **Make a pipeline** at the lower-left of the page
- Name the pipeline as **RNAseq-Pipeline-[username]**
- Select **Section name: Pipelines** then select the task nodes (rectangles) to include in the pipeline
- Click **Make pipeline** to create the pipeline



Click on the tasks above to include in the pipeline. Then click **Make pipeline** below.

Pipeline name:  Description:

Section name:

**Notes:** \_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

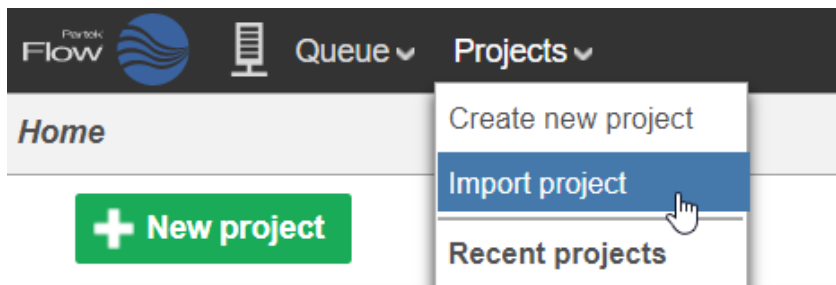
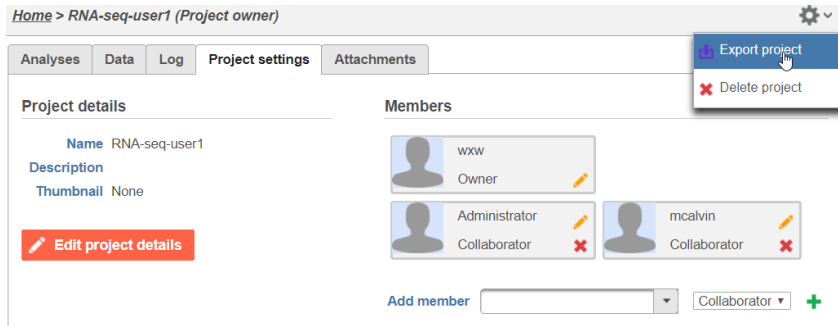
\_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

# Project Sharing

- Add collaboration on the project
- Import and export project



Notes: \_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_



# Further Training

---

## Self-learning

- Check out <http://www.partek.com/flow-resources> for documentation and additional resources
- Recorded webinars available on <http://www.partek.com/webinars>

## Regional Technical Support

- [www.partek.com/PartekSupport](http://www.partek.com/PartekSupport)

**Notes:** \_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_