Step 1

Index reference genome

Script: ref\_gen.sh

R\_script: index\_ref\_gen.R

Command: qsub ref\_gen.sh

Step 2

Make input file to split fastq

Script: split\_fastq\_make\_input.sh

Command:

* Give execute permission to your script:

chmod +x ./split\_fastq\_make\_input.sh

* Run script:

./split\_fastq\_make\_input.sh

Output: split\_fastq.input

Step 3

Download fastp

Command:

wget http://opengene.org/fastp/fastp

chmod a+x ./fastp

Step 4

Split fastq pairs into 2 million lines (500,000 reads)

Script: split\_fastq.sh

PBS: split\_fastq.pbs

Command: qsub split\_fastq.pbs

Input: split\_fastq.input

Step 5

Create input list for parallel alignment (the generated output has duplicates)

Script: align\_make\_input\_2.sh

Command: ./align\_make\_input\_2.sh

Input: samples.config\_2.txt

Output: align.input.txt

Step 6

Use R to remove duplicates of the previous step:

> rm\_dup <- read.table("align.input.txt")

> rm\_dup$V1 <- as.character(rm\_dup$V1)

> n\_occur <- data.frame(table(rm\_dup$V1))

> # rm\_dup <- rm\_dup[rm\_dup$V1 %in% n\_occur$Var1[n\_occur$Freq == 1],]

> write.table(n\_occur$Var1,"align.input.txt", quote = FALSE, row.names = FALSE, col.names = FALSE)

Move it back to Gadi

Step 7

Align read to reference genome using NGM

Script: align\_ngm.sh

Command: qsub align\_ngm.sh

Input script: align.sh

Input: align.input.txt

Step 8

Merge the chunked bam files for each sample into one bam

Script: merge\_aligned.pbs

Command: qsub merge\_aligned.pbs

Input script: merge.R

Step 9

Split bams by chromosome

Script: split\_bams.pbs

R\_script: split\_bams.R

Command: qsub split\_bams.pbs

Step 10

Mapping octopus

Script: octopus\_group.sh

R\_script: octopus\_group.R

Command: qsub octopus\_group.sh