**RNA-seq analysis pipeline**

**Note: All of the tools mentioned should already be included in the Great Lakes software library.**

1. Check reads for base quality with [fastqc](https://www.bioinformatics.babraham.ac.uk/projects/fastqc/).
2. If reads still contain adapter sequences or have poor quality reads, the adapters or low quality bases can be trimmed using a variety of tools. I prefer [Trimmomatic](http://www.usadellab.org/cms/?page=trimmomatic). It has a lot of flexibility but it’s not super intuitive. Other sequence trim tools are [TrimGalore](https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/), the FASTQ/A Trimmer and FASTQ Quality Trimmer in the [FASTX tool kit](http://hannonlab.cshl.edu/fastx_toolkit/), and variety of others. You basically just want to remove adapters if present and poor sequence at the ends of the reads.
3. After trimming, it’s a good idea to check quality again with fastqc.
4. Alignment – At this point, you have several million sequences but you don’t know what they are. Here, you’re going to align the sequences to either a genome or a curated set of transcripts/genes to identify what your sequences are. There are several different aligners. The most commonly used for RNA-seq is from the “Tuxedo” suite of tools from Johns Hopkins. [Tophat/Tophat2](https://ccb.jhu.edu/software/tophat/index.shtml) is commonly used but is a bit older and is not being updated. It has been replaced by [Hisat2](https://daehwankimlab.github.io/hisat2/). Hisat2 was used for the alignment in the paper I sent you. You could use either Tophat2 or Hisat2.
5. Read counts – After getting reads aligned you want to get a count of each transcript. In the paper I sent, they used the feature-counts function in the [SubRead](http://subread.sourceforge.net/) package. I actually prefer [htseq-count](https://htseq.readthedocs.io/en/release_0.11.1/count.html). But, either one is fine.
6. Differential expression – Next, once you have the reads aligned and counted you’ll have a file with all of the genes/transcripts in the transcriptome with a count of each for every sample. It will look like this:

Table

Description automatically generated

From here, you want to do some quantitative analysis. You’re going to be assessing which genes/transcripts are expressed differently, higher or lower, between sets. Your data set has 4 different groups: Ly6CHI and Ly6CLO from either ND or DIO mice. There are 2 replicates for each group. Differential Expression analysis is commonly done with R/[Bioconductor](http://bioconductor.org/) packages. In the paper I sent, they used [edgeR](https://bioconductor.org/packages/release/bioc/html/edgeR.html) but I prefer [DESeq2](http://bioconductor.org/packages/release/bioc/html/DESeq2.html). Either one is likely fine for all practical purposes. They do use slightly different algorithms, so it’s not uncommon to find things differentially expressed with one tool that you don’t find in the other, and vice-versa.

1. After DE analysis, there are a million ways to mine the data. Also, there are other avenues that you can explore prior to DE analysis such as splice variant analysis or novel transcript discovery. The splice variant analysis could be an interesting thing to explore. We can talk about that later if you want.

Let me know if there is anything I can do to help you. If you get hung up at any point, don’t hesitate to email me.