Inspect QC statistics from sequencing run. Look at the % pass filter reads, number of pass filter reads.

**OPEN INTERACTIVE SESSION ON O2**

Open Terminal

**ssh -XY -l els32 o2.hms.harvard.edu**

**srun --pty -p interactive --mem 500M -t 0-06:00 /bin/bash**

**TRANSFER DATA ONTO LAB FOLDERS ON O2 OR OTHER SERVER**

## if transferring from BPF

#enter following command in the folder where you want to transfer the raw data

**scp -r Kula@bpfngs.med.harvard.edu:./FC\_02831 ./**

**MAKE A REFERENCE FILE FOR THE INPUT LIBRARY (IF NEEDED)**

# if it’s a new library, you will need to make a new fasta file from an annotations file. Look at Make CMV references for align ment.ipynb for instructions on how to do this.

#vir3.fasta is is provided

#pep2\_ref50 is provided

#once you have the fasta file, then do this:

**module load gcc/6.2.0**

**module load bowtie/1.2.2**

**bowtie-build vir3.fasta vir3**

**COPY FILES OVER TO OWN DIRECTORY -> On scratch 3 ?**

#make a directory with the name of the number of the VirScan plate you’re running

**mkdir els.plate1**

**cd els.plate1**

**mkdir raw.data**

**cd raw.data**

*#*the following command copies everything in /n/data2/hms/genetics/elledge/gjx1/screens/illumina/FC\_02917/Unaligned\_12\_PF\_mm1/Data/Project\_eshrock2 that ends in .bz2 into the current directory (this is specified by the period at the end of “$i.”

**for i in /home/jdh19/PhageWork/NonCoV\_Proj/FC\_06852/Unaligned\_1234\_PF\_mm1/Data/Project\_alterMember/\*.bz2; do cp $i .; done**

**UNZIP FILES**

# if the sequencing raw data is in fasta.bz2 format

**for i in raw.data/\*fastq.bz2; do bzip2 -d $i; done**

# if the sequencing raw data is in fasta.gz format

**for i in raw.data/\*.gz; do gzip -d $i; done**

# if the sequencing raw data is in .tar.gz format

**tar -xzf 180918\_NB501807\_0265\_AH3VNTAFXY.tar.gz**

**ALIGNMENT TO THE REFERENCE FILE**

#edit script.align.sh and submission\_script\_align as needed

# example code bowtie -5 17 -3 25 -n 3 -l 30 -e 1000 means trim off 17 nt off 5’ and and 183 nt off 3’ end

**#if we have 75 nt reads, then we often trim 25 nt off the 3’ end and align to the 50 nt refence sequences**

#use submission\_script\_align\_vir3 parallelize job submission in o2

**./submission\_script\_align\_vir3**

**CHECK THE ALIGNMENT**

download one of the .out files, change it to .txt, and see how many reads aligned

#typically we get >85% of reads aligning to the reference file

**INDEX FILES (takes ~15 min)**

**module load gcc/6.2.0**

**module load samtools/1.3.1**

**for i in raw.data/\*.bam; do samtools index $i; done**

## should get .bam.bai files for each of the index files

**My code version :**

**module load gcc/6.2.0**

**module load samtools/1.3.1**

**for i in Raw\_Data/\*.bam; do samtools index $i; done**

## should get .bam.bai files for each of the index files

**COUNT INDEXES**

**module load gcc/6.2.0**

**module load samtools/1.3.1**

**for i in raw.data/\*.bam; do samtools idxstats $i | cut -f 1,3 | sed -e '/^\\*\t/d' -e '1 i id\tSAMPLE\_ID' | tr "\\t" "," >${i%.bam}.count.csv; done**

## this gives you the count files for all of the index files.

**My code version :**

**module load gcc/6.2.0**

**module load samtools/1.3.1**

**for i in Raw\_Data/\*.bam; do samtools idxstats $i | cut -f 1,3 | sed -e '/^\\*\t/d' -e '1 i id\tSAMPLE\_ID' | tr "\\t" "," >${i%.bam}.count.csv; done**

## this gives you the count files for all of the index files.

**GZIP THE .COUNTS.CSV FILES**

**for i in raw.data/\*.csv; do gzip $i; done**

**My codes version :**

**for i in Raw\_Data/\*.csv; do gzip $i; done**

**MAKE LOG DIRECTORY FILE**

**mkdir log\_directory**

**cd**

**IF YOU RUN THE SAME SAMPLE ON >= TWO LANES OF A HISEQ, THEN YOU NEED TO COMBINE THE TWO FILES WHOSE FILES ADD UP TO THE SAME SAMPLE, SAME INDEX**

**module load gcc/6.2.0**

**module load python**

**for i in Raw\_Data/\*L001\_2\_R1.count.csv; do python combine\_two\_lanes.py $i ${i%1\_2\_R1.count.csv}3\_4\_R1.count.csv ${i%1\_2\_R1.count.csv}1\_2\_3\_4\_R1.count.combined.csv; done**

**<Copy and paste MyScripts here to then put into terminal>**

#the python script “combine\_two\_lanes.py” must be copied to the folder you are running the command from

**My Scripts :**

**for i in Raw\_Data/\*L001\_R1.count.csv.gz; do python combine\_two\_lanes.py $i ${i%1\_R1.count.csv.gz}2\_R1.count.csv.gz ${i%1\_R1.count.csv.gz}1\_2\_R1.count.csv; done**

**for i in Raw\_Data/\*L003\_R1.count.csv.gz; do python combine\_two\_lanes.py $i ${i%3\_R1.count.csv.gz}4\_R1.count.csv.gz ${i%3\_R1.count.csv.gz}3\_4\_R1.count.csv; done**

**for i in Raw\_Data/\*L001\_2\_R1.count.csv; do python combine\_two\_lanes.py $i ${i%1\_2\_R1.count.csv}3\_4\_R1.count.csv ${i%1\_2\_R1.count.csv}1\_2\_3\_4\_R1.count.combined.csv; done**

**OG Scripts :**

**for i in raw.data/\*L001\_R1.count.csv.gz; do python combine\_two\_lanes.py $i ${i%1\_R1.count.csv.gz}2\_R1.count.csv.gz ${i%1\_R1.count.csv.gz}1\_2\_R1.count.csv; done**

**for i in raw.data/\*L003\_R1.count.csv.gz; do python combine\_two\_lanes.py $i ${i%3\_R1.count.csv.gz}4\_R1.count.csv.gz ${i%3\_R1.count.csv.gz}3\_4\_R1.count.csv; done**

**for i in raw.data/\*L001\_2\_R1.count.csv; do python combine\_two\_lanes.py $i ${i%1\_2\_R1.count.csv}3\_4\_R1.count.csv ${i%1\_2\_R1.count.csv}1\_2\_3\_4\_R1.count.combined.csv; done**

**QC**

open up a counts.gz file locally, sum up how many counts there are for all the peptides, and then divide by total number of reads for that sample. that will tell you for that sample, how many reads mapped to the reference.

**GZIP EVERYTHING**

**for i in raw.data/\*1\_2\_3\_4\_R1.count.combined.csv; do gzip $i; done**

My Commands :

**for i in Raw\_Data/\*1\_2\_3\_4\_R1.count.combined.csv; do gzip $i; done**

**----------------------------------------------------------------------------------------------------**

**CONTINUE ANALYSIS IN JUPYTER IPYTHON NOTEBOOK**

open terminal

go to path: VirScan\_data\_elledge folder inside ellensh directory

**jupyter notebook**

##should launch the nb

open FileZilla

host: o2.hms.harvard.edu

username: els32

pass: ecommons

port: 22

type in “remote site” bar: /n/data2/hms/genetics/elledge/gjx1/screens/results/els.plate1/raw.data

find the .count.csv.gz and the .zipval.csv files, move them into a newly made directory in els/Desktop/VirScan\_data\_Elledge/els\_plate\_8\_9\_10\_11/

go to ipython notebook

**new** → **python3**

**CALCULATE VIRUS SCORES**

#in anaconda jupyter notebook, run this:

**for col in hits\_combined.columns:**

**temp = hits\_combined[col]**

**temp.to\_csv('hits/hits\_' + col + '.csv', header = True)**

#gzip all the csv files. also gzip the VIR\_3\_clean.csv file

**for i in hits/\*.csv.gz; do python calc\_scores\_nofilter.py $i VIR3\_clean.csv.gz Species 7 >virus\_scores\_$i; done**

**python calc\_scores\_nofilter.py hits\_ebola\_plate\_23/hits\_S1.csv.gz VIR3\_clean.csv.gz Species 7 >hits\_ebola\_plate\_23/virus\_scores\_S1.csv**

then delete the “.gz” from the files and change the second column header to “SAMPLE\_ID’

python concat\_tables.py TABLE1.csv.gz TABLE2.csv.gz TABLE3.csv.gz …

&gt;COMBINED.csv.gz

**APPENDIX**

#to gain permission to use scripts

**chmod 744 filename**

#to combine lots of fastq.gz files into one fastq file

**zcat \*R1\_\*fastq.gz > all\_MHCI\_R1.fastq**

#convert fastq to fasta

**seqtk/seqtk seq -a LIB030880\_GEN00110259\_S291\_L001\_R1.fastq > LIB030880\_S291\_R1.fasta**