

James Madison University

From the Selected Works of Ray Enke Ph.D.

November, 2018

FASTQC Analysis & HISAT Alignments Using CyVerse (part 1)

Ray A Enke



This work is licensed under a [Creative Commons CC BY-SA International License](https://creativecommons.org/licenses/by-sa/4.0/).



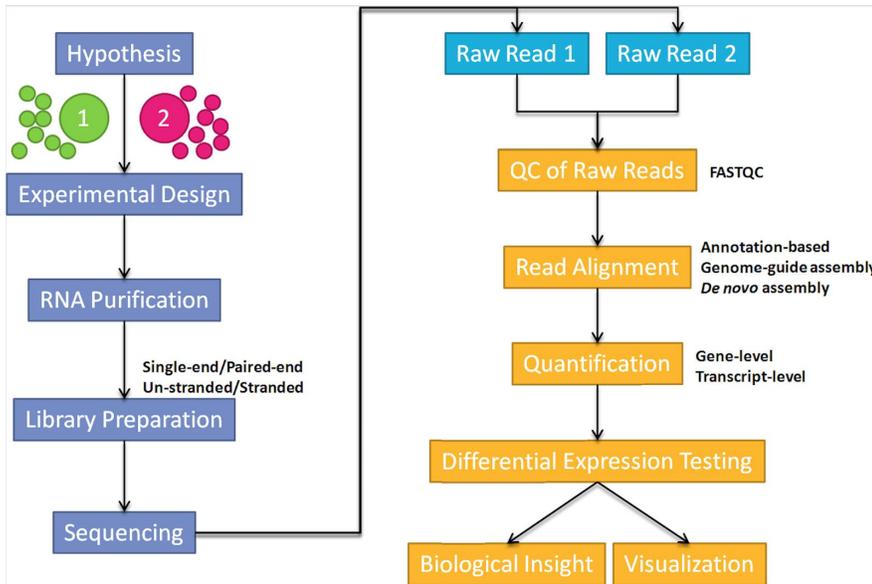
Available at: https://works.bepress.com/raymond_enke/104/

RNA-Seq Analysis Bioinformatics Pipeline

Part 1: Navigating CyVerse Discovery Environment

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you will learn how to perform several steps in a RNA-sequencing (RNA-seq) bioinformatics data analysis pipeline using unpublished Illumina NGS dataset from my lab.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.

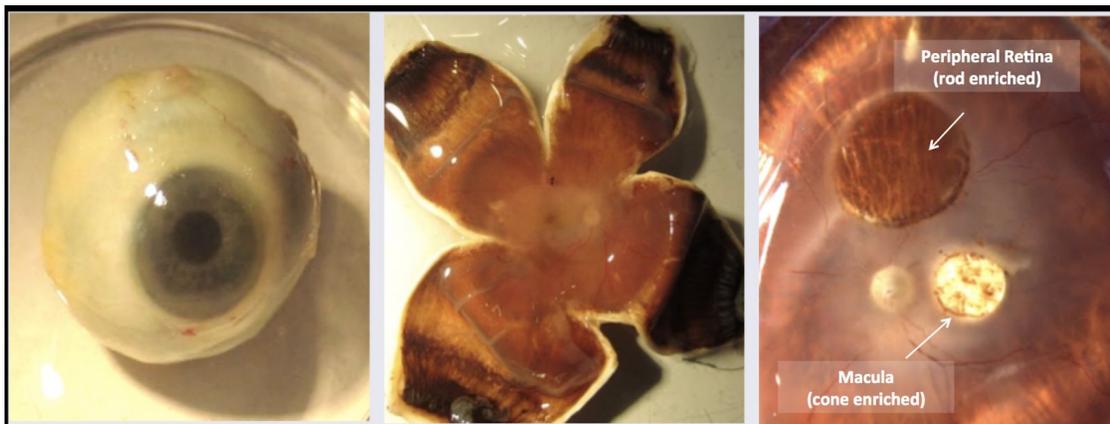


Figure 2. Post-mortem human eye (left) dissected eye exposing retina (middle), and sampled retina (right). Central macular retina contains 99% cone PRs, peripheral retina contains 99% rod PRs.

1. Access Data in CyVerse Discovery Environment (DE): CyVerse is an NSF-funded virtual organization that provides free data storage, bioinformatics analysis and other cyberinfrastructure to registered users. All of you already have a free CyVerse account when you signed up for DNA Subway, one of CyVerse’s tools. Within CyVerse, we will use a space called the Discovery Environment (DE) to store, analyze, and share data.

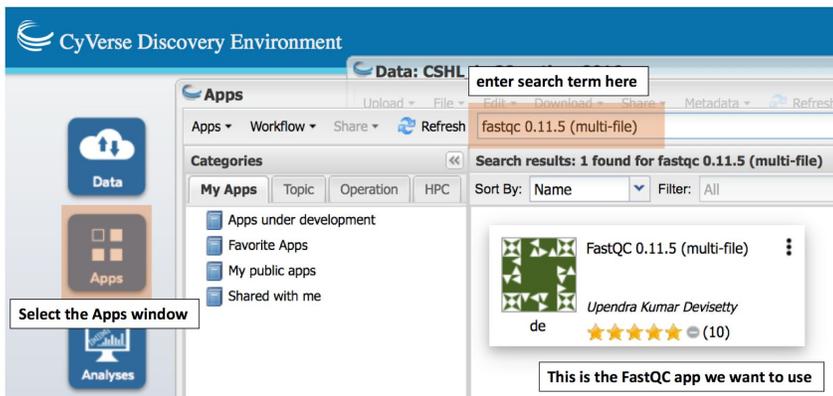
FASTQ files are pretty big (range from 1 - 5 GB) and are not easy to move around. One of the features of CyVerse DE is that files can stay in one physical location in the cloud and be shared rather than moved. Within DE, I’ve shared 2 Illumina FASTQ files from our RNA-seq experiment with you. You should be able to open your “Shared With Me” folder to view the 2 fastq files.



These **fastq.gz files** contain the raw Illumina sequencing data for two RNA-seq samples from post-mortem donor #7. The 1st file is data from RNA extracted from donor #7’s macular retina (cone PR rich) and the 2nd file is from RNA extracted from donor #7’s peripheral retina (rod PR rich). The .gz extension indicates that the fastq files are compressed for easier storage and sharing.

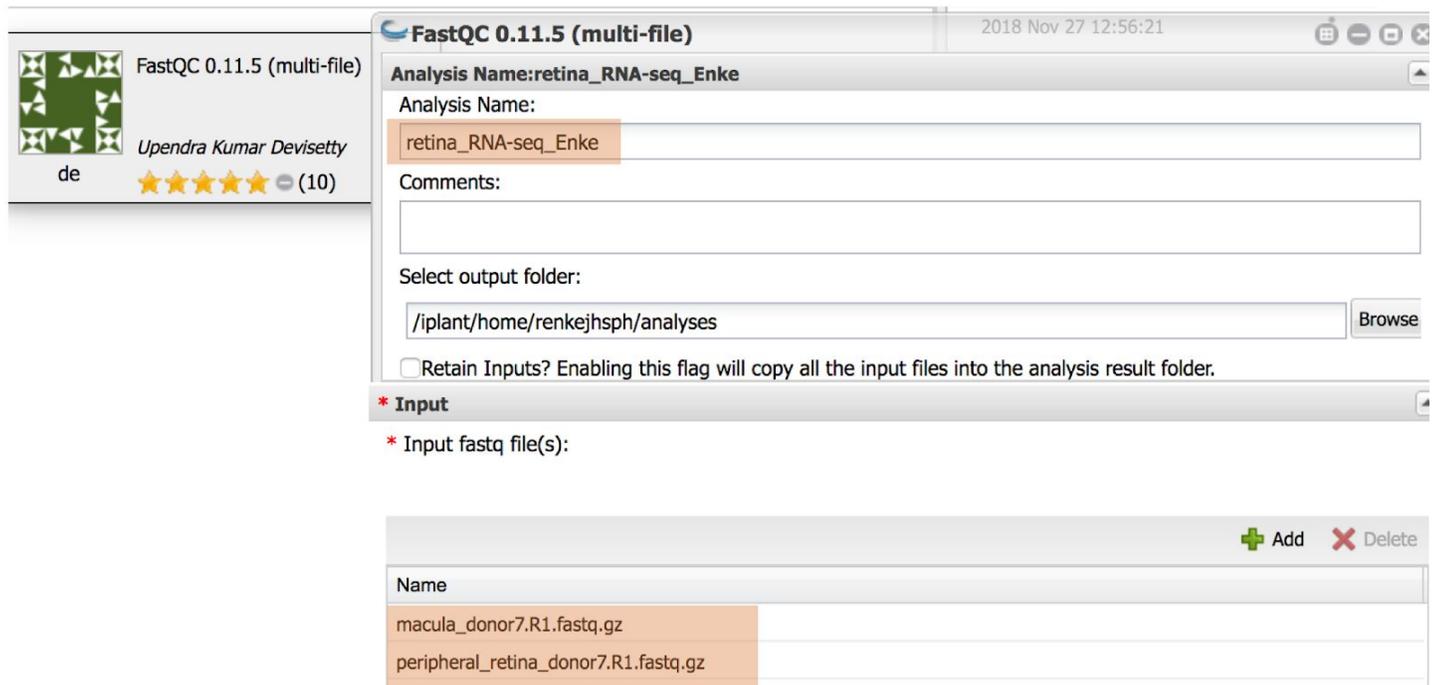
- Use your DNA Subway login to log into CyVerse DE (<https://de.cyverse.org/de/>)
- Once logged into DE, locate the 2 files
- Click on one of them to open
 - What does the open file look like?

2. FastQC Analysis of Raw Sequencing Reads: The 1st step of any NGS analysis, including RNA-seq, is checking the quality of the raw sequencing reads in each FASTQ file. If the sequence quality is poor, then your resulting analysis will by default be poor. A [software package called FastQC](#) developed by the Babraham Institute is almost universally used for this initial analysis. FastQC is one of many bioinformatics software apps that have been incorporated for easy use in CyVerse DE. A short video overview of FastQC can be viewed from the Babraham Institute’s YouTube page: <https://www.youtube.com/watch?v=bz93ReOv87Y>



- Click on the Apps icon and search for “fastqc 0.11.5”
- Click on the FastQC 0.11.5 (multi-file) app to open
 - Change “Analysis Name,” to “retina_RNA-seq_FastQC” (must use underscores, no spaces)
 - Next, open the “Input” window and click on “Add”
- Navigate to the folder containing the

- 2 shared FASTQ files, select them both and hit OK
- Hit launch analysis to start this job



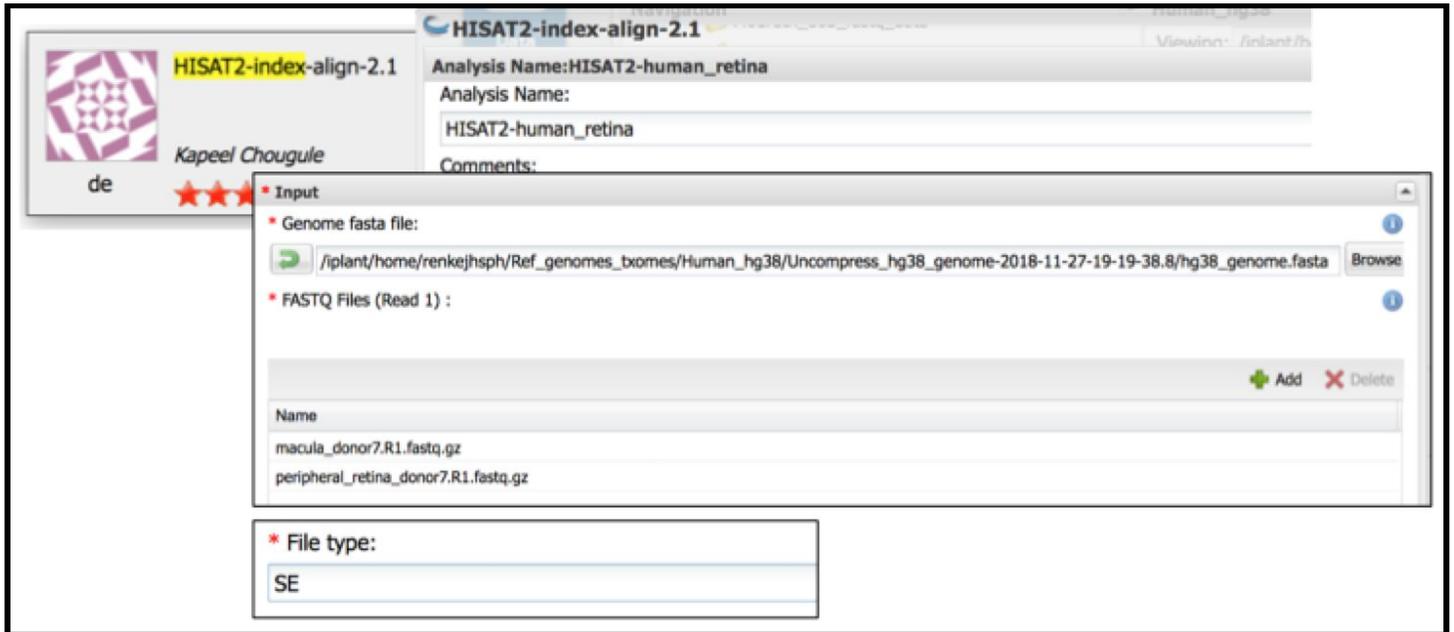
The FastQC analysis will take ~1 hour to run. The analysis is not occurring on your computer though. The jobs are being sent to the [Texas Advanced Computing Center](#) (TACC) which offers cloud computing support for CyVerse cyberinfrastructure. This means that you can close or turn off your computer while this job is running on TACC's supercomputer, [Stampede2](#). You will receive a CyVerse notification when the job is complete. We will take a look at the FastQC output next class.

3. HISAT Alignment of FASTQ Data. The next step in our analysis pipeline is to take the millions of reads in each FASTQ file and align them to the a genome assembly. You can align to any human genome assembly you want, in this case it makes the most sense to use the most recent hg38 assembly. We previously talked about an aligner called **TopHat** which used to be the standard software for aligning sequencing reads generated from eukaryotic mRNAs to a DNA genome. This is a special instance where the reads will correspond only to coding exons and not intervening introns which poses a significant computational barrier. TopHat is an example of a **splice aware aligner**. TopHat has been replaced by a very similar software called **HISAT** which pretty much does the same exact thing only faster and more accurately. We will use the HISAT app in DE to align the millions of reads in our macular and peripheral retina fastq files to the human genome.

You will need a 3rd file that I've shared with you called "**hg38_genome.fasta**" which is a FASTA version of the entire human genome. This is a pretty massive file as you can imagine (>53 GB!) and it's much easier to share rather than download and import into your CyVerse account.

- Click on the Apps icon and search for "HISAT2-index"
- Click on the **HISAT2-index-aligner-2.1** app to open
- Change "Analysis Name," change the folder name to "**HISAT2-human_retina**" (must use underscores, no spaces)
- Next, open the "Input" window and in the "Genome fasta file" field browse for the human genome file that I shared with you (in your "Shared With Me" folder; file is called "**hg38_genome.fasta**"
 - This tells HISAT which reference genome to align the sequencing reads to
- In the "FASTQ Files" field, select the 2 compressed FASTQ data files that I shared with you (macula7 and peripheral retina7)

- Scroll down and for “File type” select **SE** (for single end reads)
- Leave everything else as default settings and hit launch to send the analysis to the TACC supercomputer.



HISAT is computationally fairly complex and can take several hours to run. No worries though, since we're running in these analyses in the cloud and not on your computer, feel free to shut your laptop and wait until you get an e-mail notification from CyVerse telling you that your analyses are complete. We will take a look at the HISAT output files next class by adding them as custom tracks to the UCSC Genome Browser.