Step

Index reference genome

Script: ref\_gen\_oryx.sh

R\_script: index\_ref\_gen\_oryx.R

Command: qsub ref\_gen\_oryx.sh

Step

Download fastp

Command:

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

chmod a+x ./fastp

Step

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

Script: split\_fastq\_oryx.pbs

R\_script: split\_fastq\_oryx.R

Command: qsub split\_fastq\_oryx.pbs

Step

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

Script: align\_make\_input\_oryx.sh

Give permission: chmod a+x ./align\_make\_input\_oryx.sh

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

Input: samples.config\_oryx.txt

Output: align.input\_oryx.txt

Step

Use R to remove duplicates of the previous step:

> rm\_dup <- read.table("align.input\_oryx.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\_oryx.txt", quote = FALSE, row.names = FALSE, col.names = FALSE)

Move it back to Gadi

Step

Align read to reference genome using NGM

Give permission: chmod a+x align\_oryx.sh

Script: align\_ngm\_oryx.sh

Command: qsub align\_ngm\_oryx.sh

Input script: align\_oryx.sh

Input: align.input\_oryx.txt

Step

Merge the chunked bam files for each sample into one bam

Script: merge\_aligned\_oryx.pbs

Command: qsub merge\_aligned\_oryx.pbs

Input script: merge\_oryx.R

Step

Split bams by chromosome

Script: split\_bams.pbs

R\_script: split\_bams.R

Command: qsub split\_bams.pbs

Step

Mapping octopus

Script: octopus\_group.sh

R\_script: octopus\_group.R

Command: qsub octopus\_group.sh