Introduction to Bulk RNA Analysis using Partek Flow

Alexandra L Emmons Ph.D. & Joe Wu Ph.D.
BTEP/GAU/CCR/NCI/NIH - email ncibtep@mail.nih.gov
Bioinformatics Training and Education Program

Table of Contents

Class Overview

Signing onto Partek Flow

Create new Partek Flow project

Import data to project

Pre-alignment QC

Removing adapters

Pre-alignment QC after Adapter Trim

Mapping sequences to genome

Post-alignment QC

Quantifying expression

Normalization

Principal components

Similarity plot

Expression heatmap

Filtering normalized counts

Gene set enrichment analysis

Differential expression analysis

Over representation analysis

Class Overview

This class introduces participants to bulk RNA sequencing analysis using the point-and-click software Partek Flow. Partek Flow enables researchers to build comprehensive workflows for analyzing data derived from many sequencing modalities on the bulk and single cell level (ie. RNA, ChIP/ATAC, CITE, and spatial transcriptomics). NCI holds an institutional license for this package. Please see https://bioinformatics.ccr.cancer.gov/docs/getting-started-with-partek-flow/ *(https://bioinformatics.ccr.cancer.gov/docs/getting-started-with-partek-flow/)* to learn how to get access and about different methods for transferring data to the NIH Partek Flow server. NHGRI also holds an institutional license to this software (see https://research.nhgri.nih.gov/bi-training.shtml *(https://research.nhgri.nih.gov/bi-training.shtml)*). Scientists not affiliated with NCI or NHGRI can inquire with the NIH Library (https://www.nihlibrary.nih.gov/resources/tools/partek-flow *(https://www.nihlibrary.nih.gov/resources/tools/partek-flow)*).

Class Expectations / Learning Objectives

This class will not turn participants into expert bioinformaticians or Partek Flow users. However, concepts learned can be applied to other sequencing type using Partek Flow and provides a foundation for continual learning.

After this class, participants will become familiar with steps for analyzing bulk RNA sequencing data using Partek Flow, including:

- Creating new analysis project and importing FASTQ files into an analysis project.
- Performing quality control and cleanup of FASTQ files for use with downstream analyses.
- Aligning sequences in FASTQ files to reference genome and interpret alignment quality metrics.
- Generating gene expression table from aligned reads.
- Performing differential gene expression analysis.
- Constructing plots such as PCA, heatmap, and volcano plot to visualize RNA sequencing data.

Required Course Materials

This class is not hands-on. Experience using or access to Partek Flow is not required for participation.

> **Tip**
>
> For a review or introduction to RNA sequencing, see An Introduction to RNA-Seq: Overview of Expression Data Analysis *(https://cbiit.webex.com/cbiit/ldr.php?RCID=8d9dc8dba83ddf766fba789b29e45c55)*.

Self Learning Material

- BTEP Getting Started with Partek Flow at NIH *(https://bioinformatics.ccr.cancer.gov/docs/getting-started-with-partek-flow/)*
- Check out the BTEP Video Archive for recordings of previous Partek Flow trainings *(https://bioinformatics.ccr.cancer.gov/btep/btep-video-archive-of-past-classes/)*
- Information regarding Partek Flow on Biowulf *(https://partekflow.cit.nih.gov)*
- Partek Flow documentations *(https://documentation.partek.com)*

Contacts for Help

- BTEP: ncibtep@nih.gov
- Biowulf: staff@hpc.nih.gov
- Partek: support@partek.com

Download class data and try on your own

Click here to download the class data as a zip files to local computer. Macs should automatically unzip upon download but Windows users will have to unzip after download. Follow the instructions at https://bioinformatics.ccr.cancer.gov/docs/getting-started-with-partek-flow/data_transfer_pf_web/ *(https://bioinformatics.ccr.cancer.gov/docs/getting-started-with-partek-flow/data_transfer_pf_web/)* to learn how to transfer data from personal computer to the NIH Partek Flow server.

The content in the zip file are:

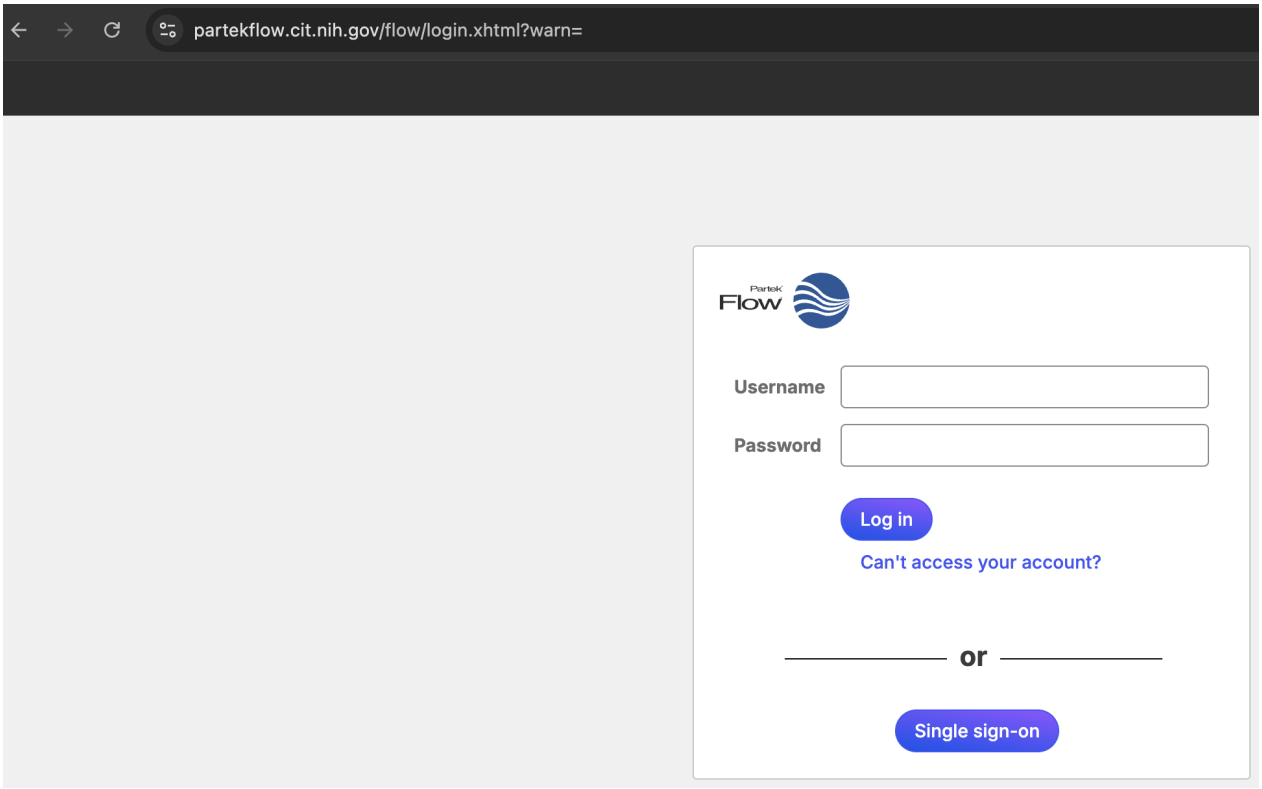
- `22.fa`: human chromosome 22 genome (source: Biostar Handbook: RNA-SEQ by Example, December 2021 *(https://www.biostarhandbook.com/computer-setup.html)*)
- `22.gtf`: human chromosome 22 gtf annotation (source: Biostar Handbook: RNA-SEQ by Example, December 2021 *(https://www.biostarhandbook.com/computer-setup.html)*)
- `c6.all.v2024.1.Hs.symbols.gmt` and `h.all.v2024.1.Hs.symbols.gmt` are the C6 and hallmark gene sets from msigdb *(https://www.gsea-msigdb.org/gsea/msigdb)*
- `hcc1395_fastqs`: paired end fastq files for hcc1395 (source: Griffith lab RNA bio *(https://rnabio.org/module-01-inputs/0001/04/01/Indexing/)*) - renaming of the files obtained from this source was required to fit the formats for some Partek Flow outputs (ie. tables and visualizations)

Tip

Upload `22.fa`, `22.gtf`, and the MSigDB gene sets to the user's Partek Flow library files. See https://documentation.partek.com/display/FLOWDOC/Library+File+Management *(https://documentation.partek.com/display/FLOWDOC/Library+File+Management)* to learn how.

Signing onto Partek Flow

Once a Biowulf and Partek Flow account has been obtained, use https://partekflow.cit.nih.gov/flow *(https://partekflow.cit.nih.gov/flow)* to connect to the Partek Flow server. Users can either supply username and password to authenticate or use NIH single sign-on.



Create New Partek Flow Project

Upon signing in, users will see a table containing links to existing projects. To create a new project, click on the "Add project" button in the Partek Flow landing page. For this class, a project called hcc1395_rna_sequnencing will be created.

Importing Data to Project and Assigning Metadata

The next step is to import data to the project. Click on the "Add data" button and select "Bulk". RNA sequencing is the default option and since FASTQ files will be imported, leave the "fastq" radio button selected. Click "Next" when ready. In the next page, users can navigate the Partek Flow folder of their own Biowulf account to select the needed files. Specify that the data is mRNA and hit "Finish" when ready. As the data is importing, users will see a rectangular task node. Once the data has successfully imported, the rectangular task node will turn into a circular data node.

After the FASTQ files have been imported, it is time to assign metadata to the files to help keep track of what condition each file came from. To do this, click on the "Metadata" tab in the project analysis page. Once in the "Metadata" page, click on "Show data files" and users will see the two paired end FASTQ files associated with the sample. Partek Flow uses the portion of the filename before "_R1.fq" and "_R2.fq" as the sample name. This class will assign metadata using the "Assign values from file" options as this is more convenient. The metadata are available in the tab delimited file "hcc1395_phenotype.txt" in the instructor's `./PartekFlow/uploads/hcc1395` folder. The contents of the file are below. Samples that start with "n" are normal and those starting with "t" are tumors, thus in this dataset there are 3 normal and 3 tumor samples. In either case, select "hcc1395_phenotype.txt" and click on "Next" when finished. In the next page, check the import box associated appropriated with the "Attribute name" or variable, which in this case is "disease_type" as there is already a column name "sample" containing the sample names. Click import when ready.

```
sample  disease_type
n1      normal
n2      normal
n3      normal
t1      tumor
t2      tumor
t3      tumor
```

Pre-alignment QC

The first step in analyzing RNA sequencing is to perform quality assessment of the FASTQ files. This step ensures that the quality of the data is good and there no issues with contaminations such as those arising from adapter read through.

To run pre-alignment QC, just click on the FASTQ data node and select QA/QC in the menu panel on the right of the analysis page. From there, select "Pre-alignment QA/QC" and make sure "All reads" is checked so that QC is performed on all reads in the FASTQ files. Then, run QC with the defaults.

> **Note**
>
> The K-mer length option when checked can be used to determine whether there are contamination such adapters in the sequencing data. However, because the adapter sequences are available for use during the trimming procedure, this option will not be used.

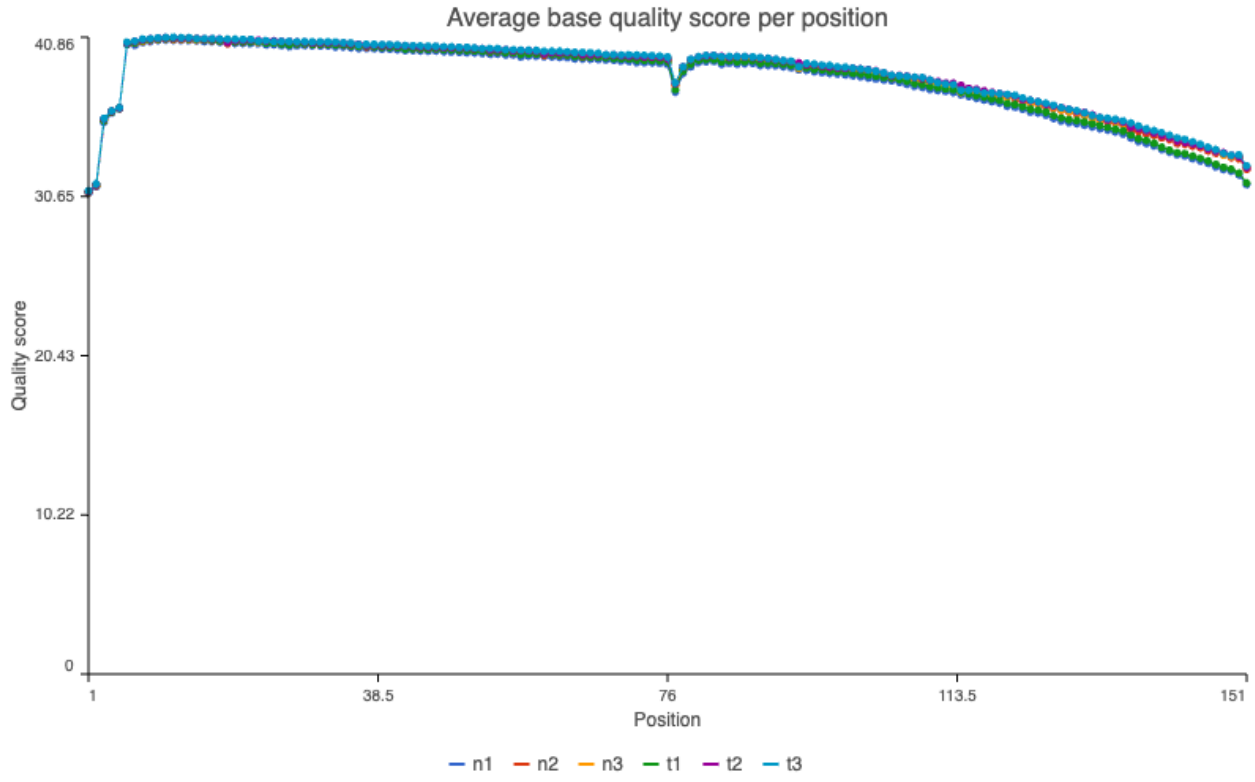
When the "Pre-alignment QA/QC" step completes, double click on the task node to view the results.

The first item in the "Pre-alignment QA/QC" report is a summary table and the columns in the table can be interpreted as follows:

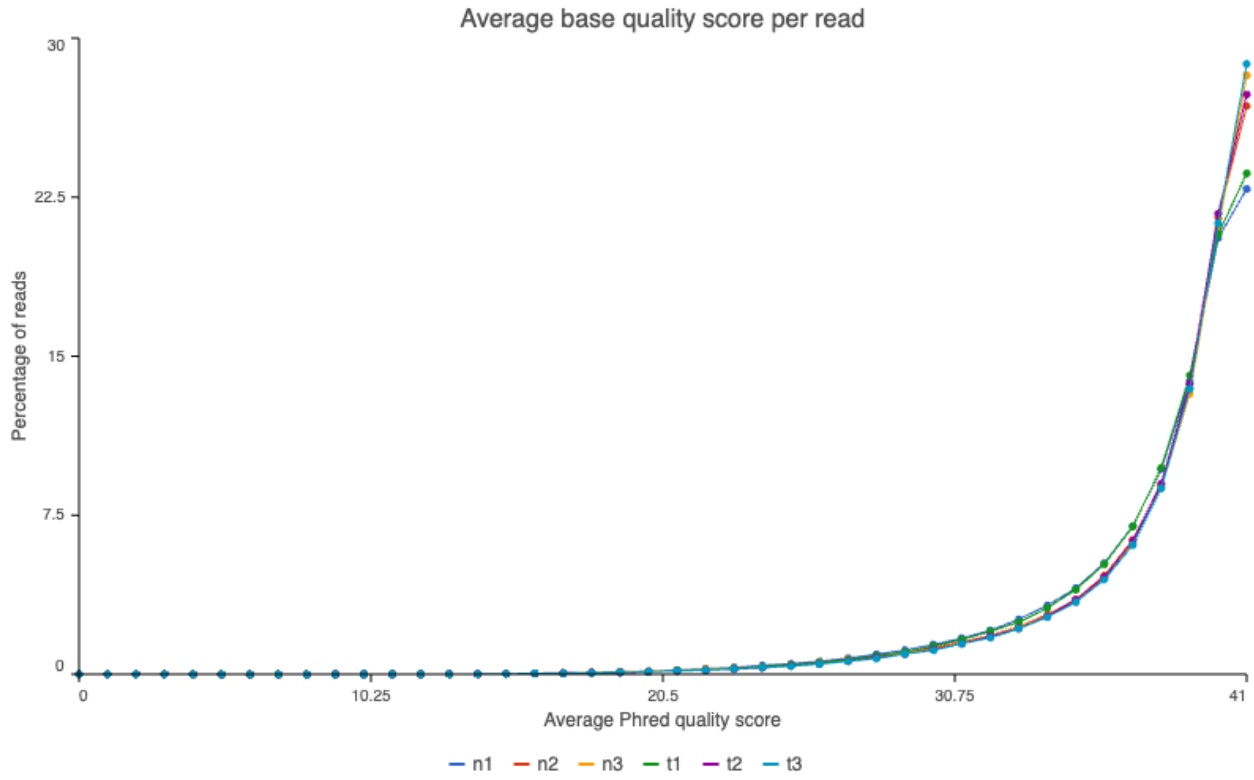
- `Sample name`: This indicates name of the sample. Partek Flow gives QC results on a per sample basis.
- `Total reads`: This is the total number of reads (or sequences) in the sample. For paired end sequencing, this refers to the total number of read pairs in the sample.
- `Average read length`: This reports the average length of the reads in the sample.
- `Average read quality`: Average of quality score for all reads in a sample is provided in this column. Higher numbers indicate that there is low likelihood for sequencing error. The samples in this dataset have high quality sequences. For instance, a quality score of 38 indicates a 0.0158% error likelihood.
- `% N`: This is the percentage of the unknown bases in the reads for a sample.
- `% GC`: This column shows the percent of the bases in the reads for a sample that are either G or C.

Sample name	Total reads	Read length	Avg. read quality	% N	% GC
n1	331,958	151.00	38.10	0%	54.22%
n2	331,958	151.00	38.42	0%	54.23%
n3	331,956	151.00	38.47	0.01%	54.23%
t1	390,607	151.00	38.25	0%	53.43%
t2	390,607	151.00	38.54	0%	53.45%
t3	390,607	151.00	38.58	0.01%	53.42%

The `Pre-alignmnet QA/QC` module averages the quality score of each base position along all sequences in a sample and results are shown the "Average base quality score per position" plot, which shows the average quality at each position for all reads/sequences in a sample. The figure below shows that each base position has an average quality of 30 or above, which indicates less than or equal to 0.1% error likelihood.



Next, the average quality score for each read in a sample is calculated and the distribution of the percentage of reads with a given average quality is generated. The image below shows that most of the sequences in the study samples have an average quality of 30 or above.

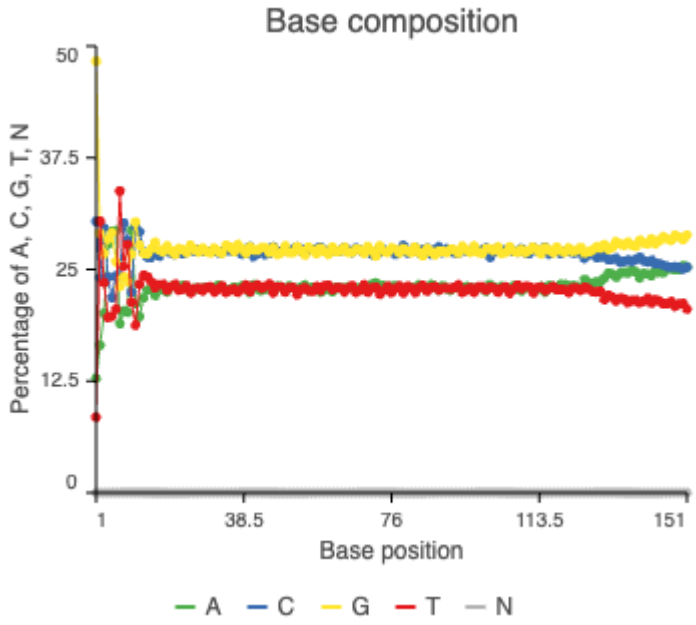


Click on any of the samples to view the sample-level QC results. This report contains a plot showing base composition for the reads in a sample.

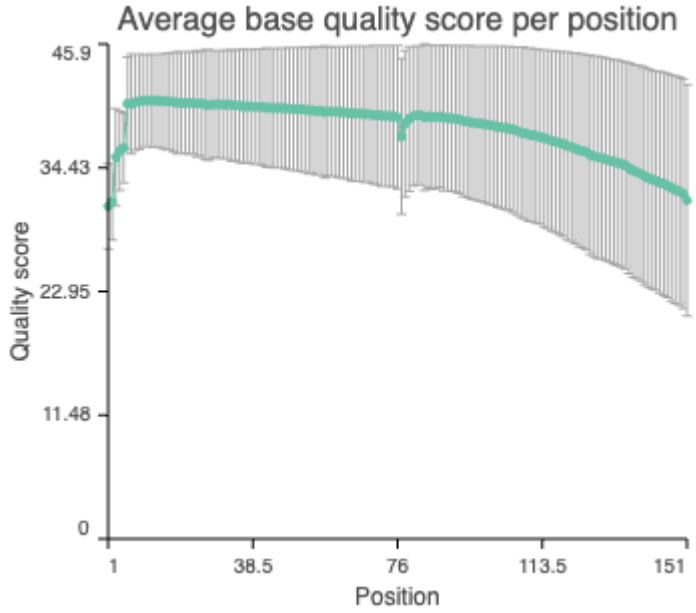
Tip

"In a random library you would expect that there would be little to no difference between the different bases of a sequence run, so the lines in this plot should run parallel with each other. The relative amount of each base should reflect the overall amount of these bases in your genome, but in any case they should not be hugely imbalanced from each other.

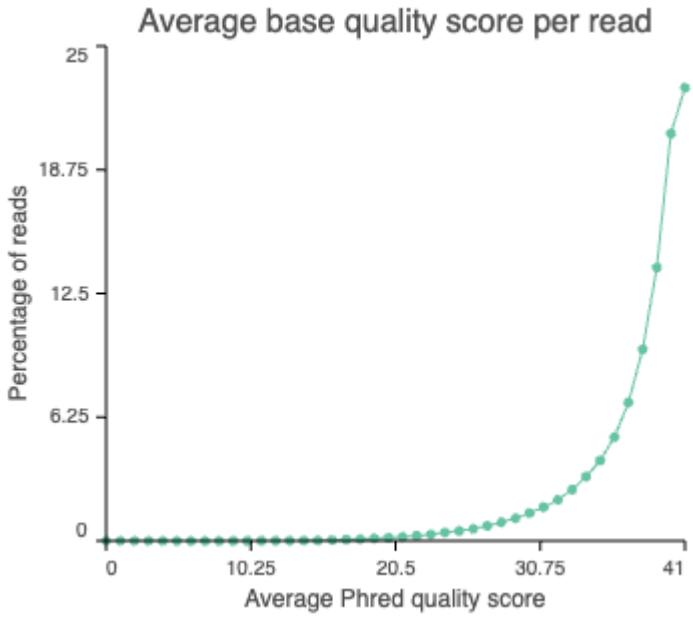
It's worth noting that some types of library will always produce biased sequence composition, normally at the start of the read. Libraries produced by priming using random hexamers (including nearly all RNA-Seq libraries) and those which were fragmented using transposases inherit an intrinsic bias in the positions at which reads start. This bias does not concern an absolute sequence, but instead provides enrichement of a number of different K-mers at the 5' end of the reads. Whilst this is a true technical bias, it isn't something which can be corrected by trimming and in most cases doesn't seem to adversely affect the downstream analysis. It will however produce a warning or error in this module." -- FASTQC manual *(https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/4%20Per%20Base%20Sequence%20Content.html)*



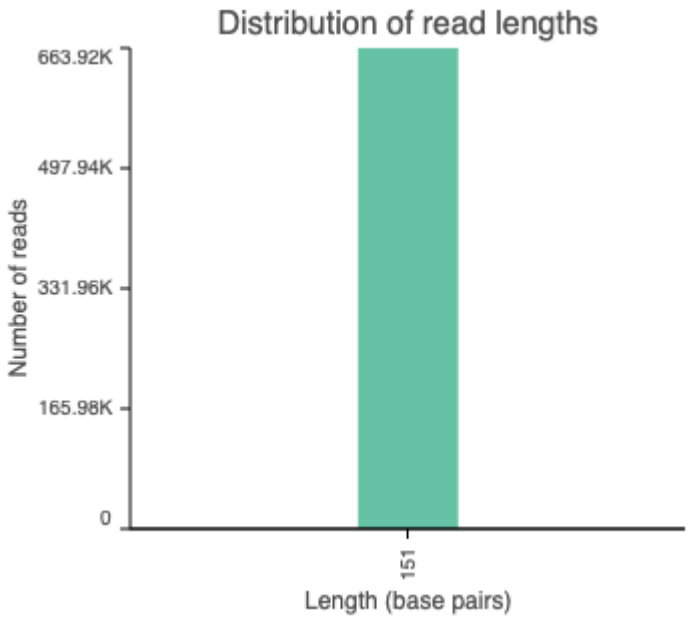
The sample-level report includes a plot showing average and range of quality score along each base position for all reads.



The sample-level read quality distribution is also provided.



A sample-level read length distribution plot is available as well. Note that prior to either quality or adapter trimming, all of the reads have the same amount of bases (151 in this example) as shown by the read length distribution plot.



Removing adapters

Because pre-alignment QC shows that the sequencing data have low error likelihood, the next step will be to remove adapter contamination by clicking on the FASTQ file data node, selecting "Pre-alignment tools" and then "Trim adapters". The adapters are stored in the file `illumina_multiplex.fa` located in the instructor's `PartekFlow/uploads/hcc1395/references` folder. The content in `illumina_multiplex.fa` are as follows.

```
>Multiplexing_Read_1_Sequencing_Primer_3_to_5
AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
>Multiplexing_Read_2_Sequencing_Primer_3_to_5
AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC
```

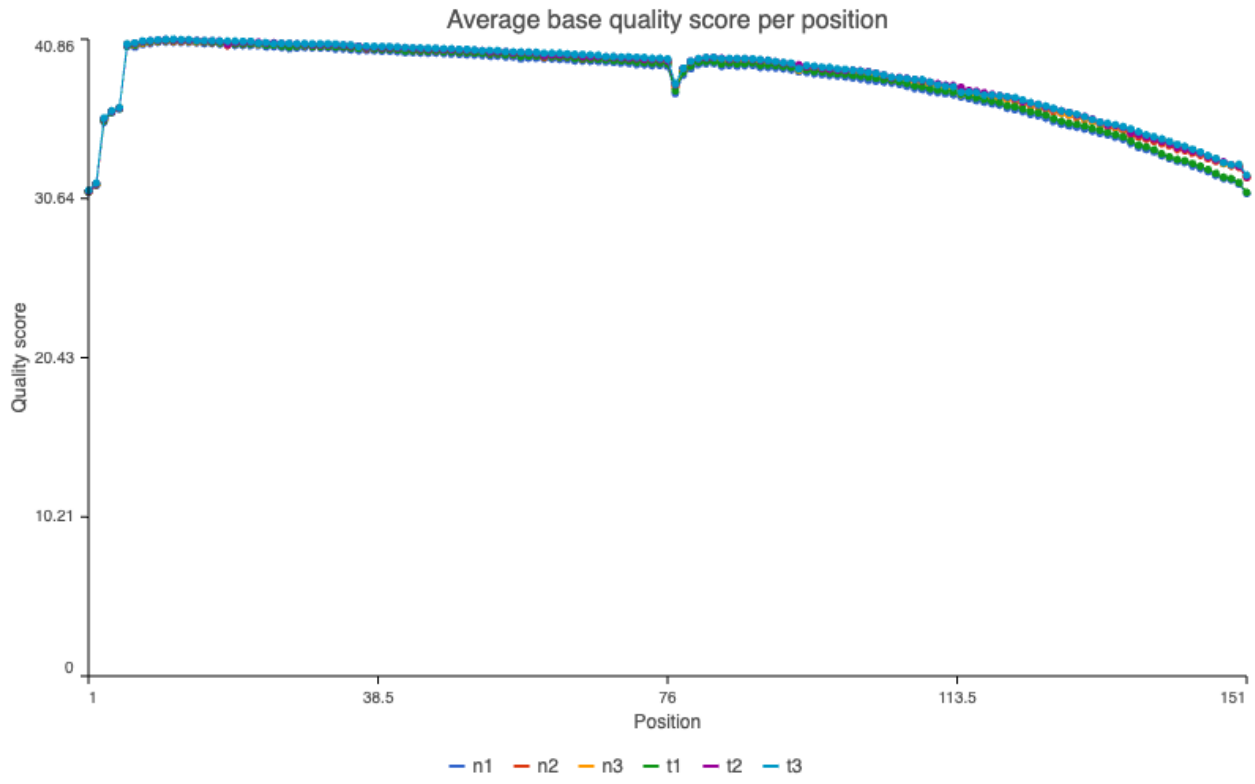
The screen recording below shows the steps for adapter trimming. Click on "Configure" under "Advanced options" to adjust the stringency criteria (ie. error rate, minimum overlap) that the trimming will use to determine whether adapter is present in a read/sequence. The minimum length for trimmed read to be kept in the analysis can be adjusted here as well.

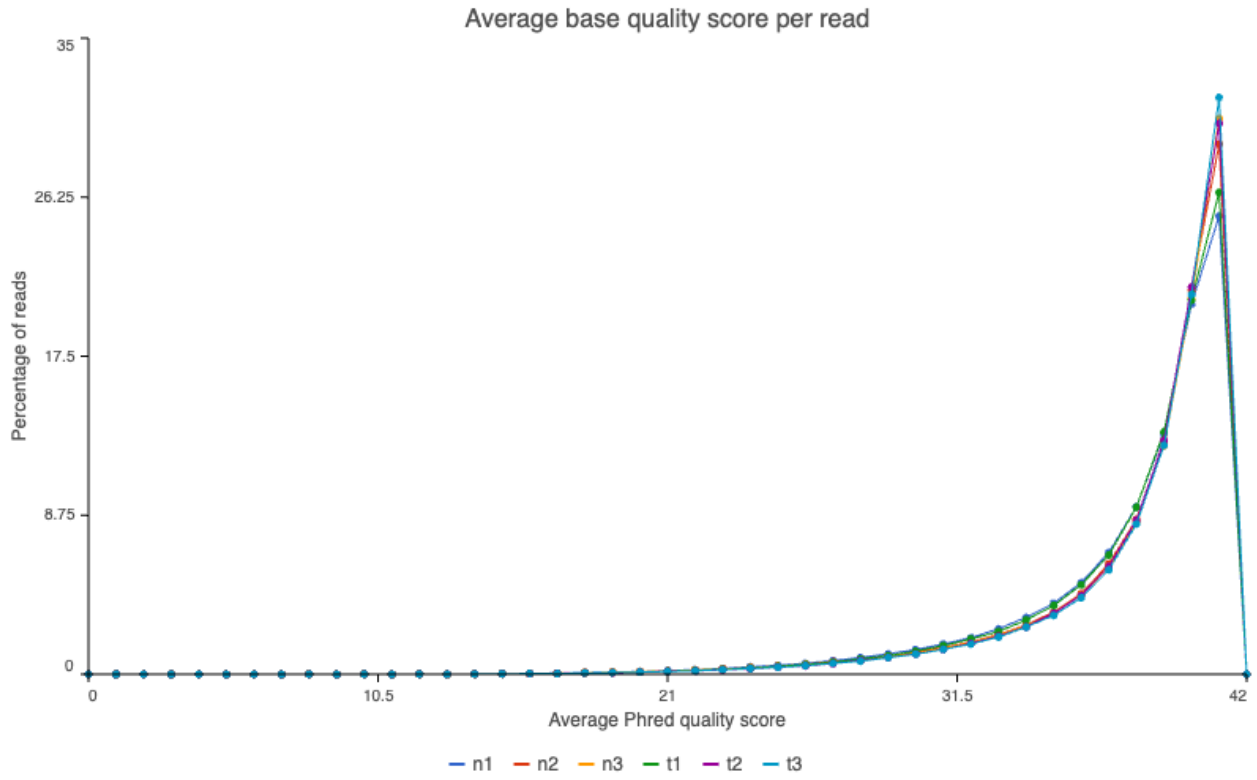
Run pre-alignment QC on the adapter trimmed reads before proceeding.

Pre-alignment QC After Adapter Trimming

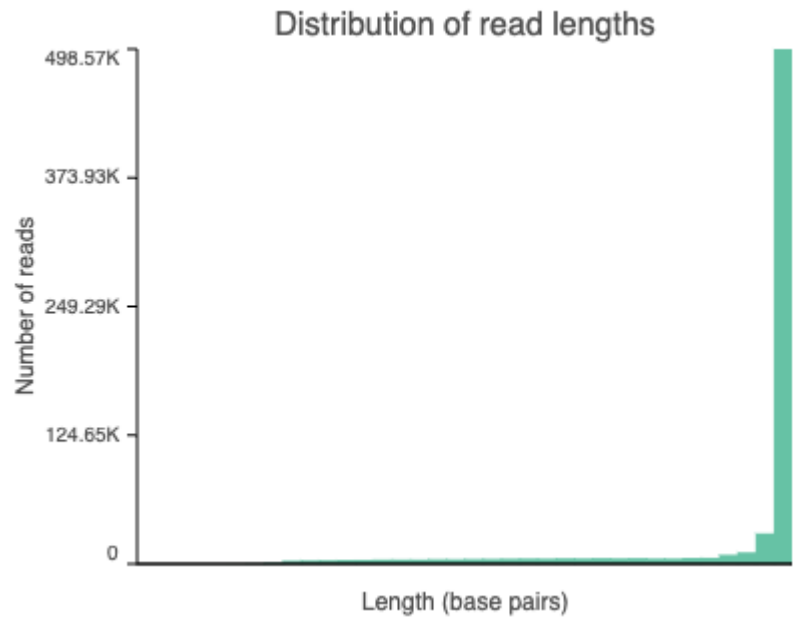
The QC results table below indicates adapter trimming resulted in some samples losing reads. This is because not all reads passed the QC criteria. Also, the average read length has been reduced as the adapters are removed from the reads. However, as indicated in the base quality and quality distribution plots, adapter trimming did not affect read quality.

Sample name	Total reads	Avg. read length	Avg. read quality	% N	% GC
n1	331,956	147.83	38.20	0%	54.27%
n2	331,958	147.75	38.51	0%	54.27%
n3	331,954	147.71	38.56	0.01%	54.28%
t1	390,604	146.96	38.38	0%	53.46%
t2	390,607	146.87	38.66	0%	53.48%
t3	390,605	146.80	38.70	0.01%	53.46%





As reads will have differing amounts of adapter contamination (ie. the number of adapter bases that appear), the trimming procedure will produce reads of varying lengths as indicated in the sample level read length distribution plot.



Mapping Sequences to Genome

After QC as well as quality and/or adapter trimming, it is time to map the sequences in the FASTQ files to the reference genome. RNA sequencing analysis requires the use of a splice aware aligner such as HISAT2 or STAR, which are both available on Partek Flow. In this class, HISAT2 will be used.

Note

The dataset, hcc1395, used for this class was subsetted to human chromosome 22. Thus, the sequences will be mapped to the hg38 chromosome 22 reference. See the Partek Flow documentations *(https://documentation.partek.com/display/FLOWDOC/Library+File+Management)* to learn how to add references and annotation files (ie. GTF files) to the user's Partek Flow account.

To map using HISAT2, click on the "Trimmed reads" data node and select "Aligners" in the menu. Then, click on "HISAT2". In the subsequent page, users will see an "Assembly" drop down box which will be used to select the index for the desired reference.

Note

HISAT2 indexes the reference prior to alignment in order to speed up the process.

If the reference index is not available in this drop down as in this example, scroll and select "New assembly". Keep the species as "Homo sapiens (human)" in the subsequent dialogue box labeled "Add HISAT2 index". Then select the reference file or assembly that needs to be indexed. In this case, "hg38_chromosome22". Keep the "Create option" as build as the HISAT2 needs index the reference prior to alignment. Click "Create" when ready. The "Index" drop down will populate with the newly built HISAT2 index when this step is done. Finally, click on "Finish" to start the alignment. HISAT2 will be run with default although users can configure it to meet their alignment stringency needs by clicking on "Configure" next to "Advanced options".

Post-alignment QC

After mapping, the next step is to perform post-alignment QC to determine things like overall alignment rate (ie. how many sequences aligned to the reference). To do this, select the "Aligned" reads data node and then select "QA/QC" from the menu. From there, click "Post-alignment QA/QC". After QC completes, click on the "Post-alignment QA/QC" task node to view results.

The first item in the "Post-alignment QA/QC" report is an alignment statistics table and an explanation of the columns is provided below.

- `Total reads`: This reports the number of reads or sequences in a sample. For paired end sequencing, this refers to the number of read pairs.
- `Total alignment`: This column indicates the number of times reads in a sample mapped to the genome. Do not confuse this with the number or percent of reads that mapped.
- `Aligned`: The percent of reads that mapped to the genome is provided here. The next four columns indicate the percentage of reads that where mapped uniquely (ie. to one location on the genome) or non-uniquely (ie. multimappers) and whether both reads in the pair or only one in the pair mapped (singleton). The value under the `Aligned` column is the sum of the `Unique-singleton`, `Unique paired`, `Non-unique paired`, `Non-unique singleton` columns.
- `Coverage`: This columns informs of the percentage or amount of bases in the genome that the reads in a sample cover.
- `Avg. coverage depth`: The average number of alignments in the region(s) of the genome covered by sequencing reads.
- `Avg. length`: The values in the column correspond to the average length (ie. number of bases) for the mapped reads in a sample.
- `Avg. quality`: This column reports the average quality of the mapped reads in a sample.
- `%GC`: The GC percentage of the mapped reads in a sample is given here.

All samples have greater than 97% alignment rate.

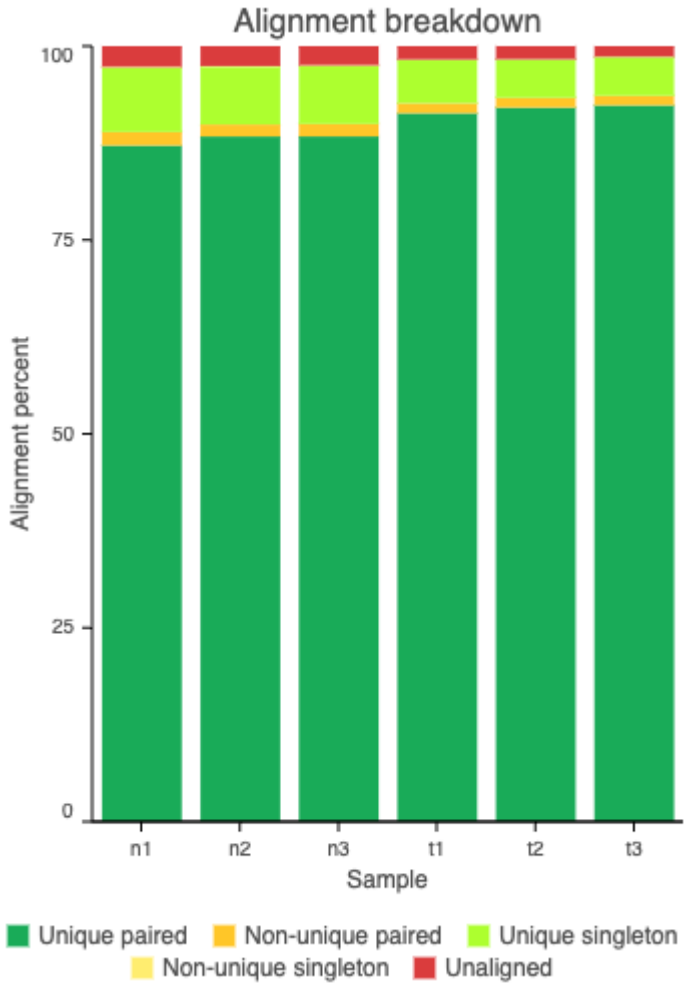
Sample name	Total reads	Total alignments	Aligned	Unique singleton	Unique paired	Non-unique paired	Non-unique singleton	Coverage	Avg. coverage depth	Avg. length	Avg. quality	%GC
n1	331,956	634,232	97.33%	8.30%	87.19%	1.74%	0.10%	11.87%	15.52	147.62	38.60	54.52%
n2	331,958	635,977	97.43%	7.44%	88.35%	1.55%	0.09%	11.81%	15.64	147.56	38.86	54.45%
n3	331,954	637,335	97.53%	7.46%	88.35%	1.61%	0.10%	11.80%	15.68	147.52	38.92	54.46%
t1	390,604	759,820	98.27%	5.57%	91.35%	1.29%	0.06%	12.99%	16.89	146.80	38.70	53.55%
t2	390,607	762,298	98.31%	4.92%	92.08%	1.26%	0.05%	12.87%	17.10	146.73	38.95	53.57%
t3	390,605	763,803	98.64%	4.99%	92.37%	1.23%	0.04%	12.94%	17.04	146.67	39.00	53.55%

The information shown in the above table are also presented as plots and among these is a stacked bar chart showing the percentage breakdown of alignment types discussed below.

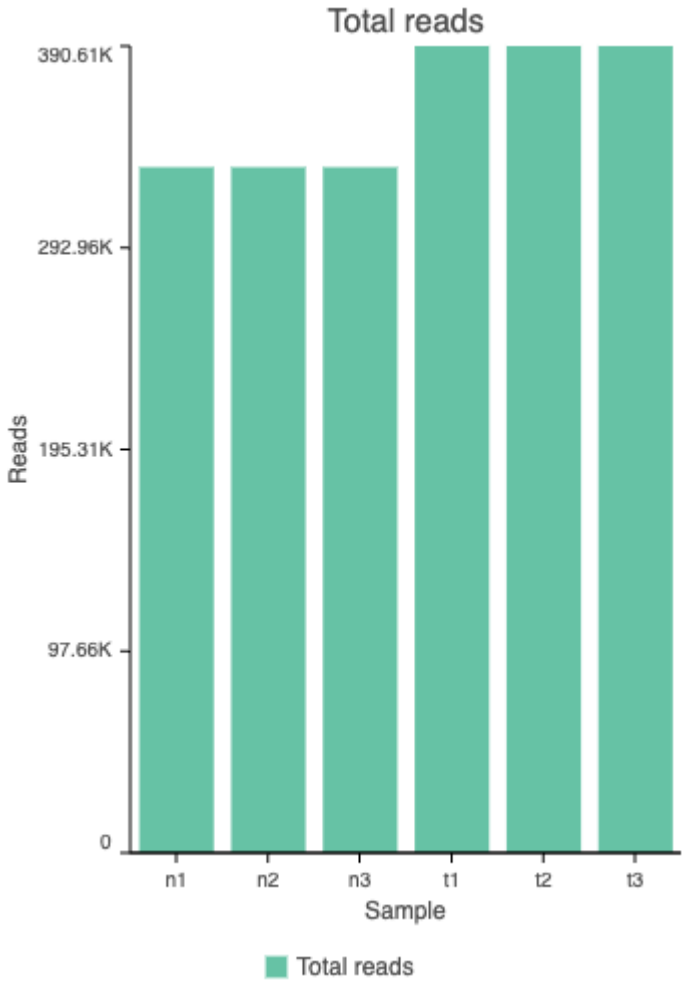
- Unique paired occurs when both reads in paired end sequencing align to only one genomic region.
- Non-unique paired happens when both reads in paired end sequencing align to more than one genomic region. These are consider multi-mappers.
- Unique singleton refers to only one read in paired end sequencing align to one genomic region.
- Non-unique singleton means that only one read in paired end sequencing aligned but to multiple genomic regions.

Note

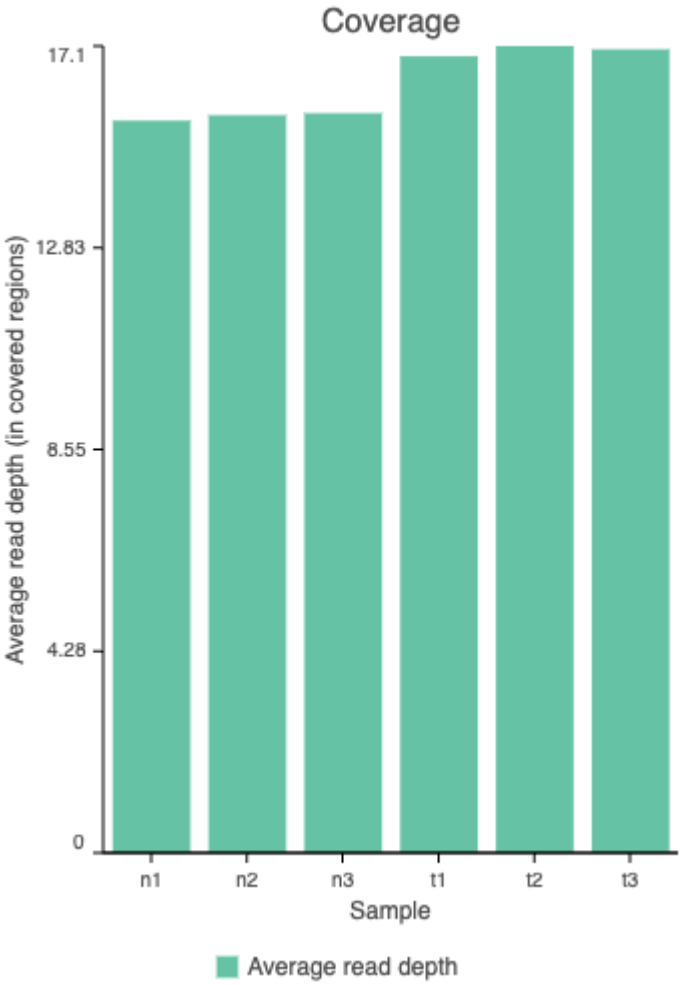
The ideal circumstance is that our reads align uniquely as this will not cause ambiguity in terms of determining which read goes to which gene or transcript when generating expression matrix.

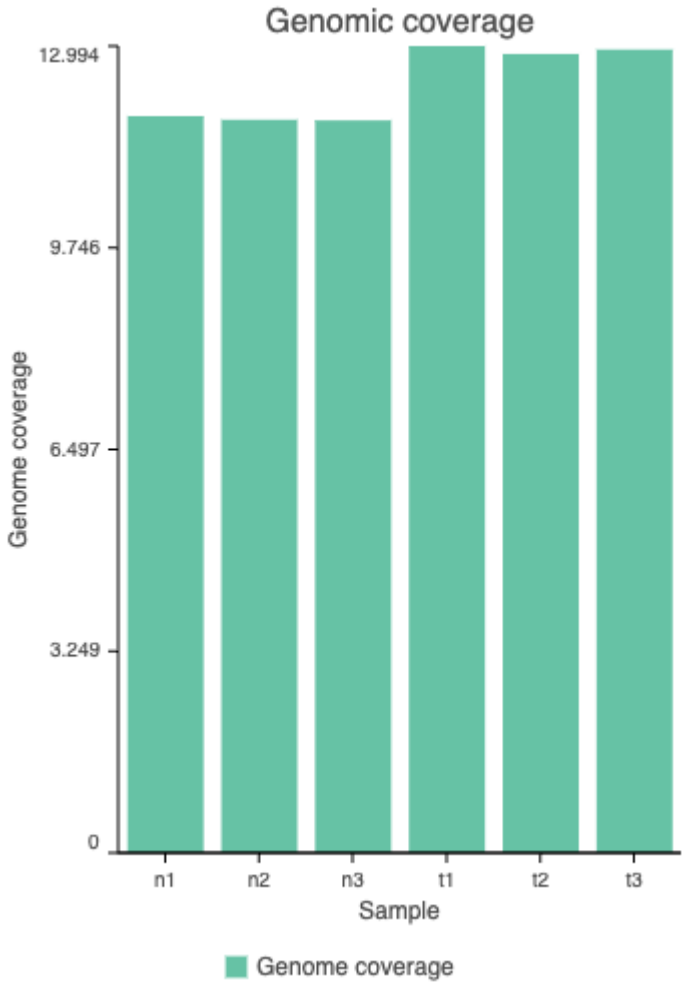


The next visualization provides the number of reads in a sample. Again, for paired end sequencing, this refers to the number of read pairs.

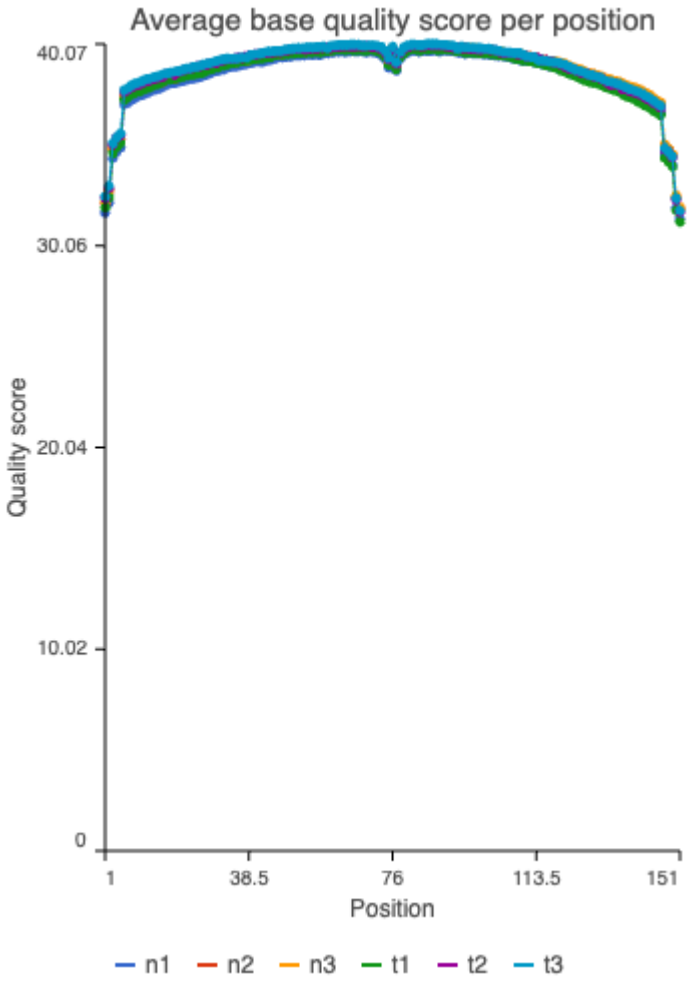


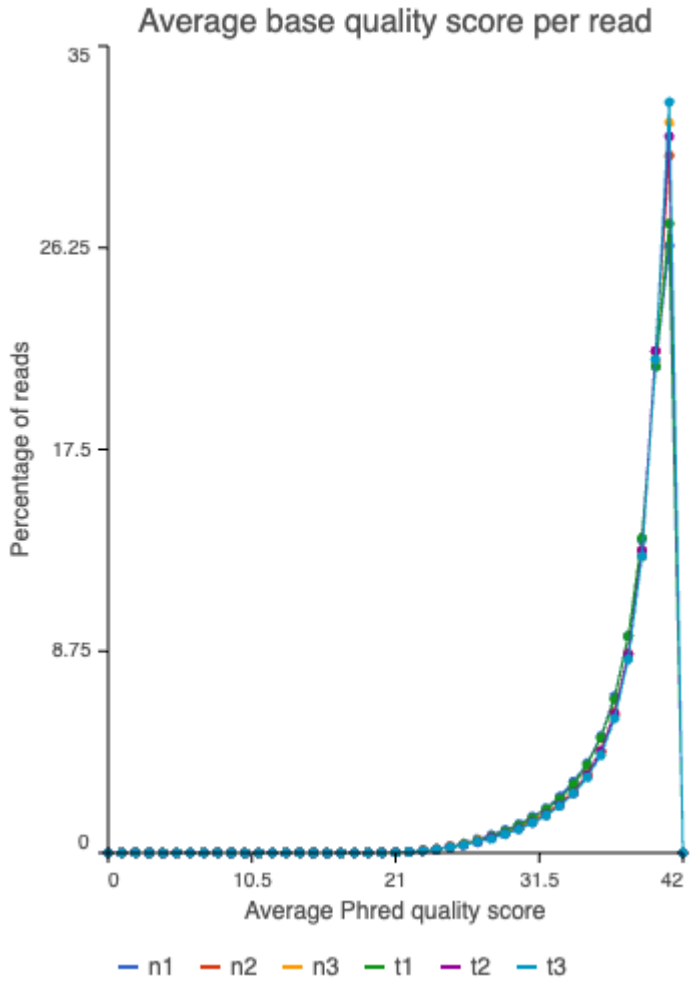
Also provided are bar charts of the average sequencing depth and the genomic coverage for each sample.


Coverage
17.1
12.83
8.55
4.28
0
Average read depth (in covered regions)
n1
n2
n3
t1
t2
t3
Sample
Average read depth

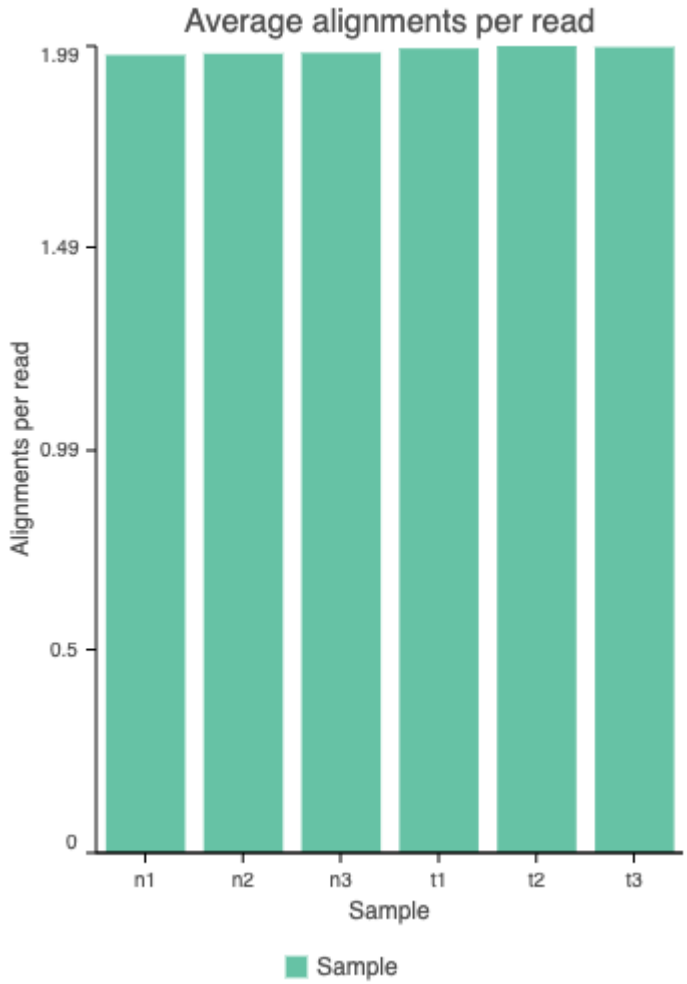


The average quality of each base and the quality distribution for the samples for all reads that aligned are also available as plots.

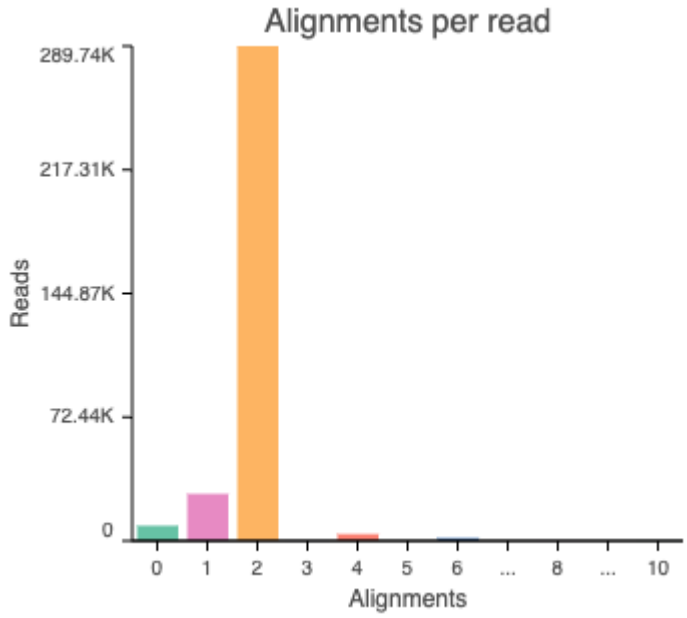

Average base quality score per position
40.07
30.06
20.04
10.02
0
Quality score
1
38.5
76
113.5
151
Position
n1
n2
n3
t1
t2
t3



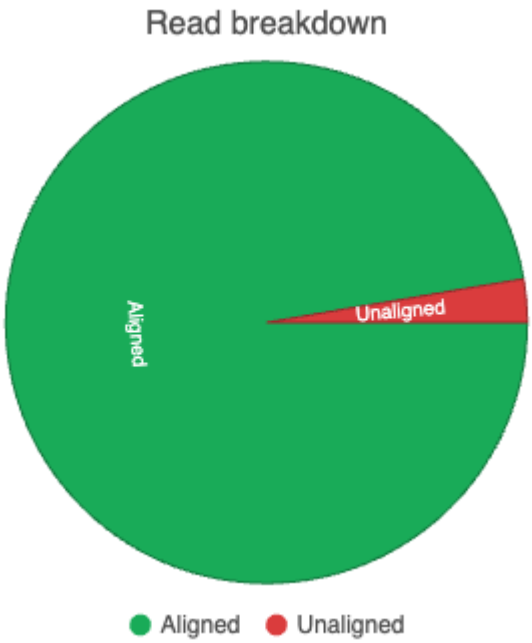
The next plot shows the alignments per read. Each sample has 1.99 alignments per read (close to two) because Partek Flow counts two alignments when both reads in pair map to the genome. Also, the alignments per read number in this dataset is not exactly two due to situations such as one read of the pair mapping.



Click on the individual samples to view its sample-level post-alignment QC results. The first plot shows the alignments per read for a sample. Most reads have two alignments in this particular example, which is ideal for paired end sequencing.



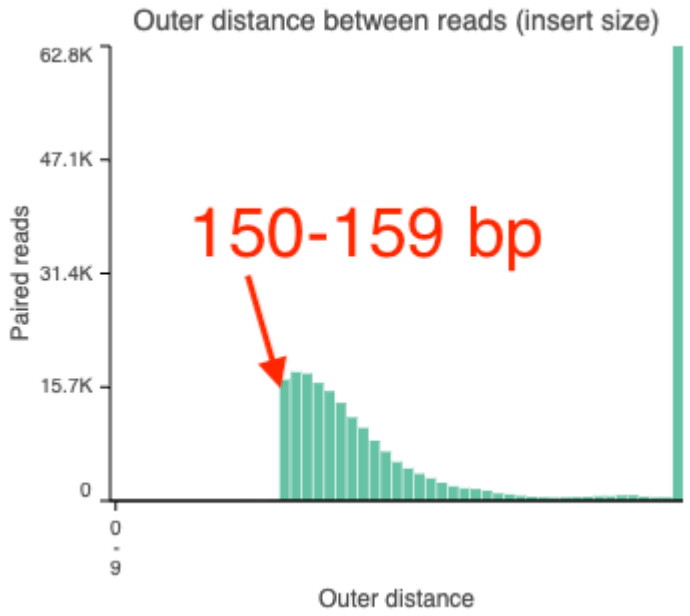
Next, a there is a pie chart that illustrates the portion of reads in a sample that aligned or unaligned.



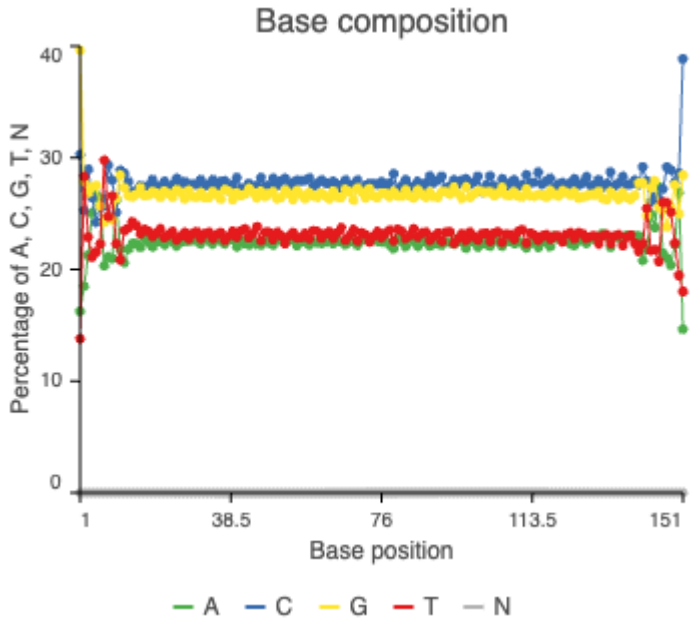
In paired end sequencing, the number of bases that span the 5' end of one read and the 5' end of another is known as the outer distance. This should be approximately equal to the nucleotide fragment length used in library preparation. Deviation of outer distance from expected could indicate the presences of structural variants such as insertions or deletions. Also, because the read length in this dataset is 151 bases, it will be expected that the two reads in the pair will overlap when aligned given the selected range for fragment lengths.

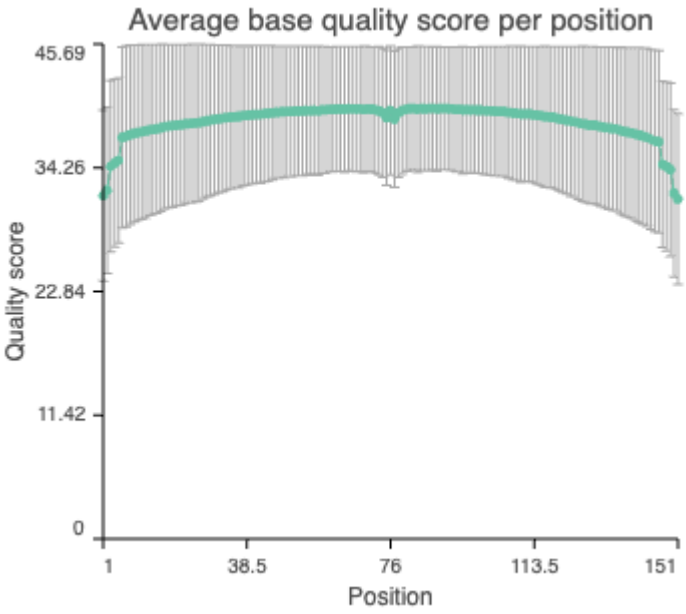
Tip

Read the article at https://www.cureffi.org/2012/12/19/forward-and-reverse-reads-in-paired-end-sequencing/ *(https://www.cureffi.org/2012/12/19/forward-and-reverse-reads-in-paired-end-sequencing/)* to get a basic idea of paired end sequencing.

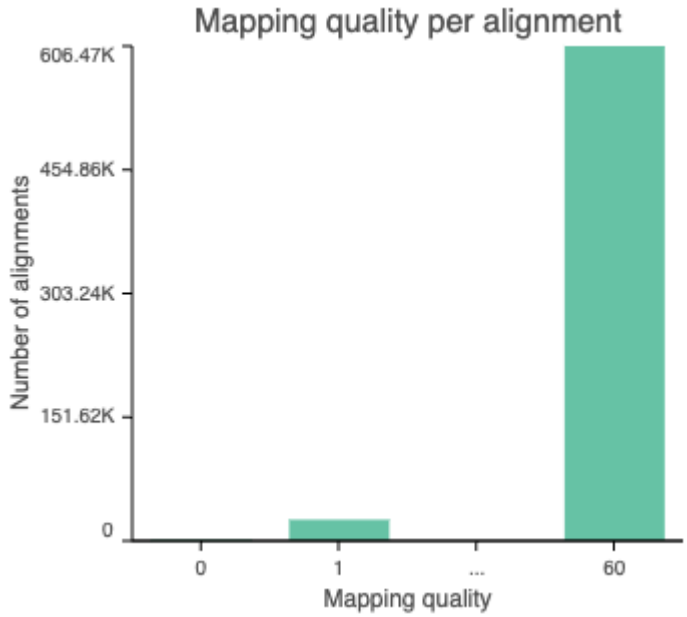


The base composition and read quality scores are also available in the sample level post-alignment QC report.

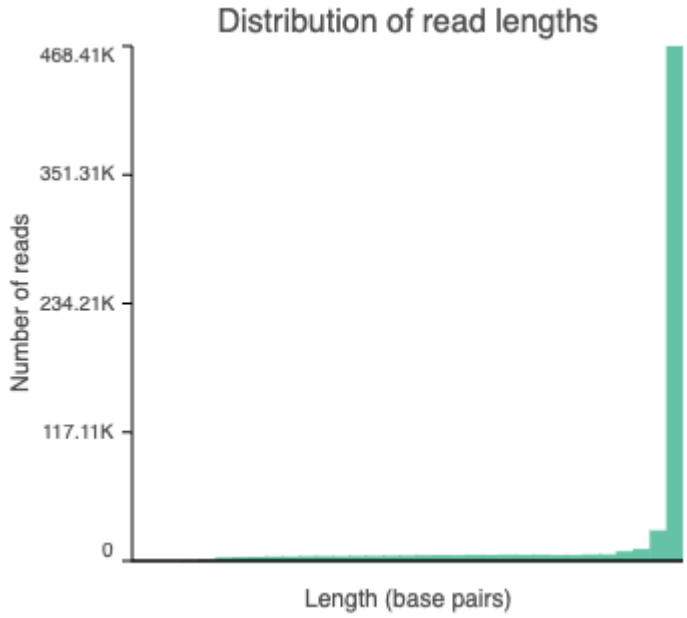




The confidence that a read was aligned or mapped to the correct location on the genome is another important post-alignment QC metric and is indicated by the mapping quality. The probability that a read was aligned incorrectly to a location on the genome can be estimated from the mapping quality through the equation below. Most reads in this dataset have a mapping quality of 60, which corresponds to 0.0001% error.



Finally, the length distribution for the aligned reads is also provided in the sample-level post-alignment QC report.

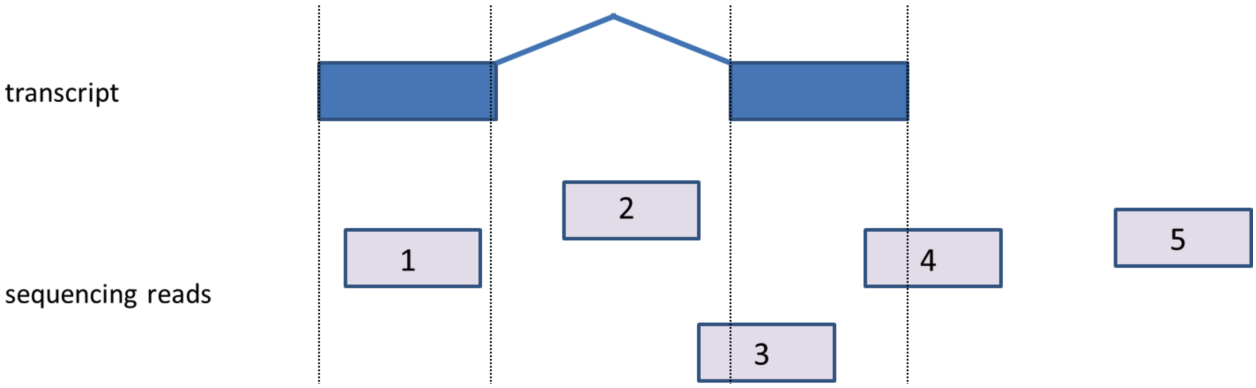


Generate Gene Expression Counts

Gene expression table can be generated from the read alignment. Options for generating an expression table. Because there a GTF annotation file is avaialable, this exercise will use the Quantify to annotation model (Partek E/M) tool, although others are available (see Partek Flow documents *(https://documentation.partek.com/display/FLOWDOC/Quantification)* to learn more).

Quantify to annotation model (Partek E/M) uses a statistical algorithm to determine how to assign multi-mapping reads to genomic features and avoids discarding these reads. When running this module, make sure that the "Strict paired-end compatibility" and "Require junction reads to match introns" options are checked.

- "Strict paired-end compatibility": In the case of paired end sequencing, this option tells Partek Flow to count only when both reads in the pair align to a transcript.
- "Require junction reads to match introns": This options deals with scenarios 3 and 4 in the image below where a part of the read maps to the intron. When checked, this option counts only when the intronic portion of the read matches the intron on the reference.



Source: https://documentation.partek.com/display/FLOWDOC/Understanding+Reads+in+RNA-Seq+Analysis *(https://documentation.partek.com/display/FLOWDOC/Understanding+Reads+in+RNA-Seq+Analysis)*

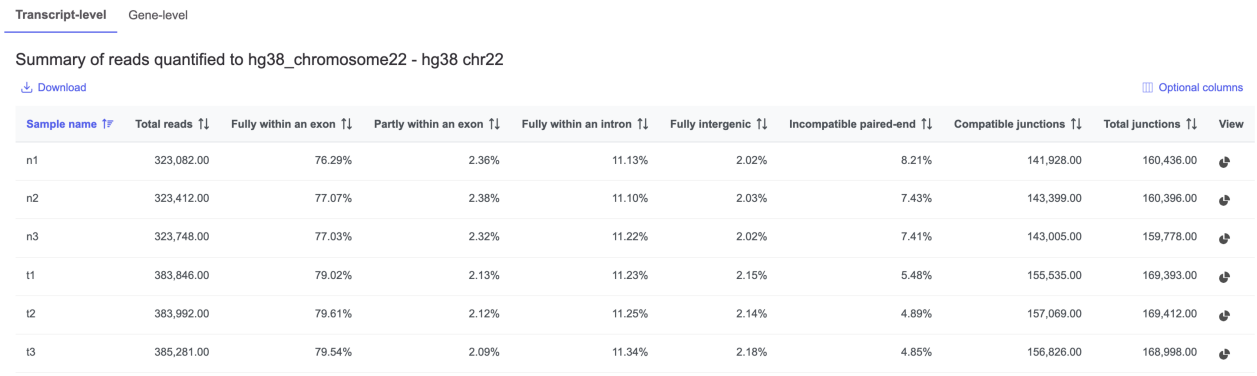
Users can also control the amount of overlap that a read has to a genomic feature (ie. gene, transcript) for it to count. Finally, the "Filter features" option allow users to remove genes or transcripts where the read counts across all samples are less than a specified threshold. This helps with filtering out low expression genes.

Under advanced options, select "auto-detect" if users do not know the strand specificity of the RNA sequencing experiment protocol.

Warning

Specifying the corrected strandedness used in a RNA sequencing experiment helps to avoid miscounting or counting the wrong gene or transcript. See https://chipster.csc.fi/manual/library-type-summary.html *(https://chipster.csc.fi/manual/library-type-summary.html)* to learn more.

The quantification step generates gene-level and transcript-level expression estimates, thus two data nodes appear upon completion of this task. These exercise will use the transcript-level data for differential expression analysis and gene-level data for GSEA. Clicking on the gene-level expression data node will pull up a summary about the quantification. The first tables shows the percentages of reads that overlapped exons, introns, and intergenenic regions, etc (click on the icon under the "View" column to view the break down of overlap types for each sample).



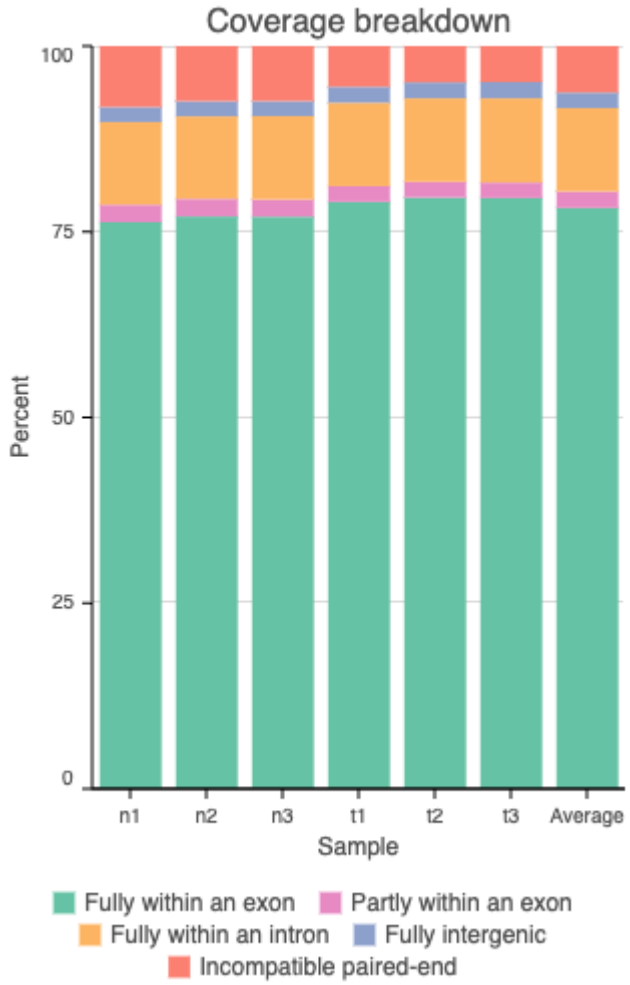
Transcript-level | Gene-level

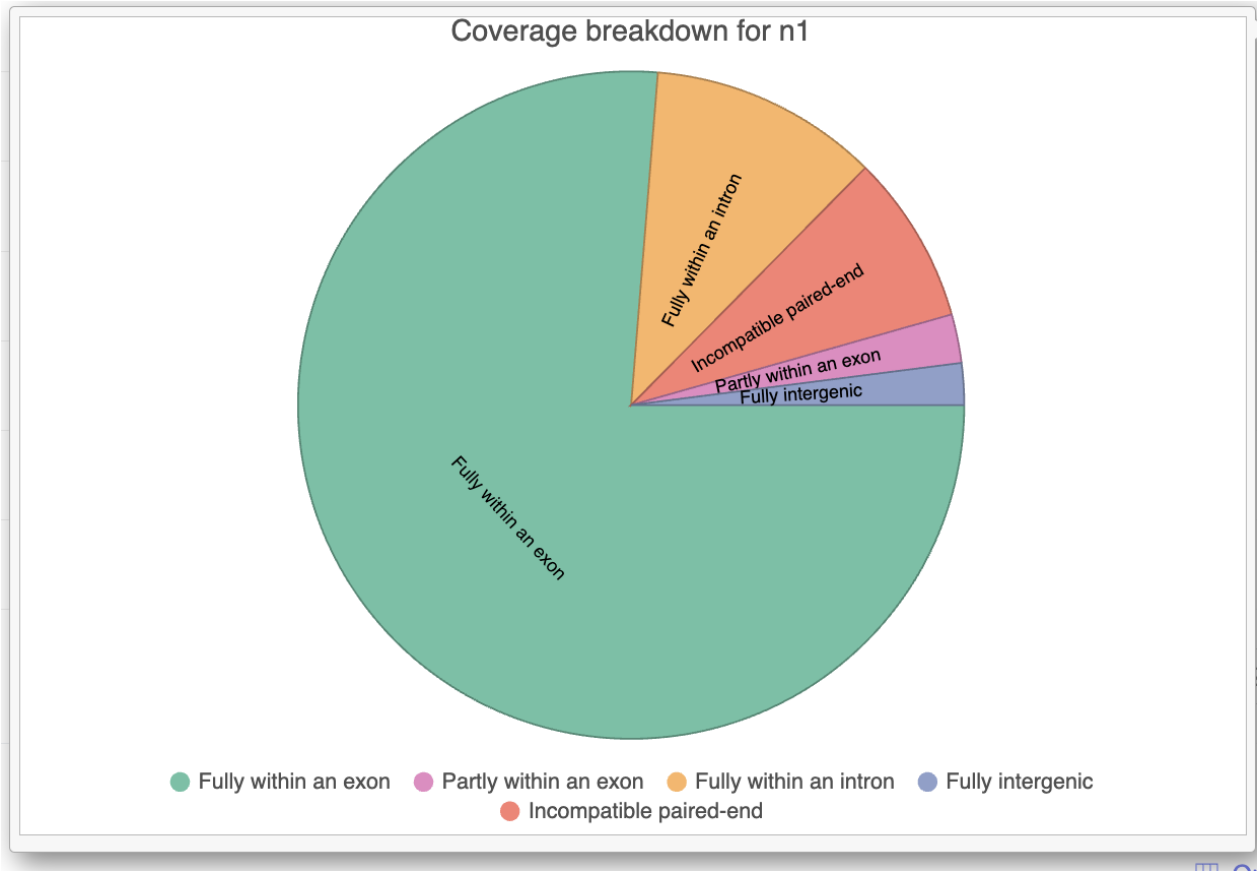
Summary of reads quantified to hg38_chromosome22 - hg38 chr22

Download | Optional columns

Sample name	Total reads	Fully within an exon	Partly within an exon	Fully within an intron	Fully intergenic	Incompatible paired-end	Compatible junctions	Total junctions	View
n1	323,082.00	76.29%	2.36%	11.13%	2.02%	8.21%	141,928.00	160,436.00	
n2	323,412.00	77.07%	2.38%	11.10%	2.03%	7.43%	143,399.00	160,396.00	
n3	323,748.00	77.03%	2.32%	11.22%	2.02%	7.41%	143,005.00	159,778.00	
t1	383,846.00	79.02%	2.13%	11.23%	2.15%	5.48%	155,535.00	169,393.00	
t2	383,992.00	79.61%	2.12%	11.25%	2.14%	4.89%	157,069.00	169,412.00	
t3	385,281.00	79.54%	2.09%	11.34%	2.18%	4.85%	156,826.00	168,998.00	

The information in this summary table is also presented as stacked bar chart as shown below.


Coverage breakdown
100
75
50
25
0
Percent
n1
n2
n3
t1
t2
t3
Average
Sample
Fully within an exon
Partly within an exon
Fully within an intron
Fully intergenic
Incompatible paired-end


Coverage breakdown for n1
Fully within an intron
Incompatible paired-end
Partly within an exon
Fully intergenic
Fully within an exon
Fully within an exon
Partly within an exon
Fully within an intron
Fully intergenic
Incompatible paired-end

The next table shows the expression distribution information such as min, max, mean, median, 25th percentile (Q1), and 75th percentile (Q3).

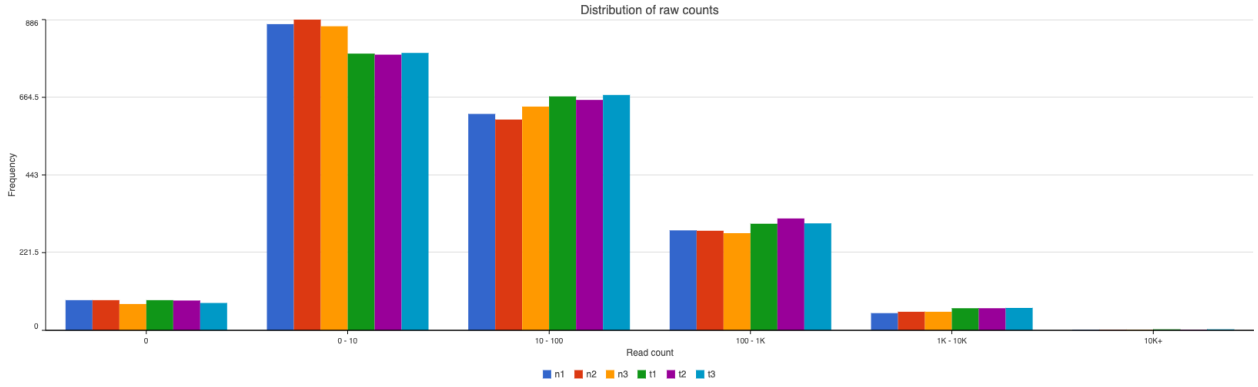
Feature distribution (6 samples; 1,912 transcripts)

Download

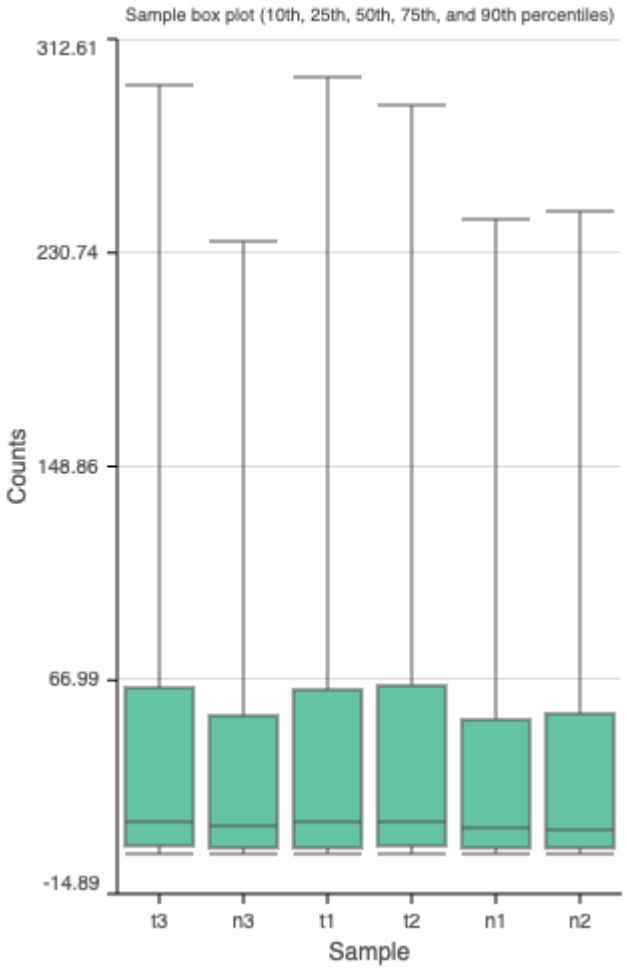
Optional columns

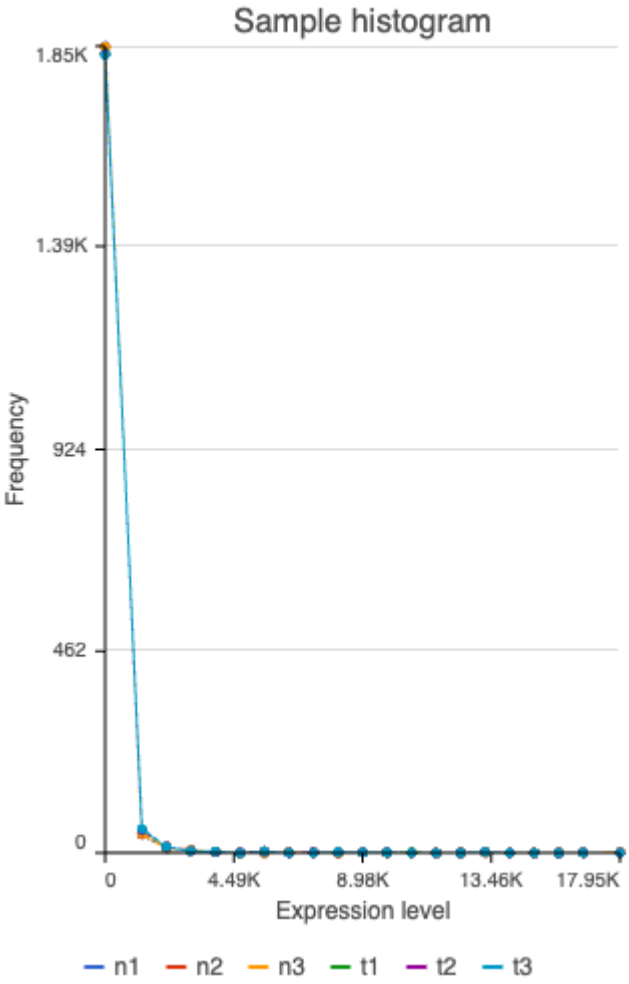
Sample name	Min	2nd min	Max	Mean	Median	Q1	Q3
n1	0	4.19E-38	17,945.20	128.57	10.00	2.15	51.11
n2	0	1.79E-43	17,711.30	130.03	9.60	2.00	53.70
n3	0	1.43E-43	17,883.80	130.10	10.51	2.30	52.97
t1	0	1.94E-40	13,646.70	158.25	12.61	2.74	63.21
t2	0	1.54E-41	13,577.30	159.49	12.70	3.00	64.31
t3	0	1.36E-37	13,605.60	159.87	12.68	3.00	63.98
All samples	0	1.43E-43	17,945.20	144.39	11.19	2.53	57.20

A histogram showing the distribution of expression estimates is available as well. Across all samples, most genes have expression counts of between 0-100 although there are some high expressing genes that have counts of between 1000-10000.



The distribution of expression estimates for samples in this dataset are shown as box and density plots.


Sample box plot (10th, 25th, 50th, 75th, and 90th percentiles)
312.61
230.74
148.86
66.99
-14.89
Counts
t3
n3
t1
t2
n1
n2
Sample


Sample histogram
1.85K
1.39K
924
462
0
Frequency
0
4.49K
8.98K
13.46K
17.95K
Expression level
n1
n2
n3
t1
t2
t3

Normalizing Gene Expression Estimation

Normalization of gene expression estimates obtained from the quantification step is important as this will remove technical or non-biological variants in the data such as:

- Differences in sequencing depth between samples (ie. not all samples have the same number of reads or sequences).
- RNA composition variations among samples (ie. samples do not have the same RNA expressed in the case where comparison of transcriptome is done between tissue from different organs or perhaps differing biological conditions such as tumor versus normal).
- Gene length (longer genes will have more reads mapping to them).
- GC content.

Ultimately, the goal is the eliminate technical variations in the sequencing experiment so that the scientist can be left with the biological variations, which are of interest.

When doing differential expression analysis between biological conditions, only the first two technical variations mentioned above are important. This is because it can be assumed that when comparing expression of the same gene or transcript between conditions, that the length and GC content would remain the same. While there are many normalization techniques available in Partek Flow (see https://documentation.partek.com/display/FLOWDOC/Normalization *(https://documentation.partek.com/display/FLOWDOC/Normalization)*), this class will use the median ratio (DESeq2 only) method as it will normalize for sequencing depth and RNA composition.

Click on the transcript normalized estimates data node to view the summary for this step in the analysis.

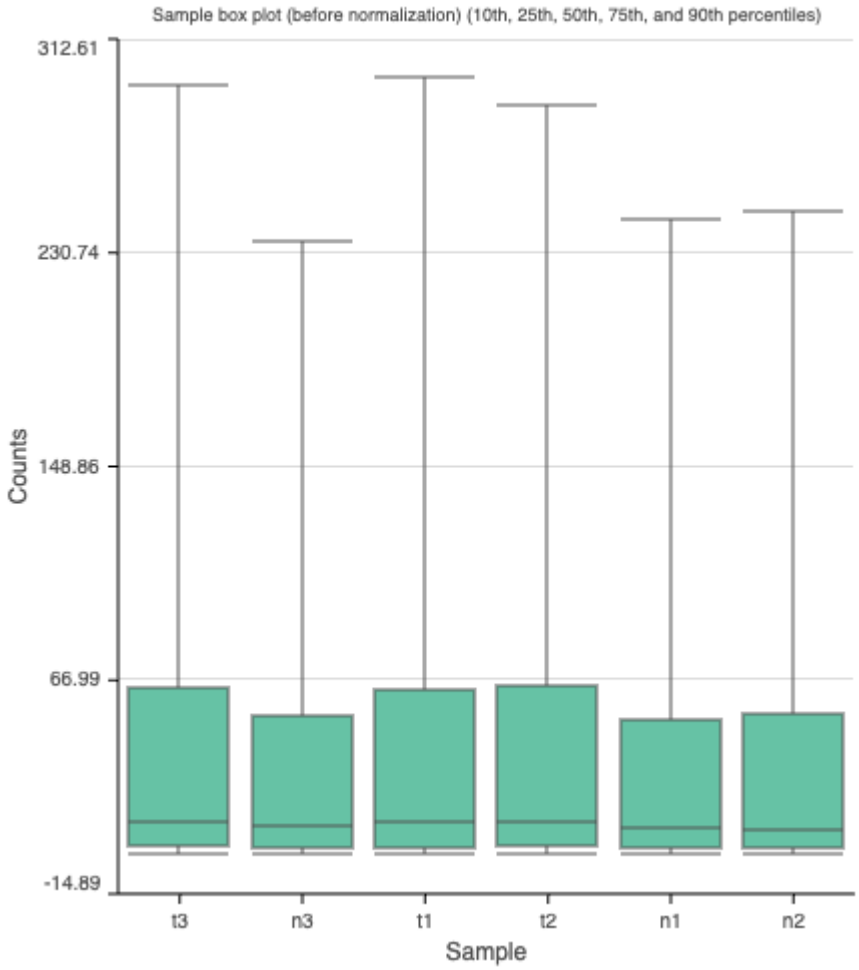
The distribution, minimum, maximum, mean, median, 25th percentile (Q1), and 75th percentile (Q3) of expression estimates are presented as a table as well as box and density plots. Note that users can compare between the pre- and post- normalized expression count box and density plots. Normalization resulted in expression estimate distribution for all samples to roughly overlap.

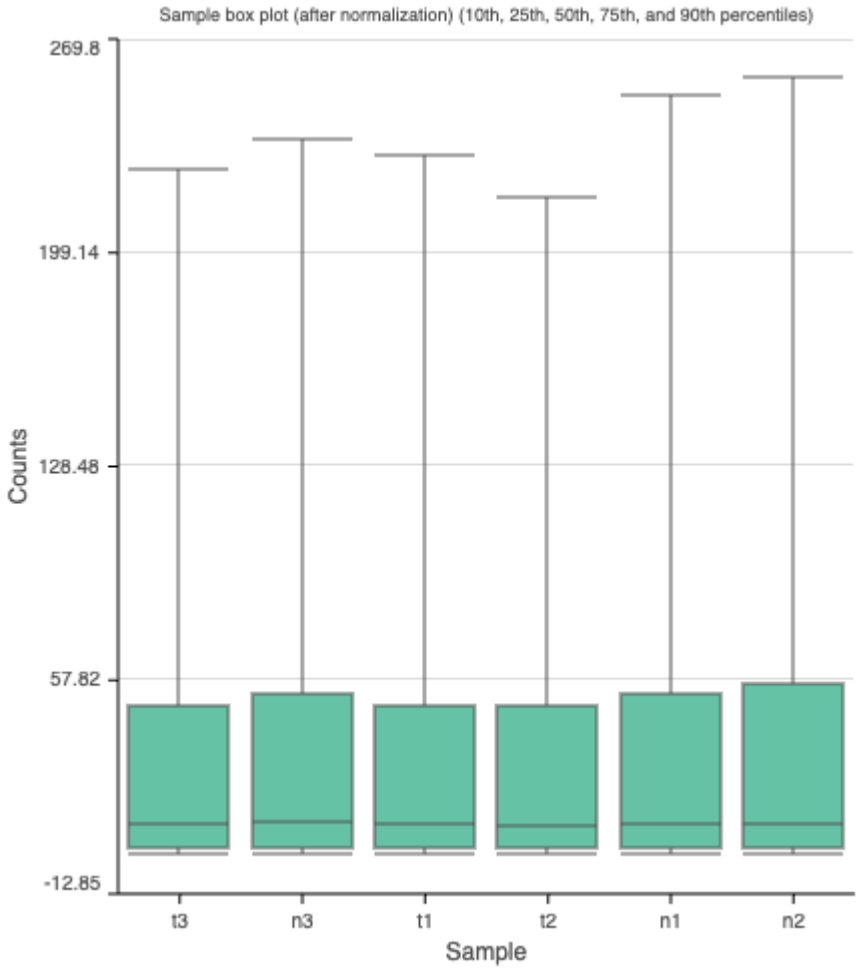
Normalization methods Median ratio (DESeq2 only)

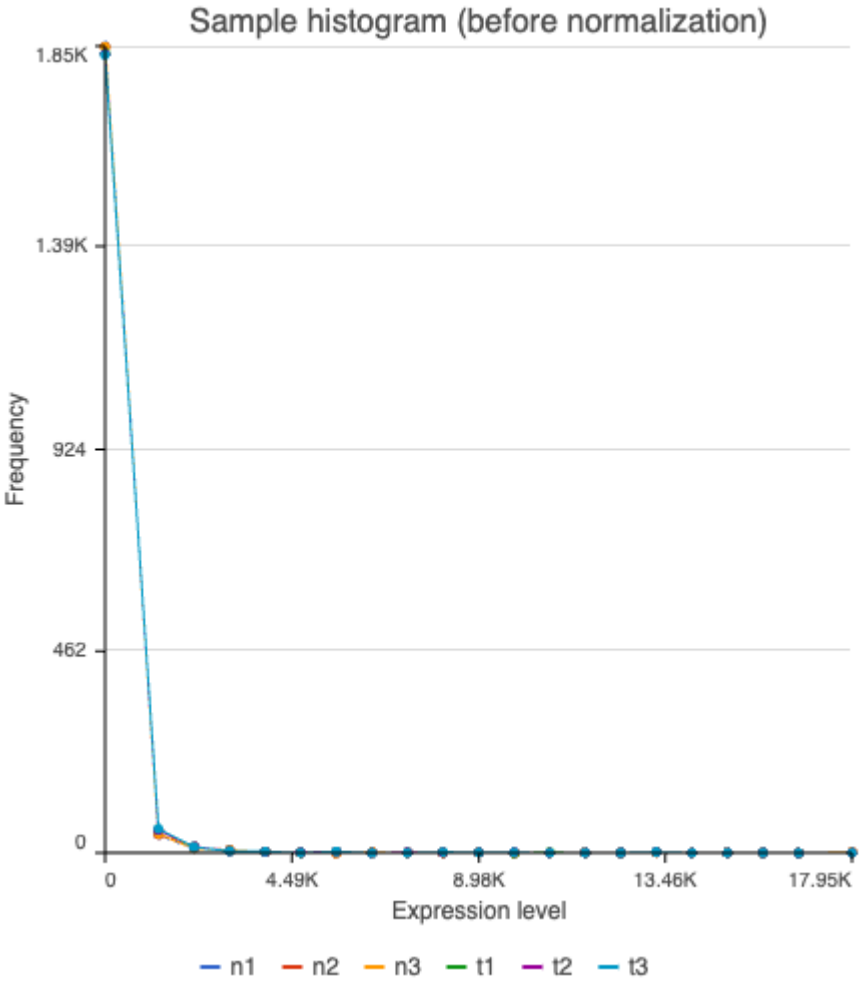
Feature distribution (6 samples; 1,912 transcripts)

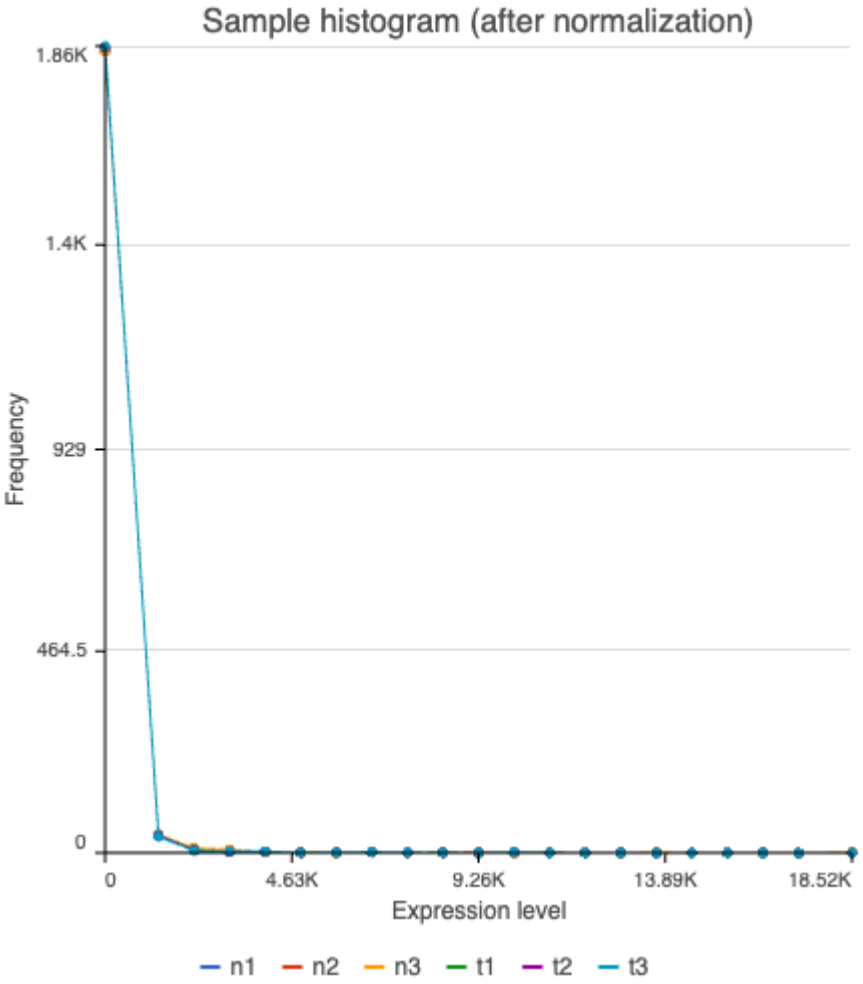
Download Optional columns

Sample name	Min	2nd min	Max	Mean	Median	Q1	Q3	Missing
n1	0	4.32E-38	18,512.70	132.63	10.32	2.22	52.72	0
n2	0	1.87E-43	18,476.20	135.65	10.01	2.09	56.02	0
n3	0	1.44E-43	18,003.60	130.97	10.58	2.32	53.33	0
t1	0	1.51E-40	10,601.60	122.94	9.80	2.13	49.10	0
t2	0	1.17E-41	10,295.90	120.95	9.63	2.27	48.76	0
t3	0	1.05E-37	10,469.00	123.01	9.76	2.31	49.23	0


Sample box plot (before normalization) (10th, 25th, 50th, 75th, and 90th percentiles)
312.61
230.74
148.86
66.99
-14.89
Counts
t3
n3
t1
t2
n1
n2
Sample


Sample box plot (after normalization) (10th, 25th, 50th, 75th, and 90th percentiles)
269.8
199.14
128.48
57.82
-12.85
Counts
t3
n3
t1
t2
n1
n2
Sample


Sample histogram (before normalization)
1.85K
1.39K
924
462
0
Frequency
0
4.49K
8.98K
13.46K
17.95K
Expression level
n1
n2
n3
t1
t2
t3

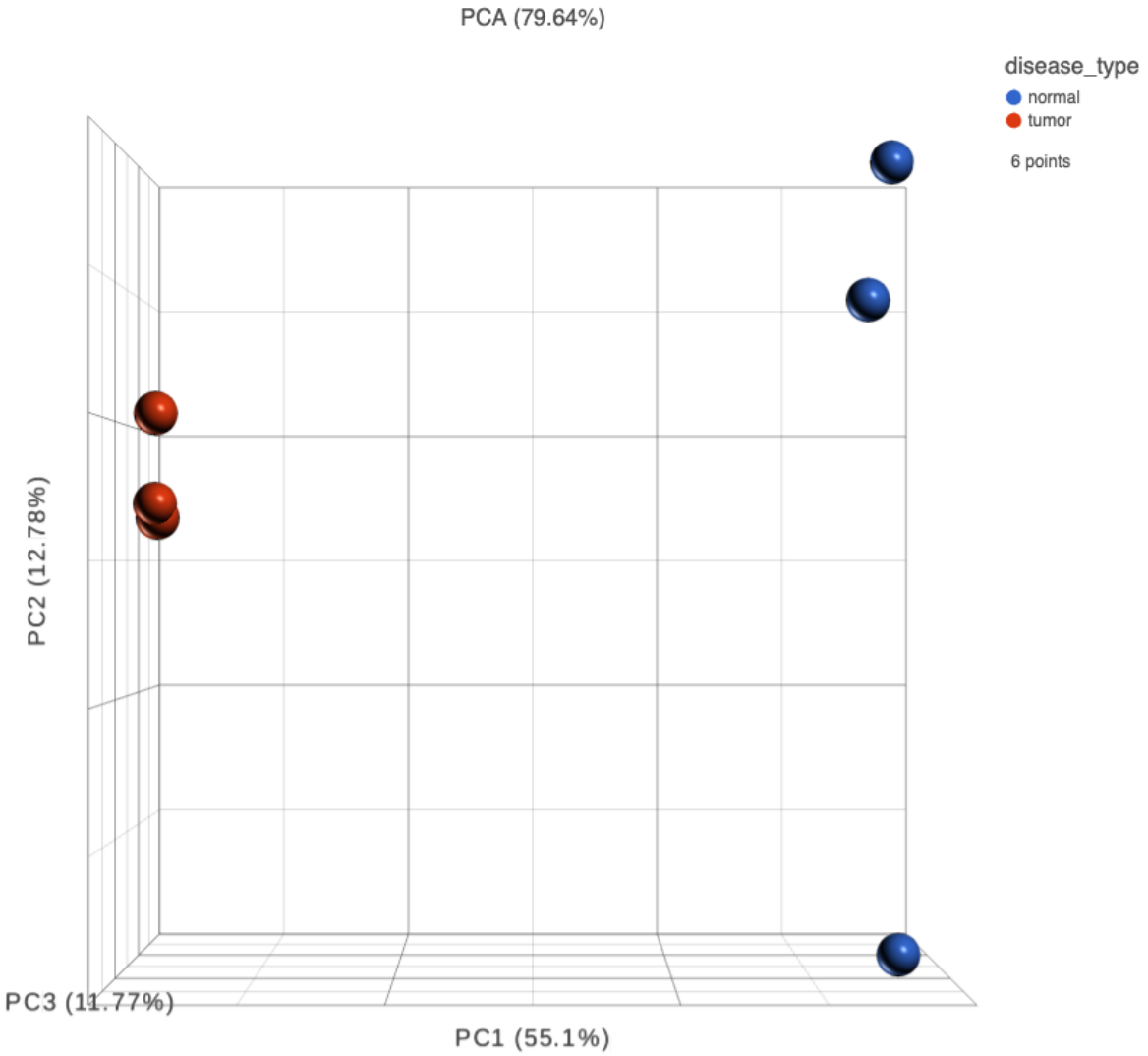

Sample histogram (after normalization)
1.86K
1.4K
929
464.5
0
Frequency
0
4.63K
9.26K
13.89K
18.52K
Expression level
n1
n2
n3
t1
t2
t3

Principal Components Analysis

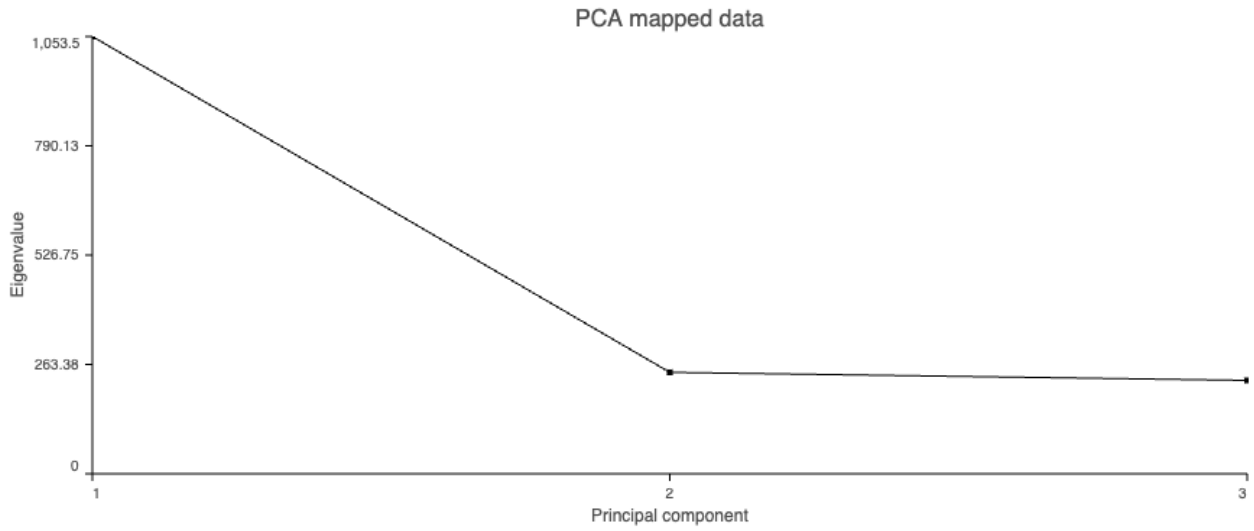
Principal Components Analysis (PCA) transforms high dimensional data such as those derived from RNA sequencing so that researchers can see how study variables cluster together. The result of PCA is that the original data is projected onto a set a perpendicular axes where each axis accounts for a percentage of variance in the data. To learn the math behind PCA, see https://www.iro.umontreal.ca/~pift6080/H09/documents/papers/pca_tutorial.pdf *(https://www.iro.umontreal.ca/~pift6080/H09/documents/papers/pca_tutorial.pdf)*.

PCA is an excellent quality assurance tool for RNA sequencing analysis as the results when plotted enable scientists to determine if samples in the same biological condition cluster together. Click on the "Normalized counts" data node and select exploratory analysis from the menu. From there, select PCA. In the subsequent PCA configuration page, lower the number of dimensions to 3 since a 3D plot is the most that can visualized.

Clicking on the PCA data node will reveal two plots and a table. First, there is an interactive three dimensional PCA plot where the axis PC1 (ie. principal component 1) accounts for the highest variance in the data (55.1%). As hoped, the normal and tumor samples are separated along this axis indicating that it is the biology (ie. normal or tumor) that differentiates the samples. The PC2 and PC3 account the second and third highest variance in the data and samples within each group are separated along these two axes suggesting that there may be differences between samples from the same condition or the existence of batch effects. Together, PC1, PC2, and PC3 explain 79% of the variation in this dataset. Here, with just 3 dimensions, scientists can visualize and interpret the data and thus, PCA is known a dimensionality reduction procedure as it reduces high dimensional data into the most relevant dimensions while enabling interpretation and conclusions to be drawn.



A scree plot showing the variance accounted for by each principal component axis is also available.



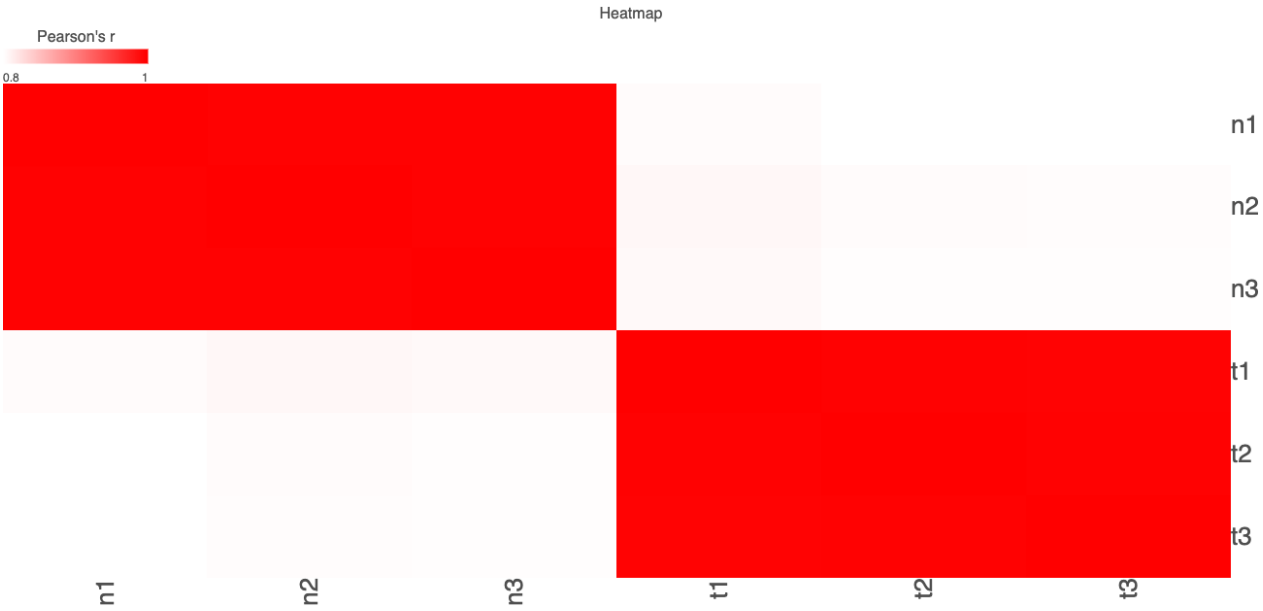
The table labeled "Component loadings" shows how the transcripts listed influence the separation of the samples along the three principal component axes.

Component loadings

Variable	PC 1	PC 2	PC 3
ENST00000006251.11>ENSG0000018...	0.95	0.22	0.21
ENST00000008876.7>ENSG00000008...	-0.77	-0.39	0.22
ENST00000043402.7>ENSG00000040...	-0.85	-0.037	-0.12
ENST00000155674.9>ENSG00000069...	0.85	-0.35	0.058
ENST00000207636.9>ENSG00000093...	0.48	0.55	0.32
ENST00000215659.12>ENSG0000018...	0.96	-0.19	0.0094
ENST00000215730.11>ENSG0000009...	0.97	-0.2	-0.09
ENST00000215739.12>ENSG0000009...	0.98	-0.19	-0.093
ENST00000215742.8>ENSG00000184...	0.1	0.47	-0.36
ENST00000215743.7>ENSG00000099...	-0.94	0.041	0.076
ENST00000215754.7>ENSG00000240...	0.99	-0.11	0.01
ENST00000215770.5>ENSG00000099...	-0.6	0.38	-0.64
ENST00000215781.2>ENSG00000099...	0.99	0.025	0.12
ENST00000215790.11>ENSG0000009...	0.71	-0.1	0.069
ENST00000215793.12>ENSG0000009...	0.99	-0.096	0.046
ENST00000215794.7>ENSG00000184...	1	0.06	-0.017
ENST00000215798.10>ENSG0000009...	-0.85	0.14	-0.2
ENST00000215829.7>ENSG00000100...	-0.9	-0.41	-0.058
ENST00000215832.10>ENSG0000010...	0.99	-0.16	-0.0053
ENST00000215838.7>ENSG00000185...	0.55	0.76	0.33
ENST00000215862.8>ENSG00000133...	-0.37	0.043	0.33
ENST00000215882.9>ENSG00000100...	0.96	0.24	0.086
ENST00000215904.6>ENSG00000241...	0.95	0.074	0.28
ENST00000215906.5>ENSG00000128...	-0.48	-0.4	0.61
ENST00000215909.9>ENSG00000100...	-1	-0.058	-0.017

Correlation Plot

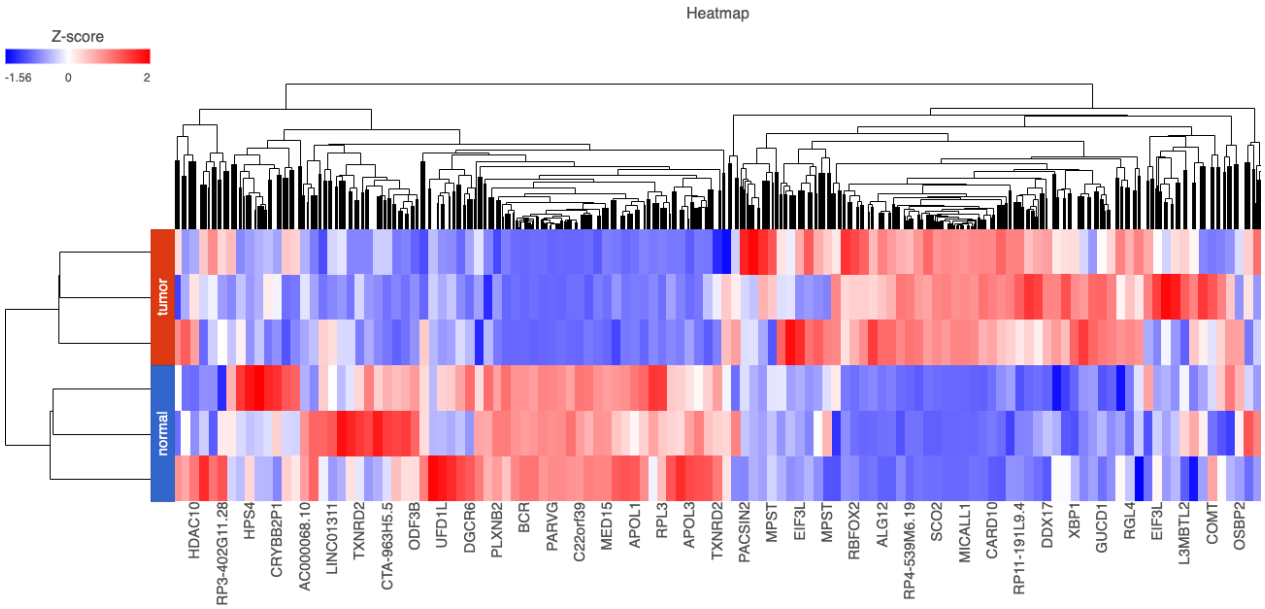
Another quality assurance measure is to determine correlation between samples. It is expected that samples from the same biological condition will be highly correlated to each other based on gene expression. As expected, correlation plot below shows that normal samples are correlated with one another while not correlated with the tumor samples.



To generate the above correlation plot, click on the "Normalized counts" data node and then "Statitics". From there, choose "Correlation". In the subsequent menu, choose "Similarity matrix" and select "Samples" as the goal is to determine how well the samples correlate with one another. In the next page, under "Calculate for" choose Pearson correlation. This will enable the calculation of how similar the samples are as a result of gene expression.

Expression Heatmap

Heatmap and dendrogram can reveal clusters of genes whose expression is up or down-regulated under certain biological conditions and from the visualization below, it is clear that there are panels of genes that are upregulated in tumor but not normal samples and those that are upregulated in normal samples but not tumors.



Filtering Normalized Expression Estimates

Prior to differential expression analysis using the transcript-level expression data, filtering is recommended to remove low expressing transcripts as these may be noise. Several filter options are available in Partek Flow, please refer to the "Filter features" *(https://documentation.partek.com/display/FLOWDOC/Filter+features)* section of the Partek Flow documentations to learn more. In this class, the "Noise reduction filter" will be used and will remove transcripts whose sum of expression across samples is less than or equal to 3.

Gene Set Enrichment Analysis

Definition

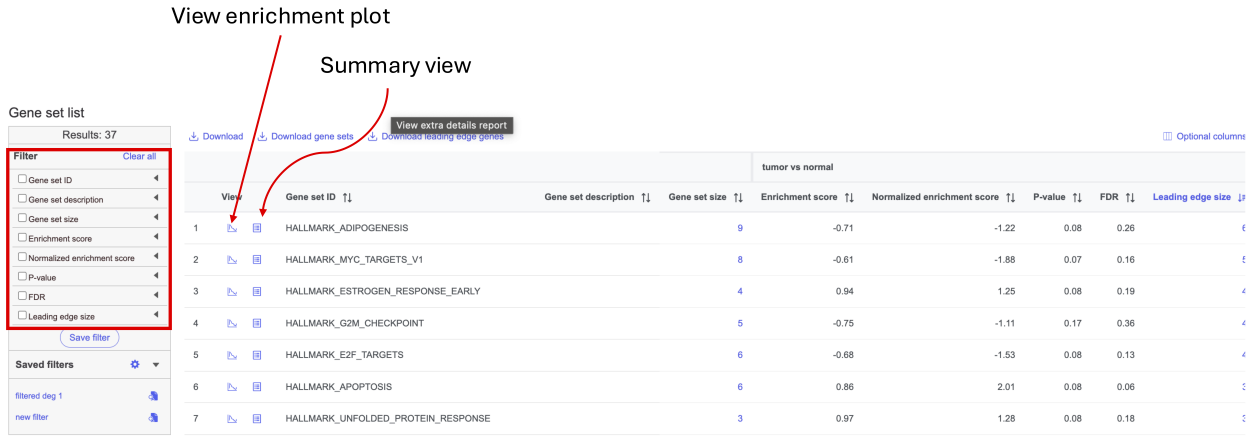
"The goal of GSEA is to determine whether members of a gene set S tend to occur toward the top (or bottom) of the list L, in which case the gene set is correlated with the phenotypic class distinction." -- Gene set enrichment analysis A knowledge-based approach for interpreting genome-wide expression profiles *(https://www.pnas.org/doi/epdf/10.1073/pnas.0506580102)*

The input for GSEA is the normalized gene expression matrix. In this example the hallmark dataset *(https://www.gsea-msigdb.org/gsea/msigdb/index.jsp)* will be used.

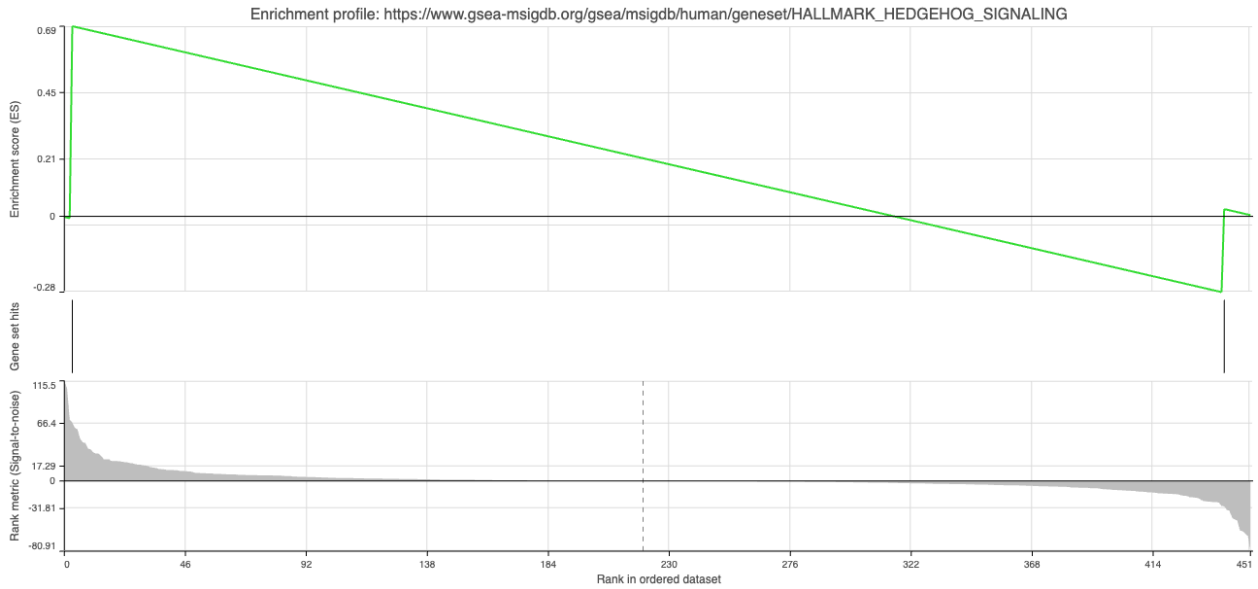
Tip

A positive enrichment score indicates that the gene set is enriched in the condition entered as the numerator during setup. A negative enrichment score on the other hand, indicates that the gene set is enriched in the condition entered as the denominator during setup.

Click on the GSEA data node to view results table after this task is completed. In this table, users can invoke enrichment view and summaries for each gene set as well as filter results.



The enrichment plot for the hedgehog signaling gene set is shown below and it indicates that this is enriched in the tumor samples (normalized enrichment score of 2.04).



The enrichment summary report reveals the genes in the hedgehog gene set that occur in the ranked expression list, with the leading edge gene being CELSR1s.

Feature information

Gene set ID	HALLMARK_HEDGEHOG_SIGNALING	**Gene set size**	2
Gene set description	https://www.gsea-msigdb.org/gsea/msigdb/human/geneset/HALLMARK_HEDGEHOG_SIGNALING		

tumor vs normal

Enrichment score	0.69	**FDR**	0.10
Normalized enrichment score	2.04	**Leading edge size**	1
P-value (F)	0.08		

Leading edge genes

Download list

CELSR1

Genes in list

Download list

CELSR1
MYH9

Refer to the following Partek Flow *(https://documentation.partek.com/display/FLOWDOC/GSEA)* to learn about the interpretation of GSEA results.

- Enrichment score. The algorithm walks down the ranked list of all the genes in the model, increasing the running sum (y axis) each time when a gene in the current gene set is encountered. Conversely, the running-sum is decreased each time a gene not in the current gene set is encountered. The magnitude of the increment depends on the correlation of the gene with the experimental factor. The enrichment score is then the maximum deviation from zero encountered in the random walk (the summit of the curve).
- Gene set hits. Each column shows the location of a gene from the current gene set, within the ranked list of all the genes in the model.

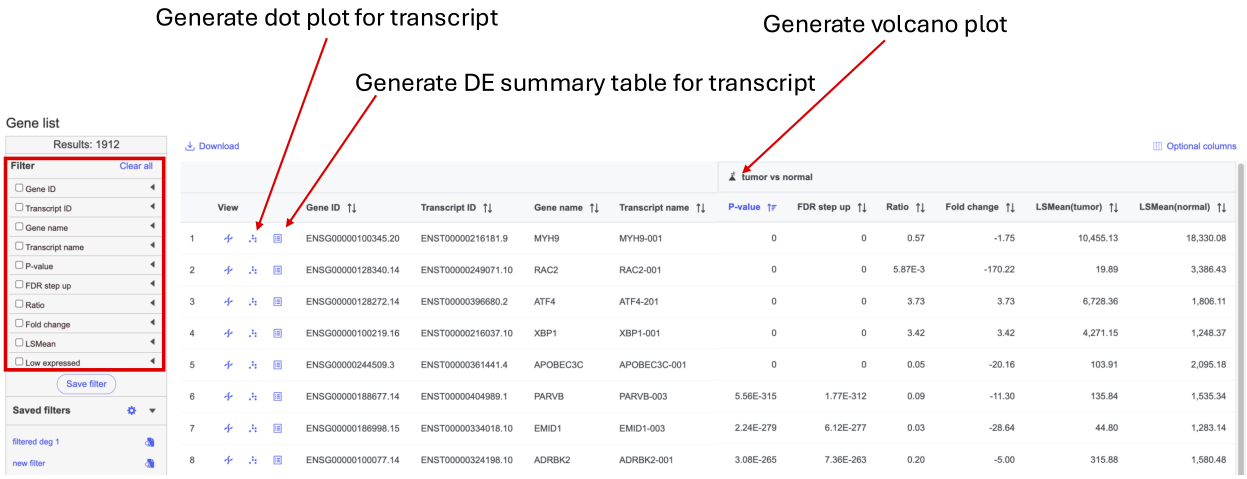
- Rank metric. The plot shows the value of the ranking metric (y axis) as you move down the ranked list of all the genes in the model (x axis). The ranking metric measures a gene's correlation with a phenotype. A positive value of the metric indicates correlation with the first category (Numerator) and a negative value indicates correlation with the second category (Denominator).

Differential Expression Analysis

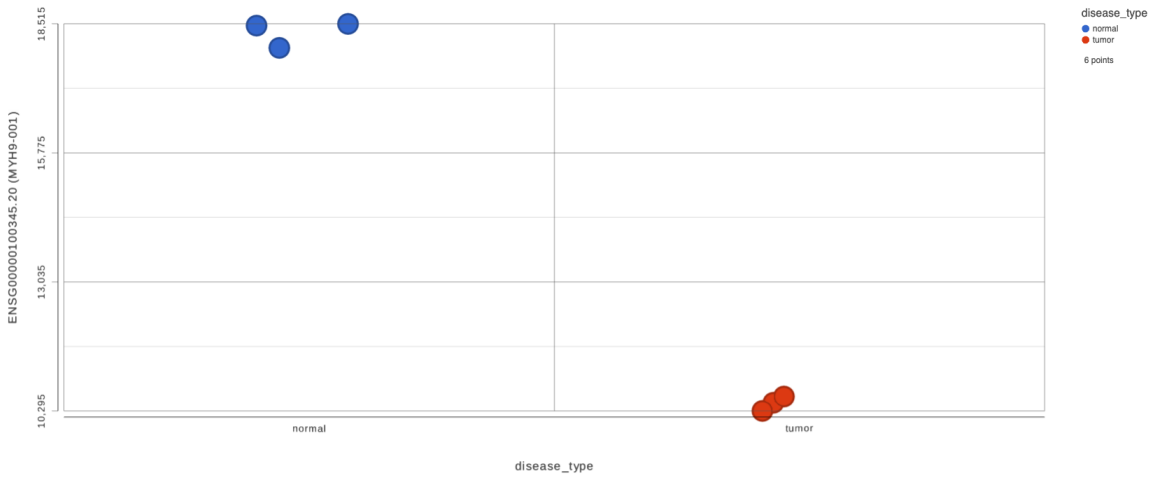
After generating and filtering out lower expressing genes from the median ratio normalized expressions data, it is time to perform differential expression analysis to see if there are genes or transcripts (trancripts will be used here) that are statistically significantly up or down-regulated between biological conditions (in this case tumor versus normal).

Refer to the "Differential Analysis" section of the Partek Flow documentations *(https://documentation.partek.com/display/FLOWDOC/Differential+Analysis)* to learn about the options for performing this task but in this example, DESeq2, which is Partek's implementation of the DESeq2 R package will be used. The video below shows the steps for completing differential expression analysis, constructing a volcano plot, and filtering out the differential expression results for use with over representation analysis.

After differential expression analysis is completed, a data node is generated. Click on it to review the results table. On the left, there is a panel where users can filter the differential expression results based on criteria such as false discovery rate (FDR) and fold change.



Under the "View" column of the differential expression results table, researchers can obtain a dot plot of the expression for the corresponding transcript or gene across all samples as well as a summary of the differential analysis results for that particular transcript or gene.


disease_type
normal
tumor
6 points
ENSG00000100345.20 (MYH9-001)
18,515
15,775
13,035
10,295
normal
tumor
disease_type

Feature information

Gene ID	ENSG00000100345.20	**Strand**	-
Transcript ID	ENST00000216181.9	**Total counts**	86,359.13
Chromosome	22	**Maximum counts**	18,512.74
Start	36,281,281	**Geometric mean**	13,843.03
Stop	36,388,019	**Arithmetic mean**	14,393.19
Length	7,501		

Model information

Model	disease_type	**AICc**	0
Model type	DESeq2		

tumor vs normal

P-value (Wald)	0	**LSMean(tumor)**	10,455.13
FDR step up	0	**LSMean(normal)**	18,330.08
Ratio	0.57	**FC 95% lower limit**	-1.80
Log2(Ratio)	-0.81	**FC 95% upper limit**	-1.71
Fold change	-1.75		

Least-Squares Mean information

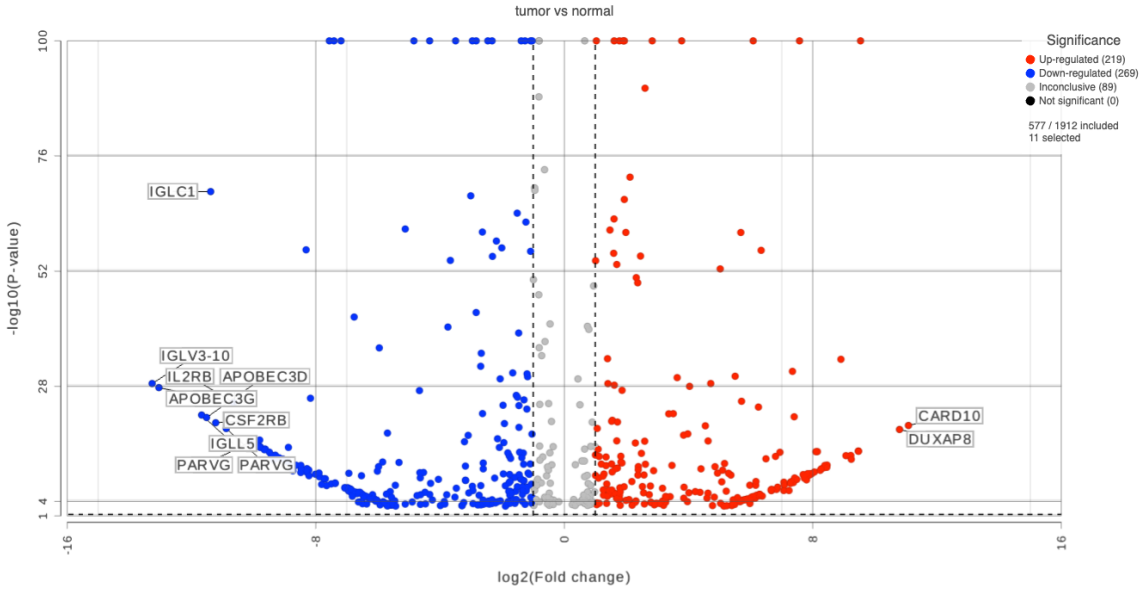
disease_type	LSMean
normal	18,330.08
tumor	10,455.13

Other columns include

- `Gene ID`

- `Transcript ID`
- `Gene name`
- `Transcript name`
- `P-value`
- `False discovery rate`
- `Ratio`: expression ratio obtained by dividing the mean expression of a gene in one condition by the mean expression of a gene in another (ie. numerator/denomintor, in this case tumor/normal as set during configuration of differential expression analysis).
- `Fold change`: this is equal to ratios that are greater 1; when ratio is less than 1, then `Fold change` is equal to -1/ratio.
- `LSMean(tumor)`: mean expression for a given gene/transcript in the tumor samples
- `LSMean(normal)`: mean expression for a given gene/transcript in the normal samples

Click on the "volcano" on top of the differential expression results table to view a volcano plot. In RNA sequencing, the volcano plot displays log2 of fold change on the horizontal axis and -log10 of the p-value on the vertical axis. The plot below was filtered such that the points labeled correspond to genes whose log2 fold change value are less than -10 or greater than 10 while p-value is between 0 and 0.001.



Over Representation Analysis

In addition to GSEA, over representation analysis is another method for determining which molecular biology functions, component, etc. are affected as a result of gene expression between biological conditions. Essentially, statistics such as Fisher Exact test are used to determine whether the genes in a differential expression list occur in or overlap with those participating in specific biological functions by chance. To learn about over representation analysis, please refer to https://bioinformatics.ccr.cancer.gov/docs/btep-coding-club/CC2023/FunctionalEnrich_clusterProfiler/ *(https://bioinformatics.ccr.cancer.gov/docs/btep-coding-club/CC2023/FunctionalEnrich_clusterProfiler/)*.

The filtered differential analysis table will be used as input for over representation analysis (see video below for steps).

In the results table, users can click the link under the "Gene set" column to view the pathway along with information such as enrichment score, p-value, and false discovery rate (FDR).

Gene set	Description	Enrichment score	P-value	FDR step up	Rich factor	Genes in set	Genes in list	Genes not in list	Genes in list, not in set	Genes not in list, not in set
path:hsa05166	Human T-cell leukemia virus 1 infection	2.12	0.12	1.00	1.00	8	8	0	121	38
path:hsa04110	Cell cycle	1.85	0.16	1.00	1.00	7	7	0	122	38
path:hsa05170	Human immunodeficiency virus 1 infection	1.58	0.21	1.00	0.88	17	15	2	114	36
path:hsa04141	Protein processing in endoplasmic reticulum	1.58	0.21	1.00	1.00	6	6	0	123	38
path:hsa05203	Viral carcinogenesis	1.58	0.21	1.00	1.00	6	6	0	123	38
path:hsa04060	Cytokine-cytokine receptor interaction	1.58	0.21	1.00	1.00	6	6	0	123	38
path:hsa04666	Fc gamma R-mediated phagocytosis	1.31	0.27	1.00	1.00	5	5	0	124	38
path:hsa04142	Lysosome	1.31	0.27	1.00	1.00	5	5	0	124	38
path:hsa04144	Endocytosis	1.31	0.27	1.00	1.00	5	5	0	124	38

Definition

"Rich factor in the above table is the ratio of Genes in list divided by Genes in set." -- Partek Flow *(https://documentation.partek.com/display/FLOWDOC/Gene+Set+Enrichment)*