**RNASeq Analysis Pipeline on Davros**

First step is to create file-having path of fastq files for paired end data in sorted order.

Go to the directory where all fastq files are located.

cd /data/rds/DMP/DUDMP/SPRECMED/DATA/EXTERNAL/COLON/RNASEQ/human/TPU/Project\_E68

Check the fastq file extension sometimes its .fq.gz or fastq.gz

Run,

find $(pwd) –name “\*.fq.gz” | sort > rnaseq\_11\_path

Check the file path where it is order or not.

head rnaseq\_11\_path

/data/rds/DMP/DUDMP/SPRECMED/DATA/INTERNAL/Colon/RNASeq/Human/TPU/Project\_E68/Sample\_E68\_0001/E68\_0001\_R1\_001.fq.gz

/data/rds/DMP/DUDMP/SPRECMED/DATA/INTERNAL/Colon/RNASeq/Human/TPU/Project\_E68/Sample\_E68\_0001/E68\_0001\_R2\_001.fq.gz

/data/rds/DMP/DUDMP/SPRECMED/DATA/INTERNAL/Colon/RNASeq/Human/TPU/Project\_E68/Sample\_E68\_0002/E68\_0002\_R1\_001.fq.gz

/data/rds/DMP/DUDMP/SPRECMED/DATA/INTERNAL/Colon/RNASeq/Human/TPU/Project\_E68/Sample\_E68\_0002/E68\_0002\_R2\_001.fq.gz

### **Make sure that your both paired end data in the sample folder.**

Go to the directory where you want to do the analysis.

Extract/copy rnaseq\_scripts.

cd rnaseq\_scripts

# Give the argument as path where you want to do the analysis. It will create proper folder structure for further analysis

./create\_folders.sh /scratch/DMP/SPRECMED/ypatil/Internal/colon/RNASeq/Human/TPU/Project\_E68/test123

copy fastq file path created in RDS to the folder in /scratch/DMP/SPRECMED/ypatil/Internal/colon/RNASeq/Human/TPU/Project\_E68/test123/scripts and /scratch/DMP/SPRECMED/ypatil/Internal/colon/RNASeq/Human/TPU/Project\_E68/test123/rna\_quality,

1. First step is to analyze the fastq files using Fastqc.
   1. In this step, we will get to know total number of reads counts for paired end data.
   2. Quality of fastq files.

Commands use to run the fastqc steps.

cd scripts

There is fastqc.pl script, run these scripts to create the fastqc jobs for each R1 and R2 file so that we can submit the jobs on HPC.

perl fastqc.pl /scratch/DMP/SPRECMED/ypatil/Internal/colon/RNASeq/Human/TPU/Project\_E68/test123/results/QC/ /scratch/DMP/SPRECMED/ypatil/Internal/colon/RNASeq/Human/TPU/Project\_E68/test123/scripts/fastqcScriptDir/ rnaseq\_11\_path

It will create a number of jobs w.r.t R1 and R2 files in fastqcScriptDir

cd fastqcScriptDir

Here, you can submit all jobs at once as

ls \*.sh > fastqc\_jobs

vi run\_jobs.sh

for i in $(cat fastqc\_jobs)

do

bsub < "$i"

sleep 2

done

save it

chmod +x run\_jobs.sh # make it executable

submit all jobs as

./run\_jobs.sh

or can submit one job as

bsub < E68\_0001\_R1\_001.sh

The output of fastqc will be generated in QC directory and all error and output log will be created in fastqcScriptDir.

The output in QC having a file of fastq\_read\_counts for all read counts for all R1 and R2 samples.

Go to the QC directory and run

./fastqc\_avg\_qc.sh # Will create average fastqc (Q30) valus for all R1 and R2 samples.

#####################################################################

### **Alignment of fastq reads onto the transcriptome**

cd scripts

In the rsem\_run.pl script, we have already given a path for transcriptome and its indexing using RSEM/bowtie2.

Run the following command to get the number of jobs for each sample in rsemScriptDir, so that we can submit jobs same as previously to HPC.

perl rsem\_run.pl /scratch/DMP/SPRECMED/ypatil/Internal/colon/RNASeq/Human/TPU/Project\_E68/test123/results/rsem\_results/ /scratch/DMP/SPRECMED/ypatil/Internal/colon/RNASeq/Human/TPU/Project\_E68/test123/scripts/rsemScriptDir/ rnaseq\_11\_path

cd rsemScriptDir

# Submit the jobs as previously

All .bam files and .tpm files for each sample will be generated in respective sample folder in rsem\_results.

Next step is getting the mapping percentage from aligned bam file.

First step is to sort the bam file and index it.

Run the following command to get the number of jobs for each sample in samtoolScriptDir, so that we can submit jobs same as previously to HPC.

perl samtool\_run.pl /scratch/DMP/SPRECMED/ypatil/Internal/colon/RNASeq/Human/TPU/Project\_E68/test123/results/rsem\_results/ /scratch/DMP/SPRECMED/ypatil/Internal/colon/RNASeq/Human/TPU/Project\_E68/test123/scripts/samtoolScriptDir/ rnaseq\_11\_path

# Submit the jobs as previously

It will create flagstat\_mapping results for each sample in rsem\_result directory.

Go to the results/rsem\_results directory and run the following script and give file name to it for transcriptome mapping percentage.

./getperc.sh > rnaseq\_transcriptome\_mapping\_perc\_colon\_TPU.txt

############## **Alignment of fastq reads onto the genome**

Perform the same flow same as transcriptome mapping for mapping to genome in rna\_quality directory.

Rna\_rsem\_run.pl(give the respective scriptDir and rna\_resul\_dir) followed by run picard\_run.pl(give the respective scriptDir and rna\_resul\_dir) and get the mapping percentage using getperc.sh

Go to the rna\_rsem\_results.

./getperc.sh > rnaseq\_transcriptome\_mapping\_perc\_colon\_TPU.txt

./getBaseperc.sh > rna\_quality\_colon\_TPU.txt

#######################################################################