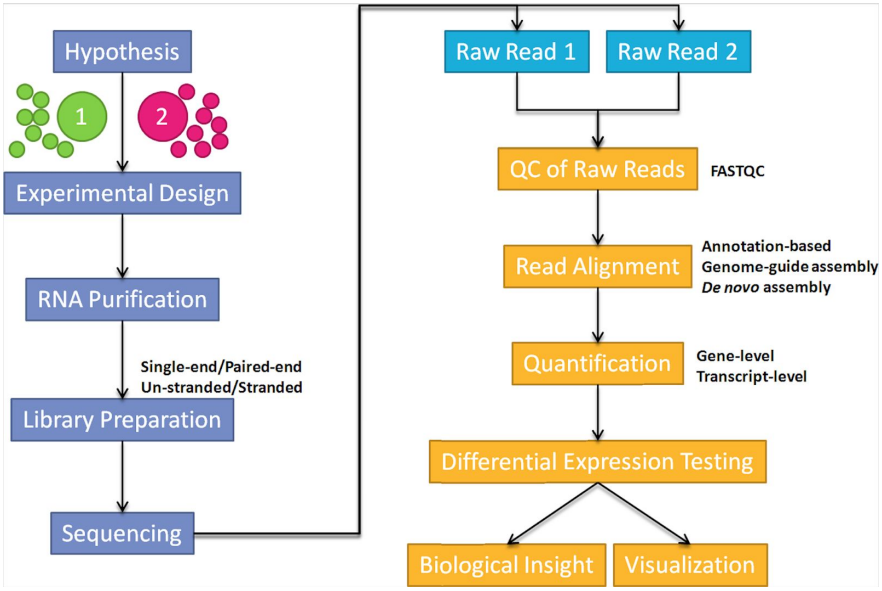
November, 2018

FASTQC Analysis & HISAT Alignments Using CyVerse (part 2)

Ray A Enke

RNA-Seq Analysis Bioinformatics Pipeline
Part 2: FastQC & HISAT analysis
Dr. Enke Bio 481 Genomics

Over the next few class meetings you will learn how to perform several steps in a RNA-sequencing (RNA-seq) bioinformatics data analysis pipeline using unpublished Illumina NGS dataset from my lab. We will use a suite of free tools provided by the virtual organization **CyVerse** to analyze the data. The steps we will use are a general pipeline that can be applied to other eukaryotic RNA-seq data sets. The figure below illustrates all of the steps in an RNA-seq analysis pipeline.



RNA-seq Bioinformatics steps:

1. QC of raw reads
2. Read alignment to reference
3. Transcript quantification and differential expression testing
4. Data visualization

Figure 1: RNA-seq analysis overview. Steps on the left in blue illustrate RNA sampling from 2 experimental groups in replicate followed by cDNA library preparation and Next-Generation Sequencing. Millions of raw sequencing reads from each sample in replicate are fed into a bioinformatics analysis pipeline represented in yellow on the right providing differential gene expression analysis between sample groups 1 and 2 (image taken from https://ycl6.gitbooks.io/rna-seq-data-analysis/rna-seq_analysis_workflow.html)

RNA-seq Dataset Overview: The Enke lab studies photoreceptor (PR) gene regulation in the human retina. Post-mortem human retina can be dissected to collect the central macular retina which is enriched in red, blue and green cone PRs and the peripheral retina which is enriched in rod PRs **(Figure 2)**. To further characterize rod and cone-specific gene regulation, we performed RNA-seq on macular vs peripheral retina collected from 7 donor eyes.

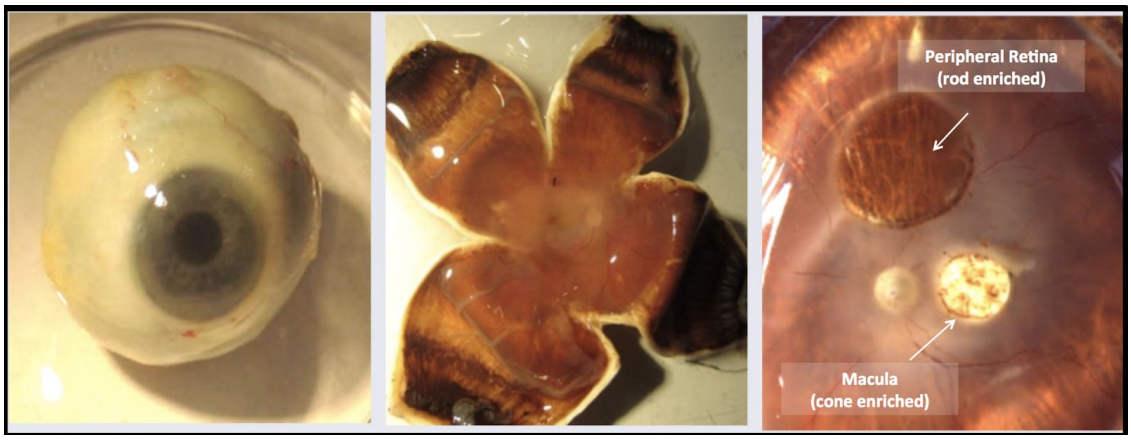


Figure 2. Post-mortem human eye (left) dissected eye exposing retina (middle), and sampled retina (right). Central macular retina contains 99% cone PRs, peripheral retina contains 99% rod PRs.

Interpretation of FastQC Reports: FastQC reports will show 12 Pass/Warning/Fail quality control metrics summarizing the overall quality of your raw sequence data for each sample. The different metrics are more/less important for different types of sequencing projects. Individual reports may generate a warning or fail, this does not mean your data are unusable. In general for RNA-seq data sets the most important FastQC metrics to consider are the following:

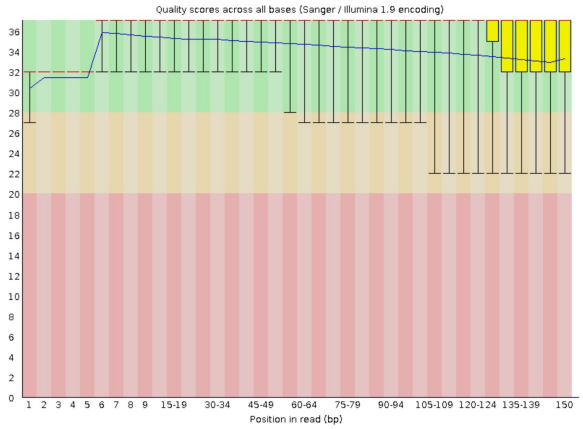
Basic Statistics

Measure	Value
Filename	E16retina1_R1.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	39324028
Sequences flagged as poor quality	0
Sequence length	32-150
%GC	47

- **Basic Statistics**

This report shows the total # of sequencing reads in the FASTQ file as well as how long the reads are and how many are poor quality. Fewer reads than expected or high # of flagged sequences would indicate a problem with the sequencing.

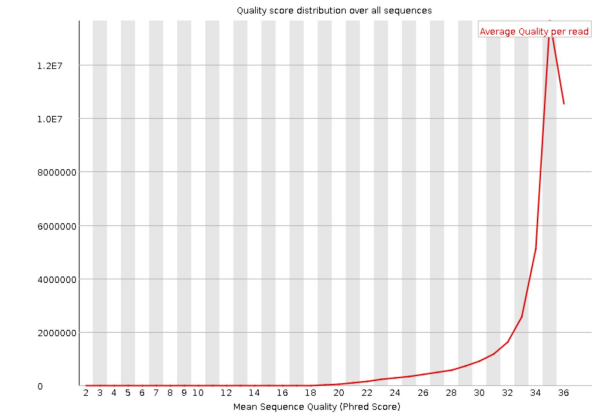
Per base sequence quality



- **Per base sequence quality**

This report shows the avg Phred quality score across the length of all reads per base. Ideally, the ave Phred quality score at each base will be >28. Low quality base calls at the 5' and 3' end can be trimmed by downstream filtering. Low on avg scores across reads would indicate a problem with the sequencing.

Per sequence quality scores

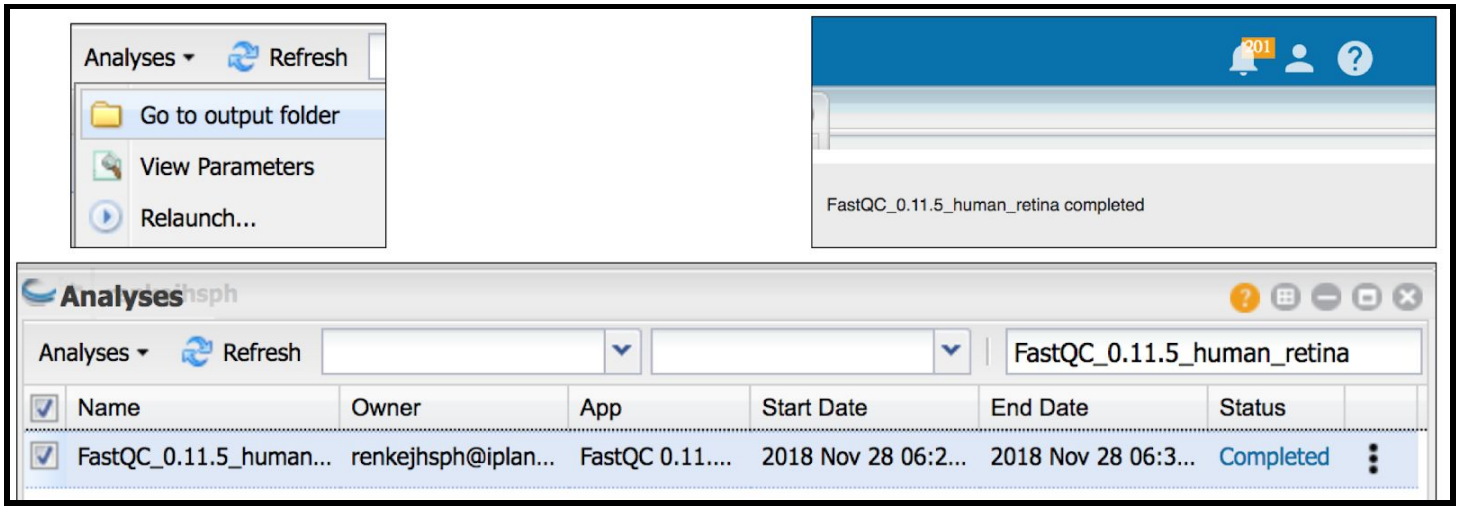


- **Per sequence quality scores**

This report indicates the distribution of Phred quality scores for all reads in the FASTQ file. Ideally, the bulk of reads will have a high avg Phred quality score (>28). Low quality reads can be removed by downstream filtering. A peak <28 would indicate a problem with the sequencing.

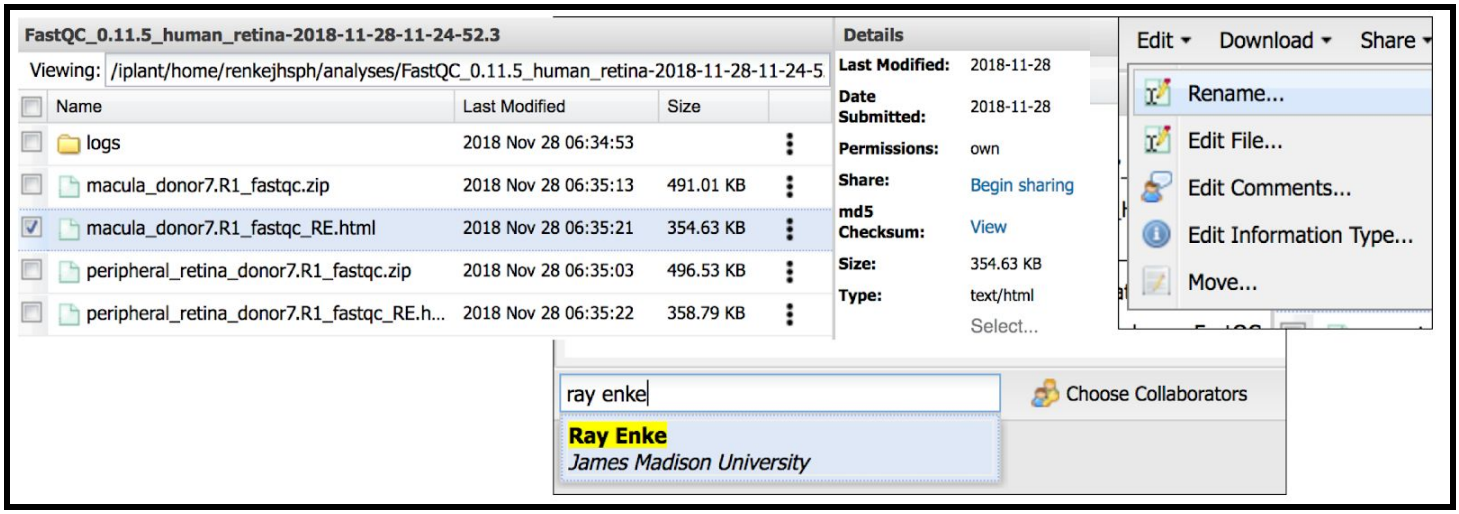
1. Viewing & Sharing FastQC Reports: Last class you ran the **FastQC** program in CyVerse DE to check the quality of the 2 FASTQ files that we're using for our analysis. The job should be complete so we can look at the FastQC report data and you can also easily share this data with me using DE's share options.

- Click on the bell icon in the upper right of your window to see your notifications
- Click on the notification informing you that your FastQC job is complete
- In the Analyses window that opens, select the FastQC job
- In the Analyses dropdown menu, select "Go to output folder to view your FastQC output data files



FastQC outputs 2 data files for each FASTQ file analyzed, 1 .html file and 1 .zip file. The html file has the data we want to view.

- Select your .html files one at a time then use the drop down menu to select Edit>Rename
  - Add and underscore followed by your initials just before the .html (no spaces!)
- Select your renamed macula .html file and click the "Begin sharing" option in the Details window
  - type my name "Ray Enke" into the search field and share you macula FastQC report with me



After sharing your data with me, open your FastQC reports and collect a few pieces of data for a figure

- Click on each of the .html files to open your FastQC reports in a web browser
  - You will have to disable your browser's popup blocker to view your reports

FastQC Report

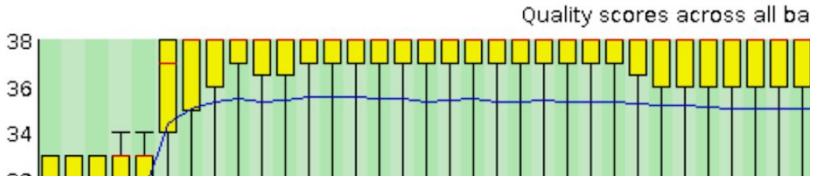
Summary

- Basic Statistics
- Per base sequence quality
- Per tile sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content
- Kmer Content

Basic Statistics

Measure	Value
Filename	macula_donor7.R1.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	22093938
Sequences flagged as poor quality	0
Sequence length	125
%GC	45

Per base sequence quality



- **Create a figure as outlined below in MS PowerPoint or similar using some of your FastQC data**

Macula basis stats

Macula per base quality

Macula per sequence quality

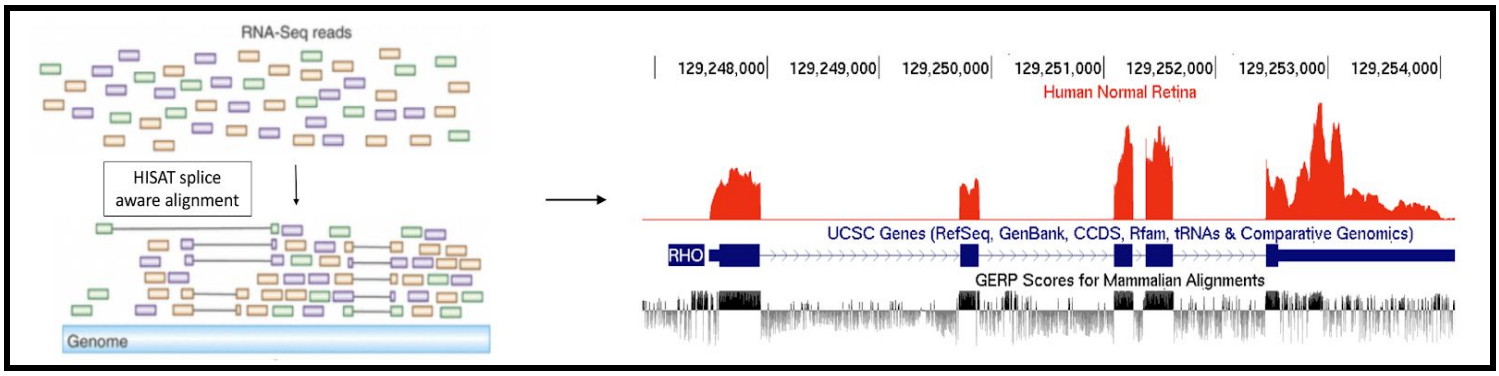
Peripheral retina basis stats

Peripheral retina per base quality

Peripheral retina per sequence

Figure legend describing your method, how many reads in each file and a brief description of the quality of the 2 files

2. **Creating HISAT Alignment Custom UCSC Tracks**

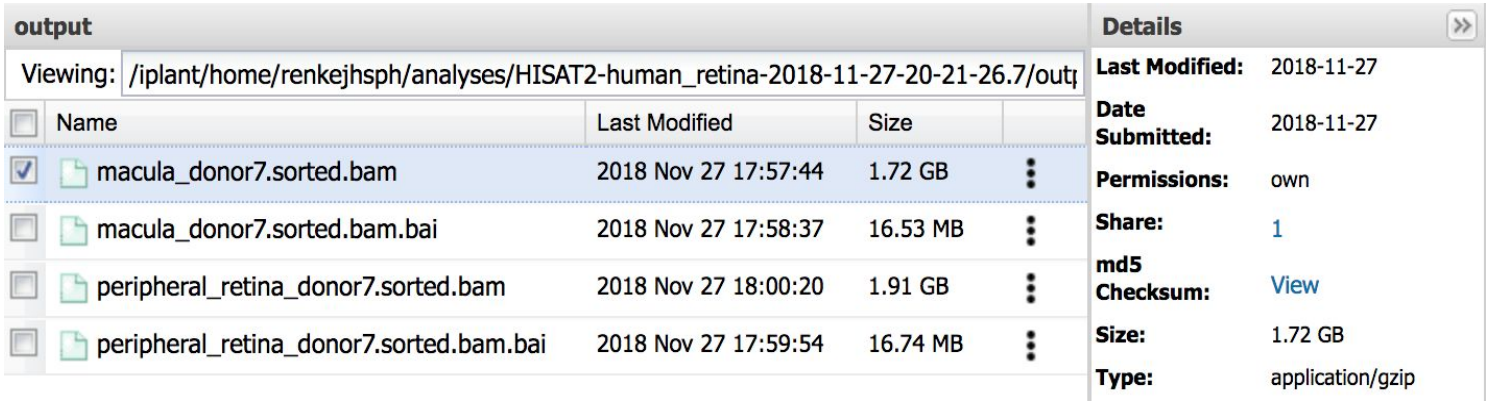


Last class you also initiated **HISAT alignments** of your 2 retina RNA-seq FASTQ files to the hg38 human genome assembly. HISAT outputs can be used as inputs for further differential transcript expression analysis in a RNA-seq bioinformatics pipeline. However, the alignment output file itself is a very useful piece of data.

- Click your notification in DE to find and open your HISAT alignment output folder
    - notification>select job>select file>select “Go to output folder”>select “output” folder

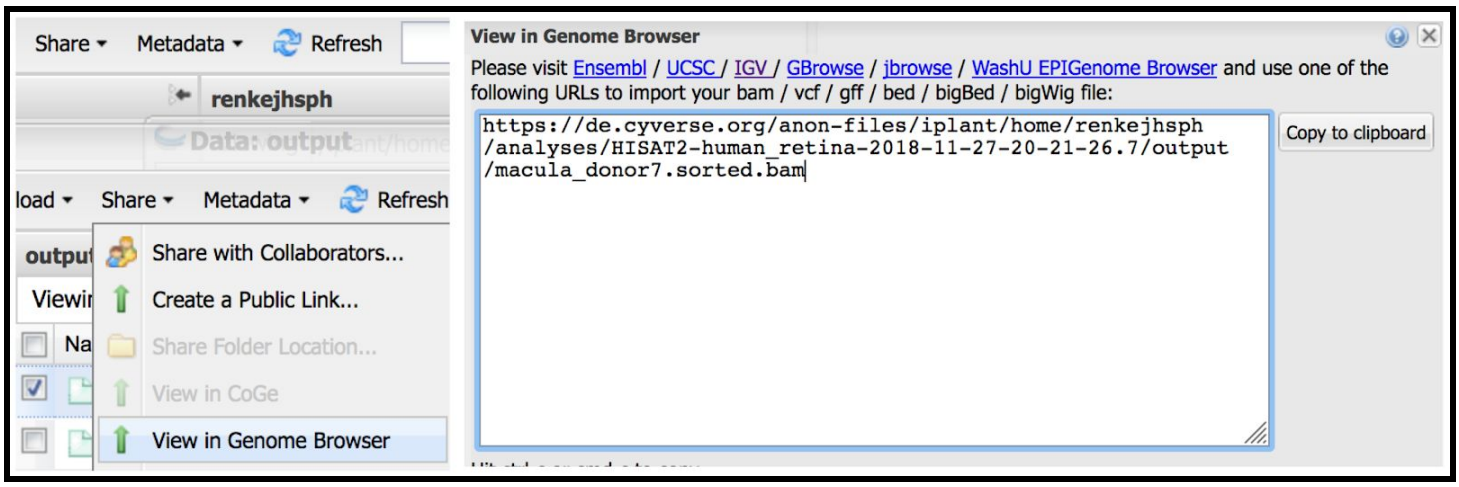
HISAT outputs 2 large files for each FASTQ file that was input and aligned

1. **.bam file** - contains all of the genome coordinates for the millions of aligned sequencing reads
2. **.bam.bai file** - an index file associated with each .bam file (let’s ignore this for now)



Bam files are sort of similar to **BED files** and can be used to create custom tracks in genome browsers such as UCSC. Rather than uploading the millions of genome coordinates in a Bam file to a browser, we can simply input the URL to where the Bam files live in CyVerse and UCSC will make our custom track using the shared data. This is a very useful tool!

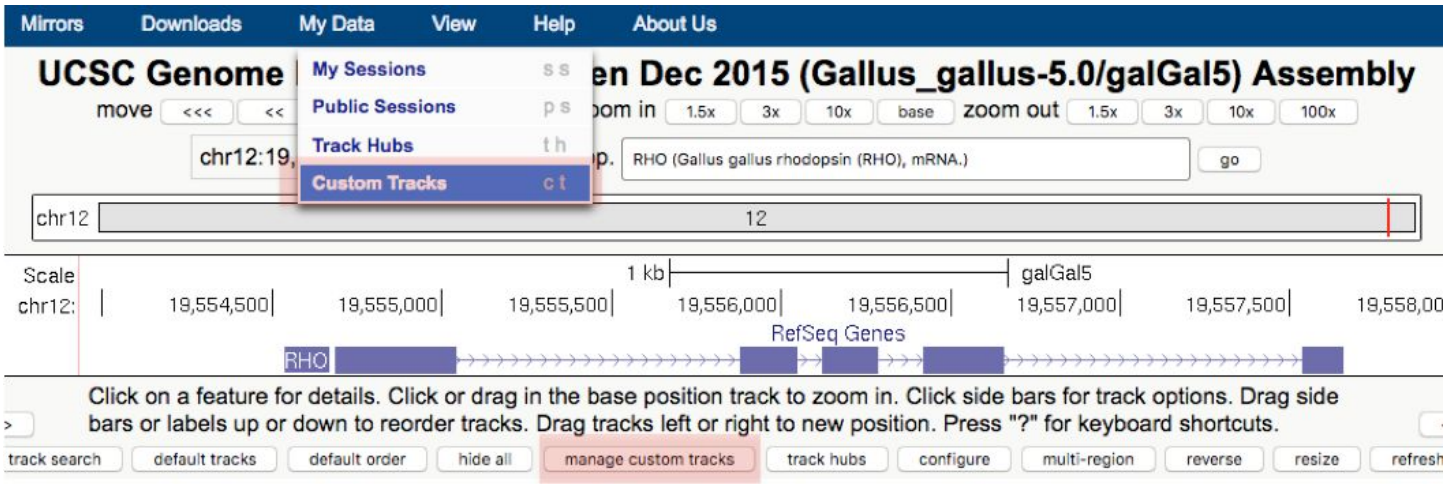
- Select the macula Bam file and then select “Share” in the DE toolbar
- Select the “View in a Genome Browser” options
- Copy the link that pops up which is the URL where the Bam file lives in DE
- Paste the URL somewhere where you can find it later and label it “macula bam”
- Repeat the process for the peripheral retina Bam URL
- Use the below tutorial to create Bam Density Plot custom tracks in the UCSC Genome Browser



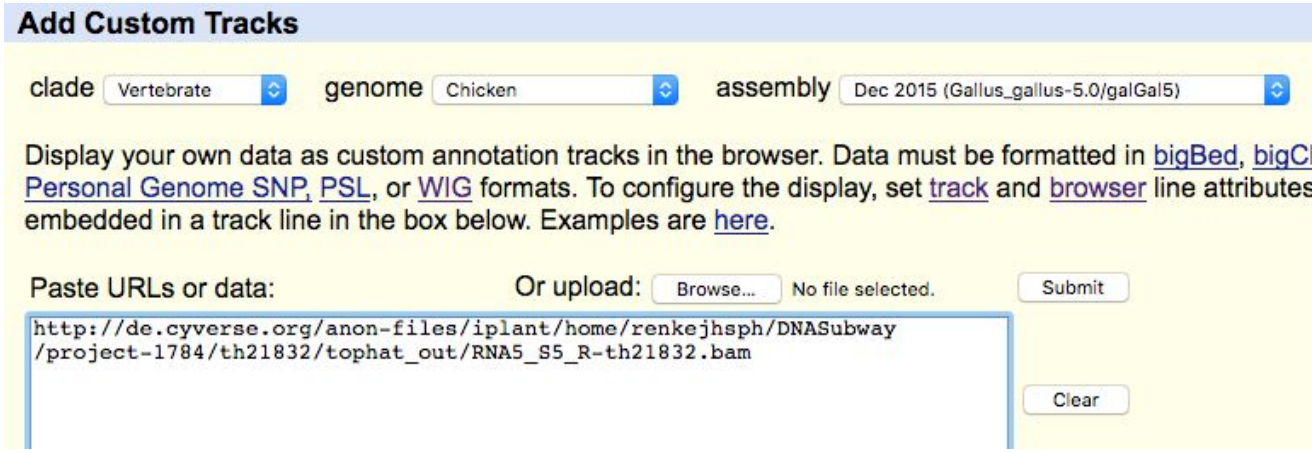
FYI, the below example is a demo on how to make Bam custom tracks using chicken RNA-seq data in the chicken genome. Use this demo to make custom tracks with your human Bam data in the hg38 human genome assembly.

Bam Density Plot Custom Track Demo

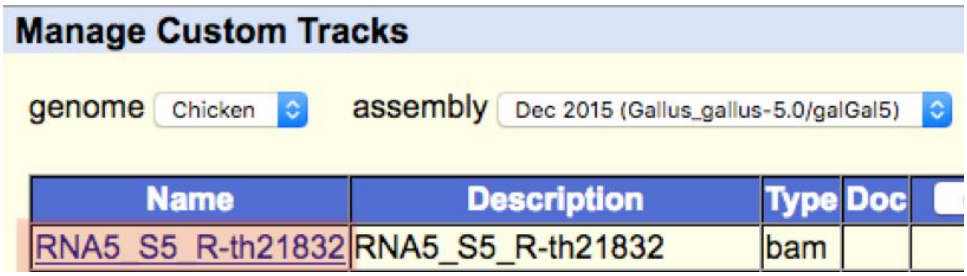
- Navigate to the Chicken 2015 galGal5 genome assembly (note, this is not the default assembly)
    - Hide all tracks then add back **RefSeq Genes (pack)**
    - Reconfigure to remove blue vertical lines and increase text size to 14
- Select either **My Data>Custom Tracks** from the top menu or **Manage custom tracks** from below the browser window



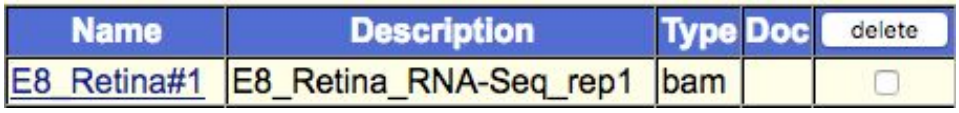
- On the Manage Custom Tracks page, select **add custom tracks**
- On the Add Custom Tracks page, copy/paste the BAM URL for E8 retina replicate #1 into the custom track window and hit submit (may take a few minutes to load)



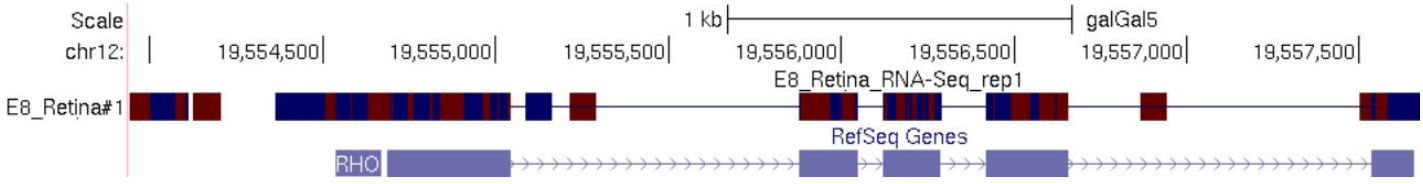
- Once the data is loaded click into the **Name link** to change the Name and Description of the data file



- change the **track name** & **description** by replacing the text in the Edit Configuration window
    - *track name=E8_Retina#1 description=E8_Retina_RNA-Seq_rep1 type=bam*
- Hit submit. Your track Name & Description should reflect this change. If not, try again (must be exact)



- Hit **Go** & navigate to the ***RHO* gene** to see the default view of your custom track



- Click into the Custom track to edit the track settings
- Change 2 things here
    - Change **Display mode** to full
    - Select the option to **Display data as a density graph**

These options will allow you to view a quantitative **BAM Density Plot** of the amount of reads that pile up at each gene in each sample. The data scaling will constrain the y-axis for each sample to a 0 - 1000 normalized read range. Feel free to play around with the other options to change the view of the TopHat alignment data.

E8_Retina_RNA-seq_Rep1 (▲All Custom Tracks)

Display mode: full Submit Remove custom track Update custom track

Display read names

Minimum alignment quality: 0

Color track by bases: different item bases Help on base coloring

Alignment Gap/Insertion Display Options Help on display options

- Draw double horizontal lines when both genome and query have an insertion
- Draw a vertical purple line for an insertion at the beginning or end of the query, orange for insertion in the middle of the query
- Draw a vertical green line where query has a polyA tail insertion

Additional coloring modes:

- Color by strand (blue for +, red for -)
- Use gray for alignment quality
- Use R,G,B colors specified in user-defined tag YC
- No additional coloring

Display data as a density graph:

Type of graph: bar Graph configuration help

Track height: 128 pixels (range: 11 to 128)

Data view scaling: auto-scale to data view Always include zero: OFF

Vertical viewing range: min: 0 max: 2500 (range: 0 to 1000)

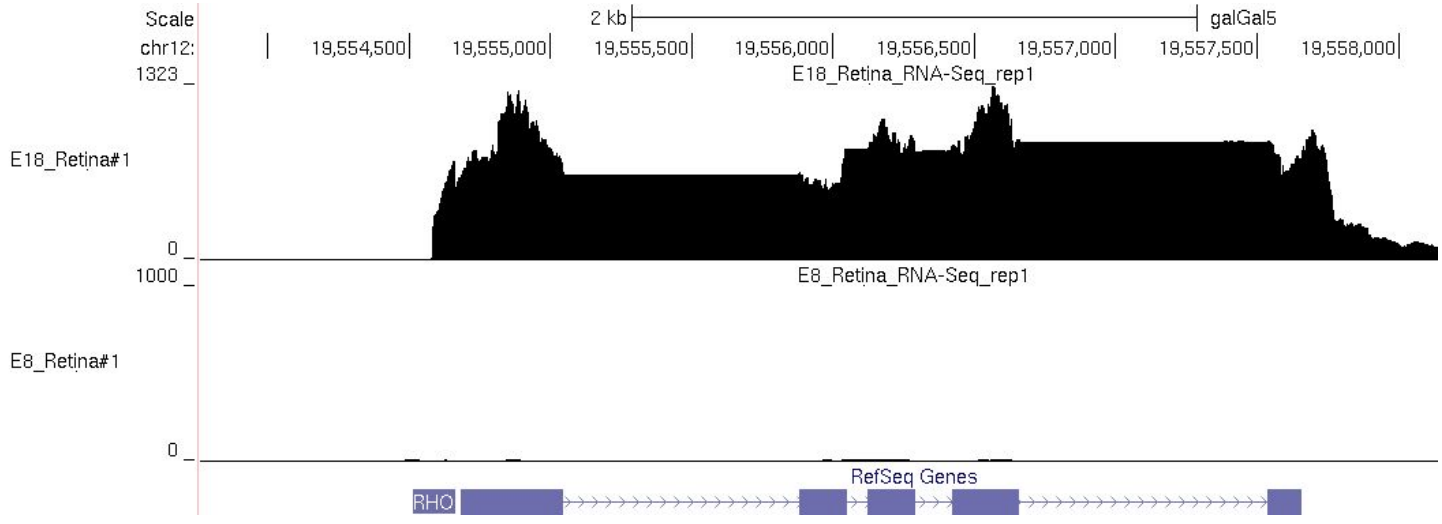
Transform function: Transform data points by: NONE

Windowing function: maximum **Smoothing window:** OFF pixels

Negate values:

Draw y indicator lines: at y = 0.0: OFF at y = 0 OFF

Here's an example of what these data would look like:



In class assignment (in groups or pairs; 10 pts; due Fri 11/30 5pm)

1. Share your macula FastQC .html output file with me as outlined in part 1
2. Create 1 PowerPoint or similar figure & legend outlining your FastQC data (as outlined above on pg 4)
3. Create 2 Bam density plot custom tracks in the hg38 human genome assembly visualizing your macula and peripheral retina RNA-seq HISAT alignments. Your session should be formatted as follows
    - Text size 14 with no blue vertical lines
    - Gencode gene (pack)
    - Conservation (full)
    - Macula custom track (full)
        - Red data track; track name=Macula mRNA; track description=Macula RNA-seq donor #7
        - Change the “Data view scaling” to “use vertical view range setting” and set the “max” number to 30575 (the same setting as the peripheral retina); this will change the scaling of the y-axis
    - Peripheral retina custom track (full)
        - Blue data track; track name=Retina mRNA; track description=Peripheral Retina RNA-seq donor #7
4. Create a 2nd PowerPoint of similar figure & legend of the human RHO gene plus the upstream enhancer and promoter region with the above tracks

Submit your 2 figures to the Canvas assignment page. Don’t forget to also share your FastQC data with me in DE.