Q5: Provide the script/pipeline you used to process raw RNA-seq data (.bcl) with clear annotation for each step.

**Aligning the sequences to a reference genome with Salmon**

1. Convert all “.bcl” files to “. fastq” files using Illumina's standalone program "bcl2fastq".
2. Remotely connect to the computing center where we will upload all necessary files and run the alignment job (At UNL we use the Holland Computing Center's Crane Cluster). This can be done through a third-party software like WINScp or via the command line/terminal (“SSH <username>@crane.unl.edu>”).
3. Upload all “.fastq” files, a reference genome (.gtf), SLURM bash instructions (HCC\_Slurm\_bash.txt), and a Salmon (Alignment software) script (salmon.slurm). The required files are below and are uploaded to the github repository for examination (excluding fastq files).
   1. RNA-seq Fastq files
   2. SLURM bash job info file (Titled “HCC\_Slurm\_bash.txt” in Github repository) – this file is annotated (hashtags “#”) with several things including the time the alignment job will take, the amount of CPU required for the job, the name of the job, what the name of the output error file will be, and where the output error file is placed. Additionally, this file tells you what commands to run in the unix shell to move to the correct directory (“cd /work/pblack/vperez3/rawdata”) start the job (“sbatch salmon.slurm”) and check the status of the job (“squeue -u vperez3”).
   3. Salmon script (Titled “salmon.slurm” in Github repository) – This is the bash script that will run the sequencing job after calling “sbatch salmon.slurm” in the terminal/command line. This is a bash script that loads the alignment software, Salmon, and executes Salmon commands to align the .fastq files to the .gtf file.
4. Once all the files, especially the .gtf file, are uploaded to the computing server, we can build the index from the .gtf and the Salmon software. The code to build the index for aligning is “salmon index -t transcripts.fa -i transcripts\_index --decoys decoys.txt -k 31”. This new index will be used by Salmon to align the sequences to.
5. After all files listed are uploaded to the computing center, and the salmon index has been build, we can run the code “sbatch salmon.slurm”, which tells the server to run the script salmon.slurm.

**Quantifying the Outputs from Salmon**

1. The output of salmon will leave us with quant.sf files, which include the transcript identifier, effective length, TPM (transcripts per million), and number of reads. These quant files can be further processed using R’s packages Tximport, Genomicfeatures, tidyverse, and edgeR (or Deseq2 in your labs case). At this step, we take the quant.sf files from the server, and bring them to our personal computer for analysis in R.
2. For the tximport to work, we need a tx2gene file, which is an annotated file built from the .gtf file that will help us convert our transcript IDs to gene IDs. We can build this with R packages genomicfeatures and tidyverse. See the script Q5\_DEG\_from\_raw\_files.R – section 1 for details.
3. After the tx2gene file is built, we can assign our transcripts gene IDs and counts. This is done using our tx2gene file and the tximport package in R. See the Q5\_DEG\_from\_raw\_files.R – section 2 script for details.
4. Lastly, tximport generates count files associated with gene names. These can be analyzed for differential expression using edgeR in R, or in your labs case, Deseq2. See the script Q5\_DEG\_from\_raw\_files.R - Section 3 for details on how this is run.

**Differential expression and pathway enrichment analysis from count files**

1. Tximport will provide count files which include the total counts for associated genes based on the transcript reads. These counts can be quantified using either edgeR or Deseq2. I prefer edgeR, see Q5\_DEG\_from\_raw\_files.R – section 3 for details or see Perez\_Mock\_Analysis.R and TCGA\_PAAD\_plus\_VMP.RData for details.
2. We can use the R package clusterProfiler to perform pathway enrichment analysis. Please see Q5\_DEG\_from\_raw\_files.R – section 4 for details.