RNA-Seq Analysis in Partek® Flow®

HANDS-ON TRAINING

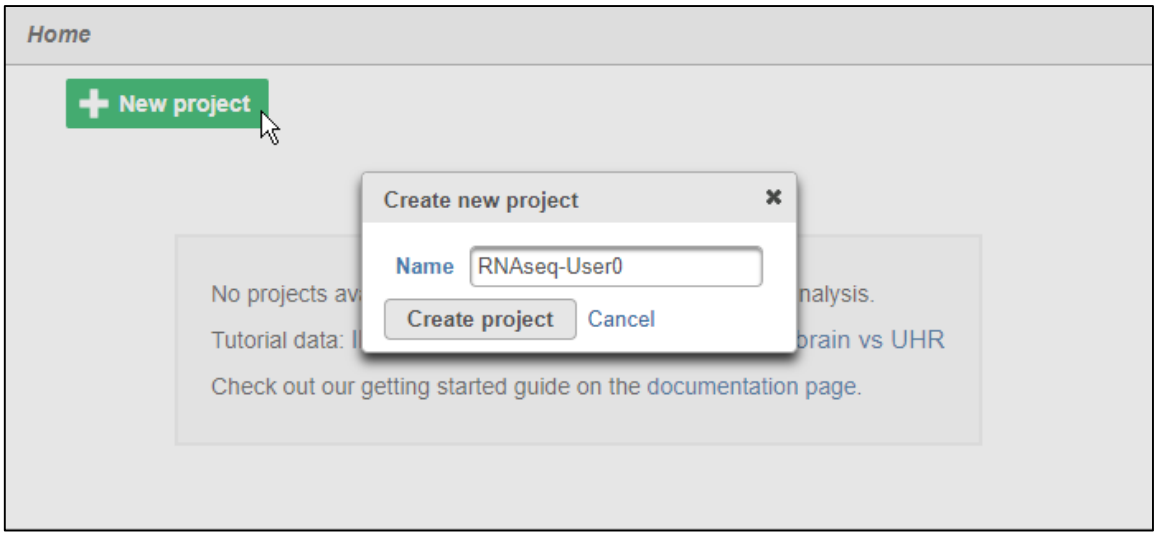
National Institutes of Health
November 2018

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

Partek®
turning data into discovery®

Login and Project Set-up

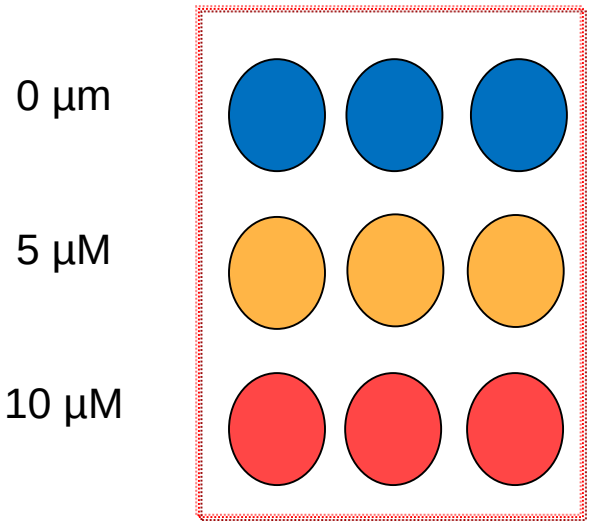
- Open Google Chrome and enter: training-server-url
- Log in using the username and password given to you
- This will open to the Partek Flow homepage
- Click **New Project** and enter project name: RNAseq-[username]
- This will create a new project



Notes: _____

Experiment Description

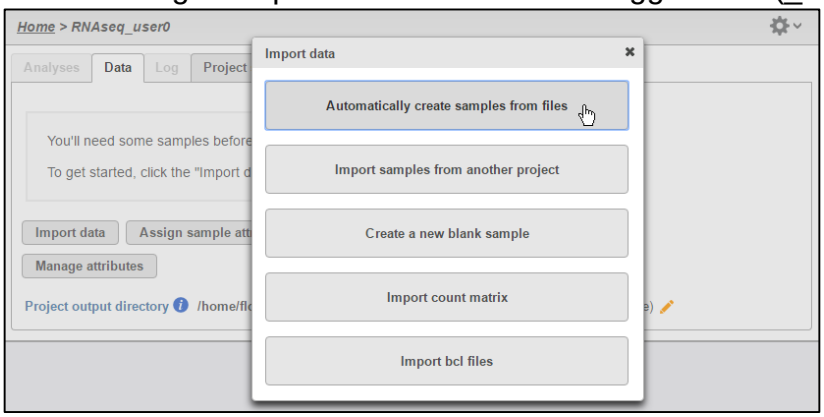
- HT29 colon cancer cells exposed to 5-aza drug with 3 different doses
    - 0 µM (Control)
    - 5 µM
    - 10 µM
- *Goal: Identify differentially expressed genes between different groups*
- mRNA purified and sequenced using Illumina HiSeq (Paired end reads)
- Xu et al. 2013 BMC Bioinformatics (PMID: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

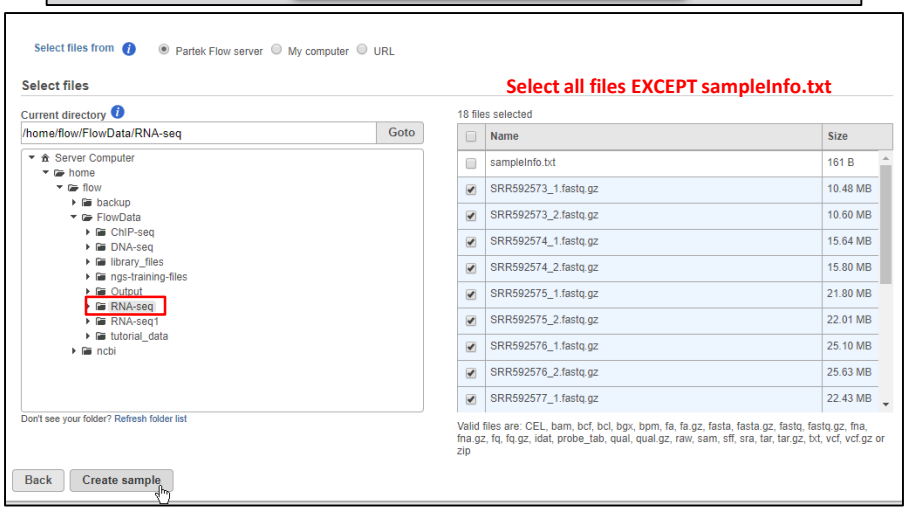


Notes: _____

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)

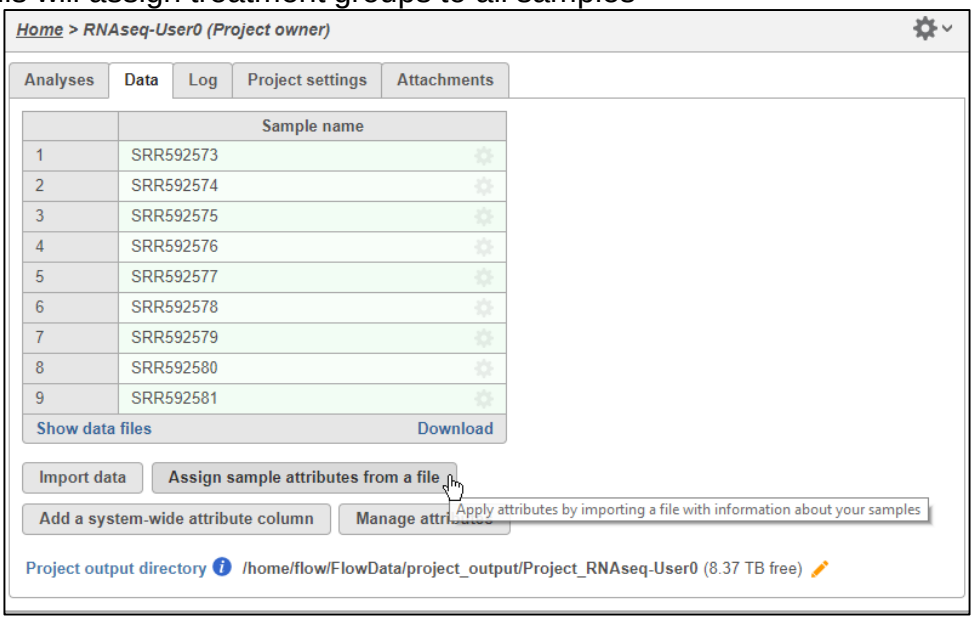


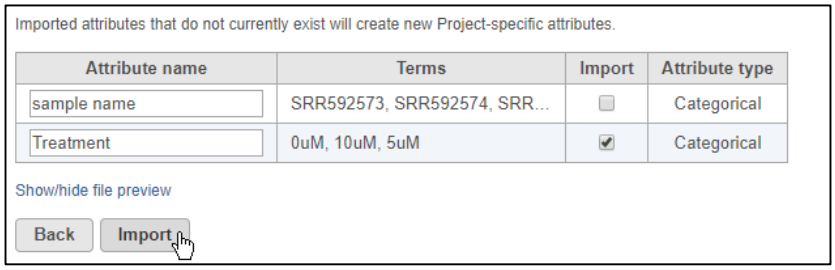


Notes: _____

Sample Attribute Assignment

- Assign sample attributes using a tab-delimited text file
  - Contains table with ID in 1st 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





Notes: _____

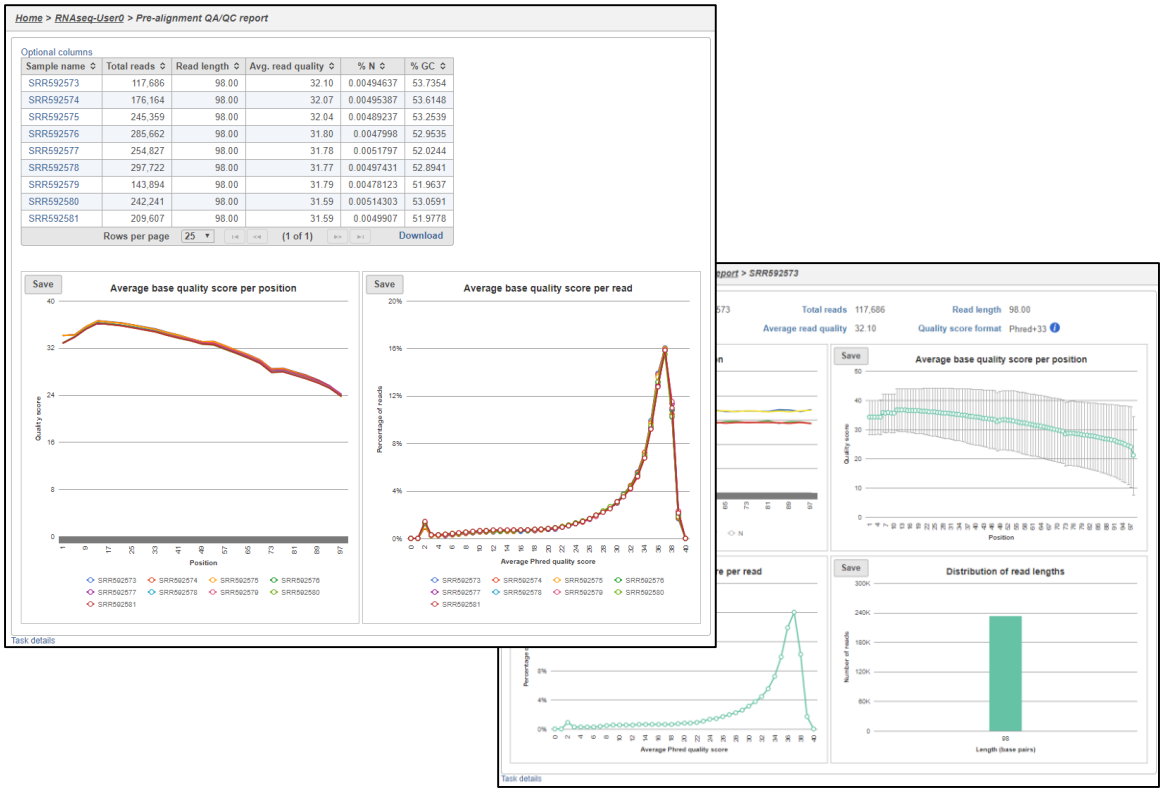
Analyses Tab Overview

- 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

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 $-10\log_{10}$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%

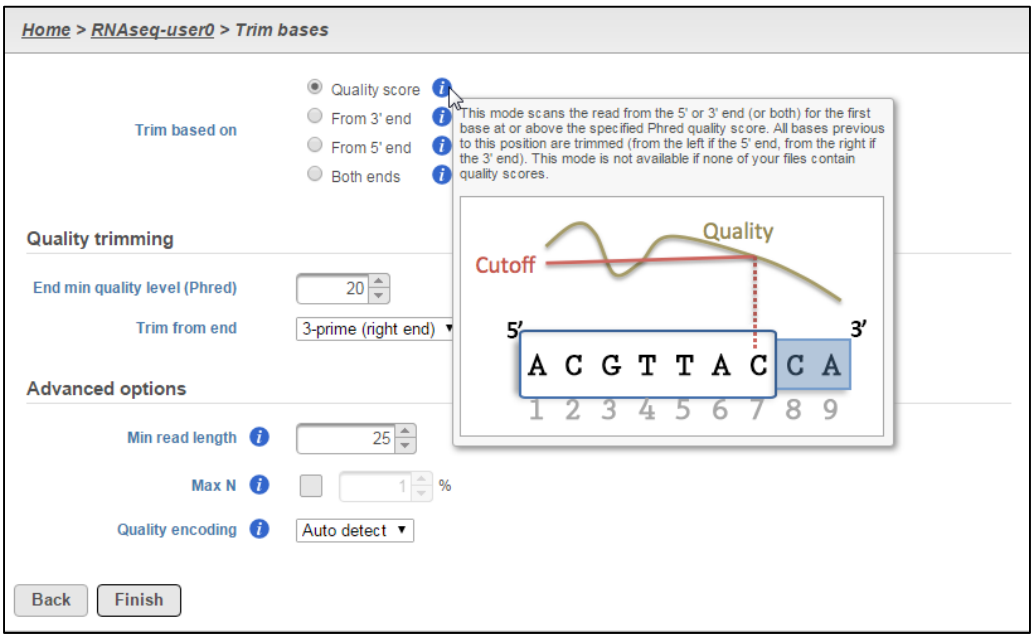
Notes: _____

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 1st new data node, the **Trimmed reads** data node

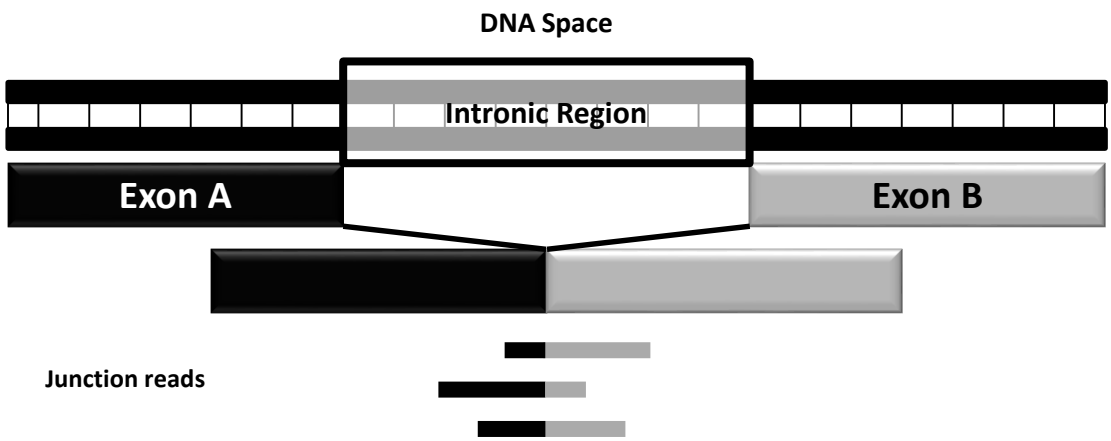
Tip: Hover over any ⓘ to get additional information about a specific option



Notes: _____

Aligning RNA-Seq Data

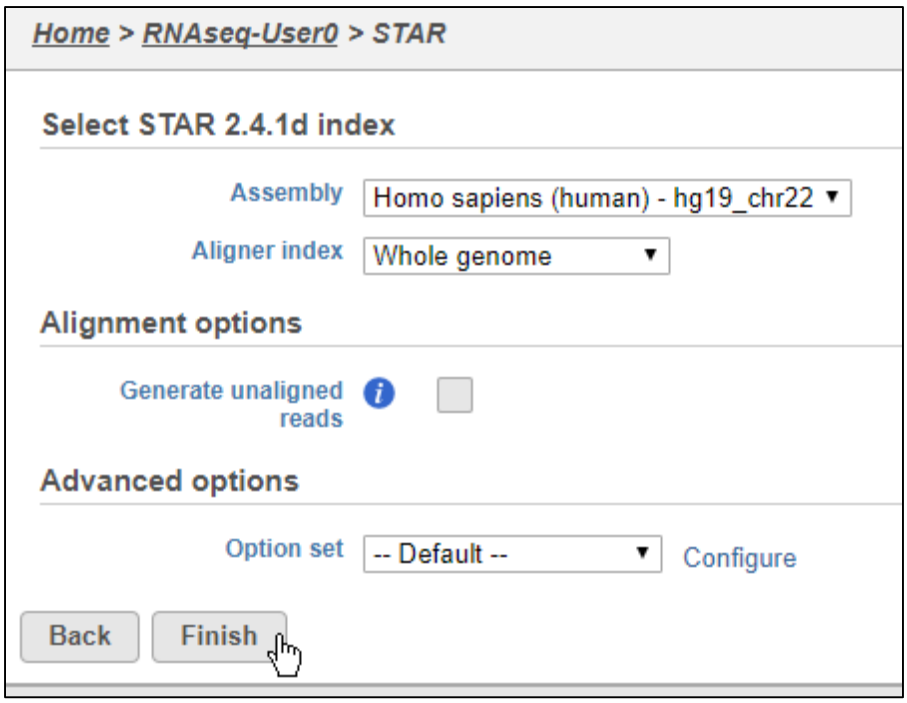
- RNA-Seq data must be aligned using an aligner that supports junction reads
- A junction read is one that spans two exons
- STAR is one of several aligners in Flow that you can use
  - Others include TopHat and GSNAP



Notes: _____

Alignment

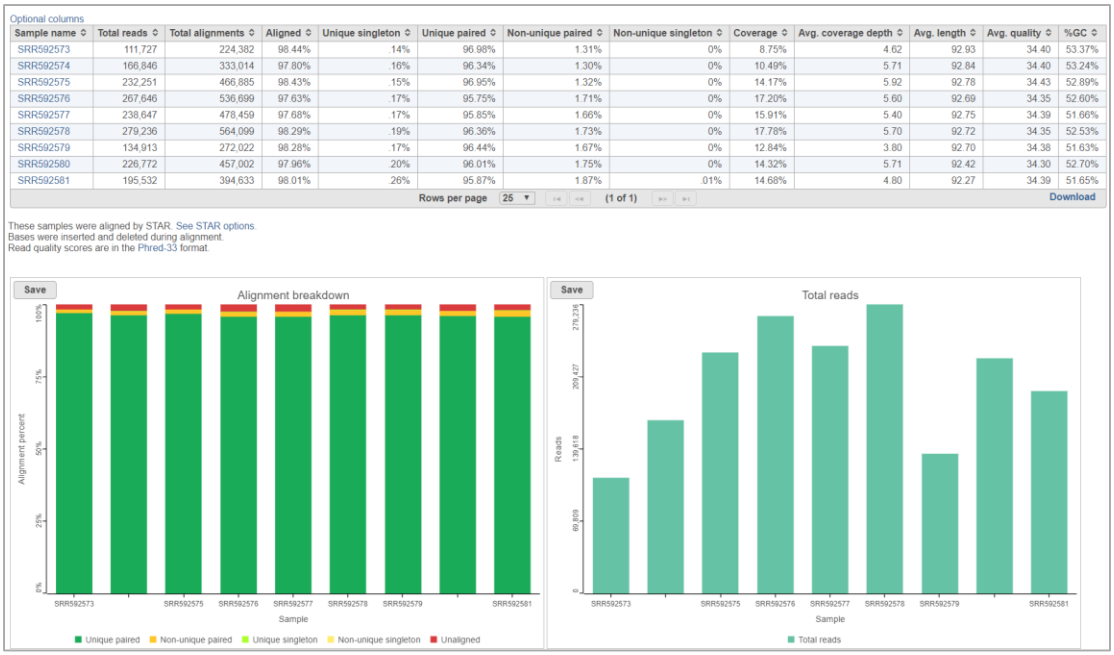
- Select the **Trimmed reads** data node
- Click **STAR** from the **Aligners** section of the menu
- Select STAR index:
  - Genome build: **Homo sapiens (human) - hg19_chr22**
  - Index: **Whole genome**
- Use the default options, click **Finish**



Notes: _____

Post-alignment QA/QC

- Perform Post-alignment QA/QC to asses 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



Optional columns

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	166,846	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%	.01%	14.68%	4.80	92.27	34.39	51.65%

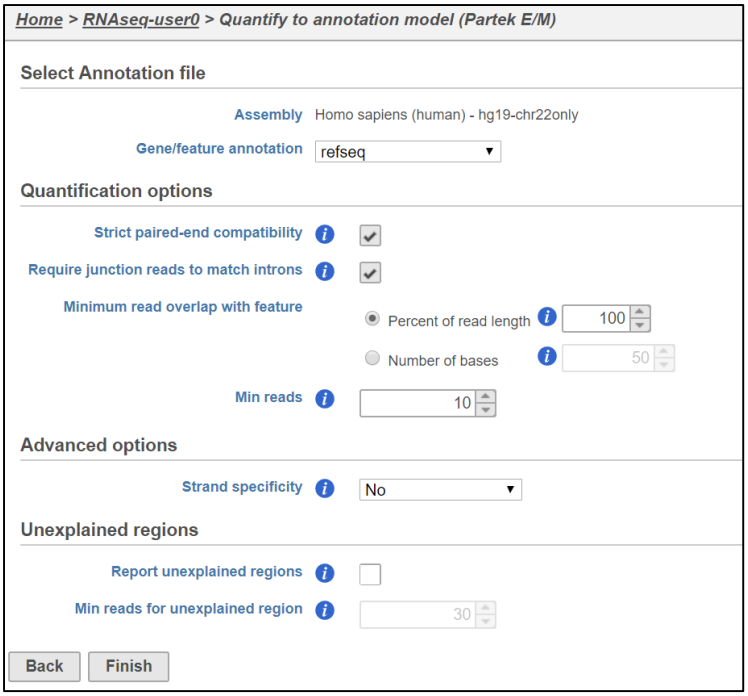
Rows per page 25 (1 of 1) Download

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



Notes: _____

Viewing Quantification Results

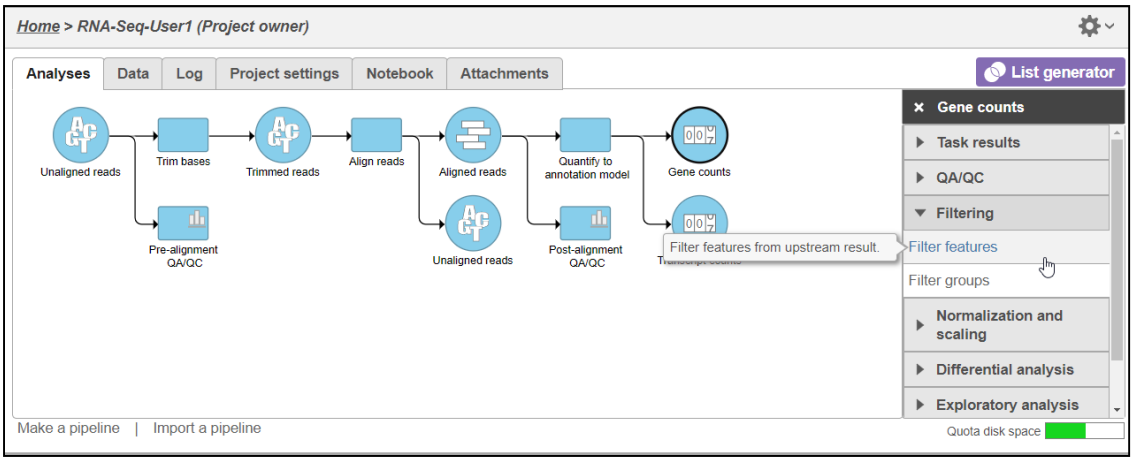
- Since the RefSeq annotation has both *gene*- and *transcript-level* information, this task will generate 2 data nodes:
  - Gene counts
  - Transcript counts
- To view the results, double-click the **Gene counts** data node
- Data at this level can be downloaded as text file containing count matrix

Sample name	Total reads	Fully within a feature	Partly within a feature	Not in a feature	Incompatible paired-end	Compatible junctions	Total junctions	View
SRR592573	109,985.00	55.12%	2.85%	35.35%	6.68%	20,435.00	25,161.00	
SRR592574	163,175.00	55.61%	2.82%	35.09%	6.49%	29,924.00	36,820.00	
SRR592575	228,614.00	50.69%	3.21%	39.39%	6.70%	38,826.00	48,447.00	
SRR592576	261,309.00	44.95%	3.85%	45.14%	6.06%	36,882.00	47,308.00	
SRR592577	233,123.00	48.13%	3.34%	42.69%	5.84%	33,702.00	42,946.00	
SRR592578								
SRR592579								
SRR592580								
SRR592581								
Average								

Notes: _____

Filter Features

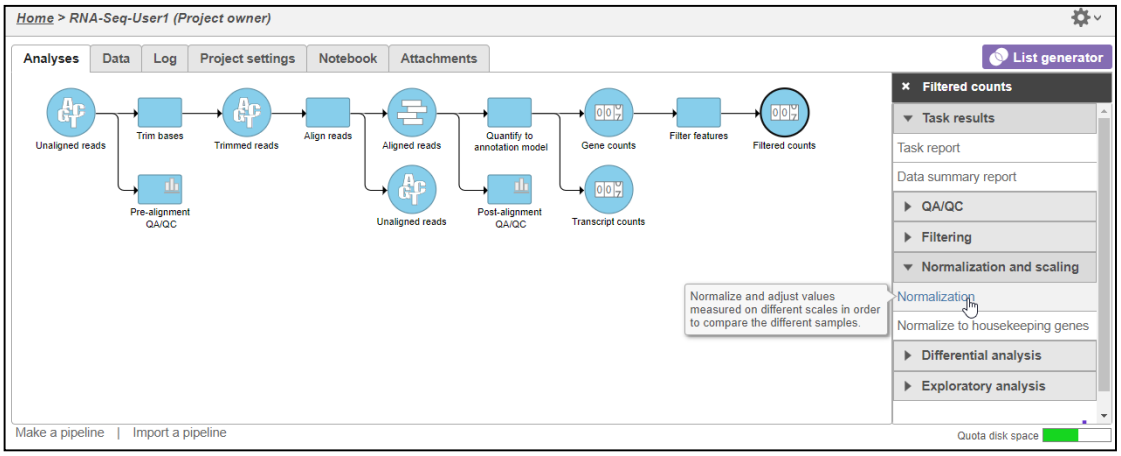
- Low expression genes maybe indistinguishable from noise, will decrease the sensitivity of DEG detection
- To filter out low expression at the gene level, select the **Gene counts** data node and click **Filter features** under the **Filtering** portion of the menu
- Choose **Filter exclude features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

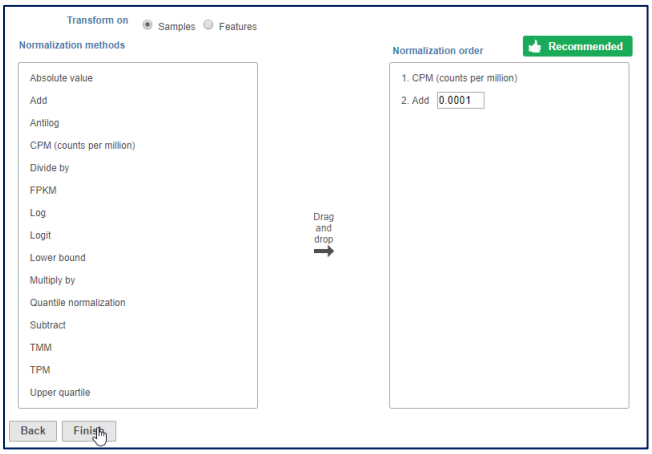


Notes: _____

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**

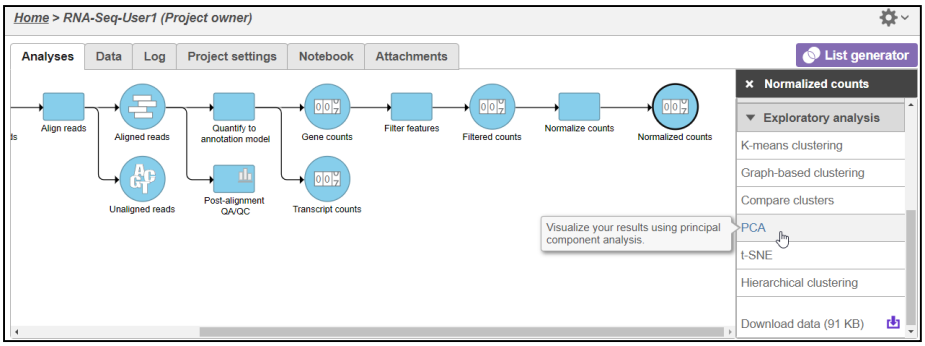


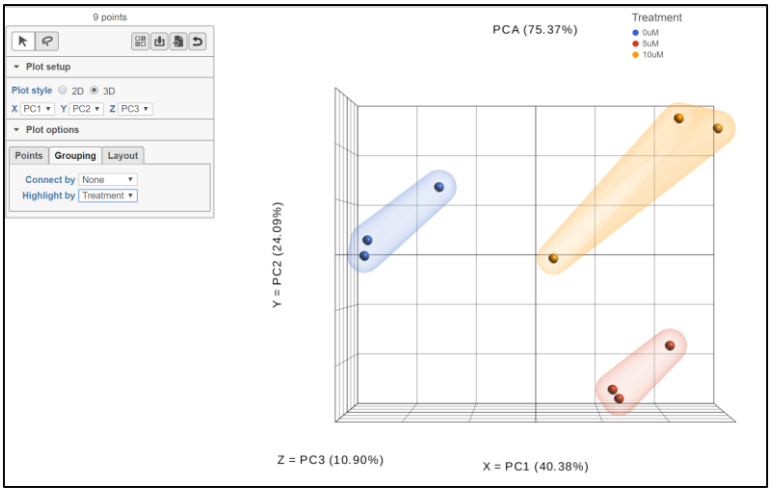


Notes: _____

Principal Components Analysis

- The principal components analysis (PCA) scatter plot allows you to assess relatedness between samples and identify outliers
- This can only 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**

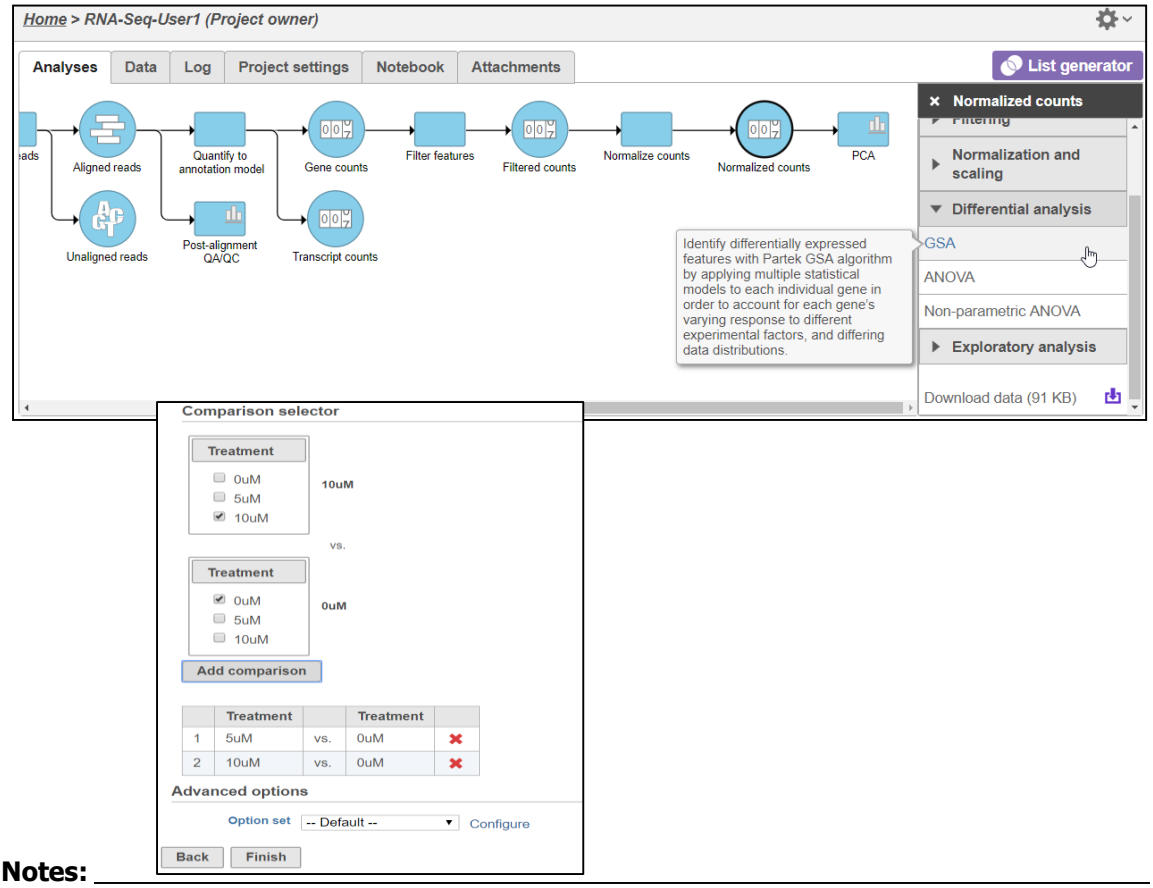




Notes: _____

Differential Expression Analysis

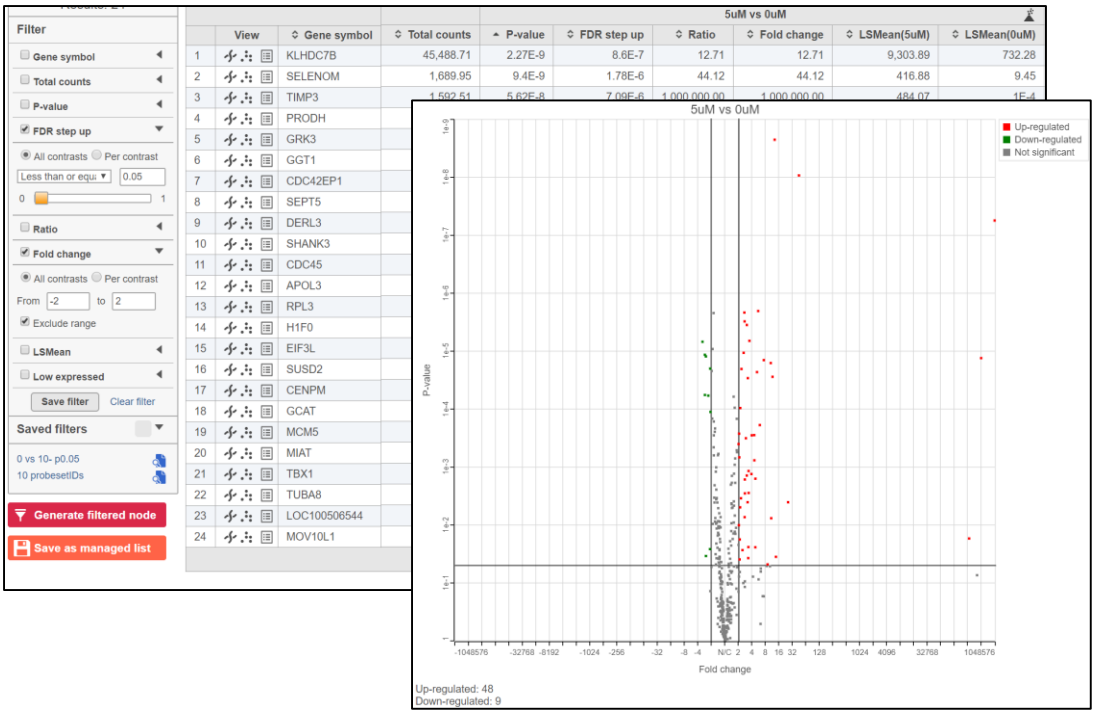
- Select the **Normalized counts** data node
- Click **GSA** from the **Differential analysis** section of the menu
- Select **Treatment** as an attribute to include in statistical test and click **Next**
- Setup the following comparisons and click the *Add comparison* button
  - **5uM vs 0uM**
  - **10uM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



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: _____

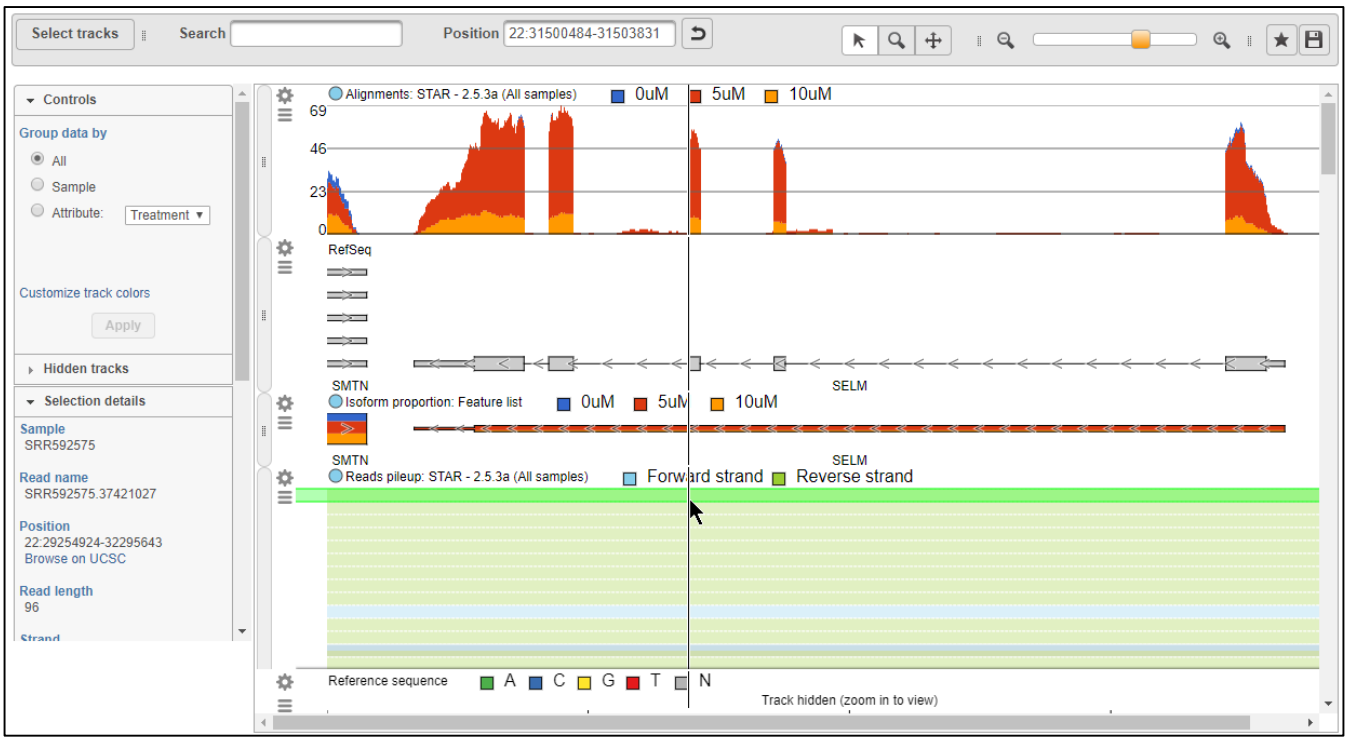
Viewing Gene/Transcript Level Results

- 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

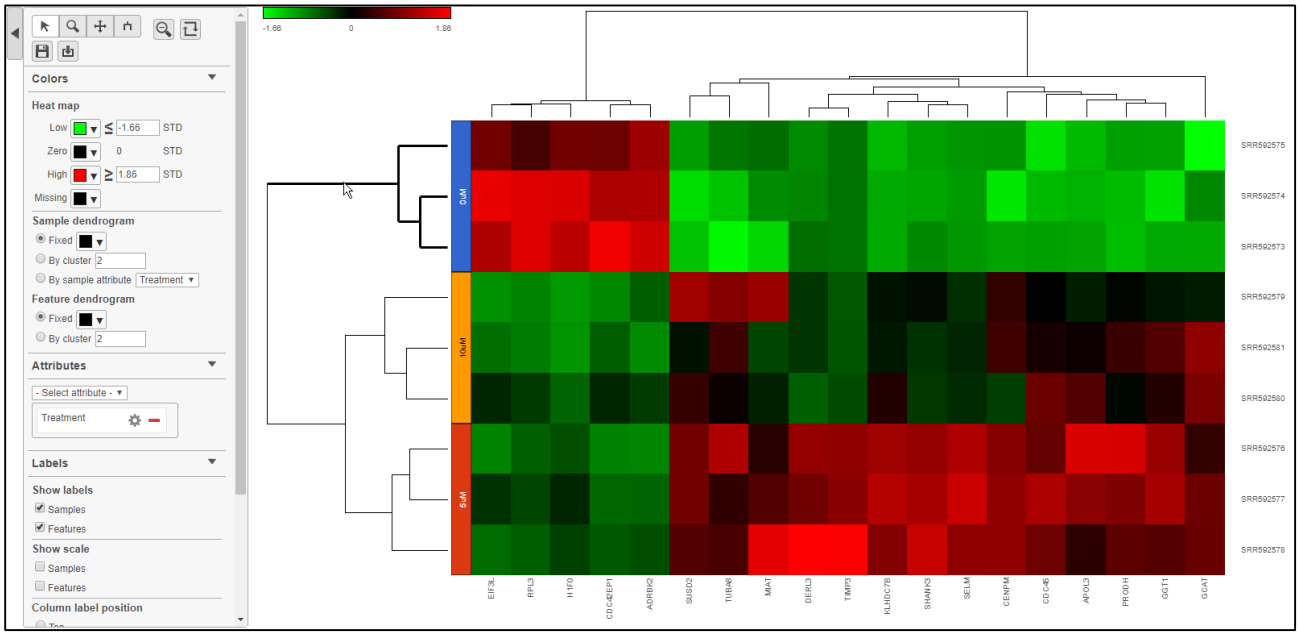
- **Select tracks** allows you to select different annotations or datasets to view together
- Sample grouping, color and transcript labeling can be edited in the **Controls** panel
- Search for any gene using the **Search** box
- Navigate to a genomic coordinate using the **Position** box
- Change and 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	View extra details
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	

Gene set GO:1901605 Enrichment score 9.65303

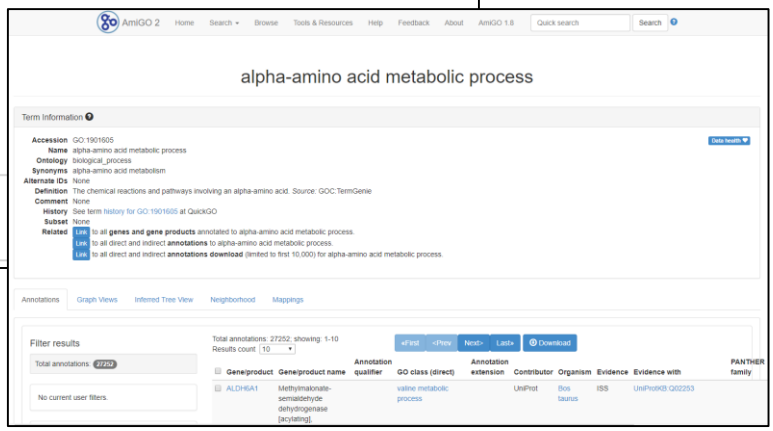
Description alpha-amino acid metabolic process P-value 6.42309E-5

Gene breakdown

	In list	Not in list
In set	3	0
Not in set	14	383

Genes in list

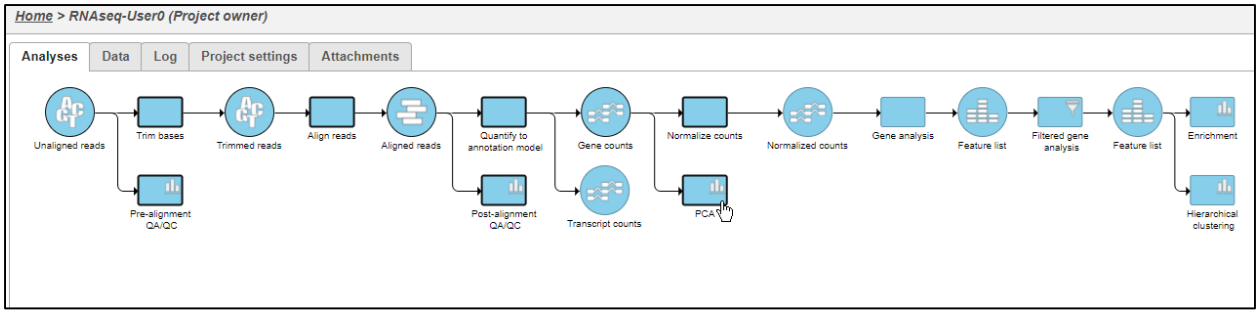
Genes not in list

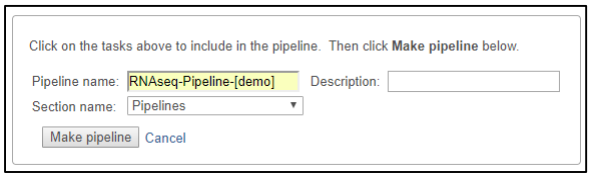


Notes: _____

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

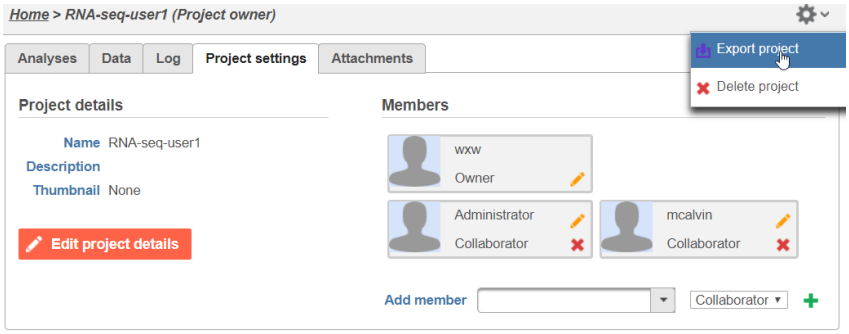


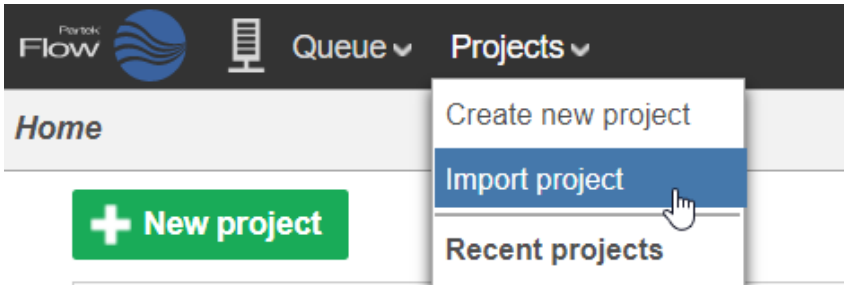


Notes: _____

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

Notes: _____