Whole Genome Sequencing

El objetivo es definir un procedimiento que procese un numero grande de secuencias WGS en el menor tiempo posible. Para ello,una vez definido el pipeline deberemos automatizar las tareas e integrarlas en el schedule manager del cluster (Como usar el cluster sin morir en el intento).

tl;dr

```
./wgs.pl -o <output_dir> [-cut <list.txt>] [-g] <input_dir>
```

Opciones:

- <input_dir> : (Mandatory) Directorio donde se encuentran todas las secuencias. El script buscara los sujetos y sus archivos dentro de este directorio.
- -o <output_dir> : (Opcional) Directorio donde se escribiran los resultados. En caso de obviarse se escribiran en el directorio desde el cual se lanza el script.
- -cut <list.txt> : (Opcional) Dice al script que analice SOLO los sujetos incluidos en el archivo que se suministra <list.txt>. Este archivo debe ser una listasimple de los sujetos a analizar.
- -g : Indica que no se borren los archivos temporales. Por defecto se borran, a no ser que se ponga este switch.

Pipeline

Primeramente hemos de definir el pipeline que se correra dentro del cluster. Aqui se ha de tener cuidado porque todas las herramientas y archivos han de ser accesibles desde cualquier nodo. En aras del siguiente paso podemos dividir el proceso en varios trozos. Tomemos por ejemplo el sujeto *seq-5*. Aqui los pasos son,

```
/nas/usr/local/bin/bwa mem -t 4 -R
"@RG\tID:V300016291_L01_549\tSM:seq5\tPL:BGI\tPI:380" -M
/the_dysk/BGI_exome/reference/Homo_sapiens_assembly38
/the_dysk/BGI_exome/F18FTSEUET0180/HUMehbE/seq-5/V300016291_L01_549_1.fq.gz
/the_dysk/BGI_exome/F18FTSEUET0180/HUMehbE/seq-5/V300016291_L01_549_2.fq.gz
> /the_dysk/BGI_exome/F18FTSEUET0180/WGS/seq-5/tmp/seq5_0.sam

/nas/usr/local/bin/bwa mem -t 4 -R
"@RG\tID:V300016291_L01_550\tSM:seq5\tPL:BGI\tPI:380" -M
/the_dysk/BGI_exome/reference/Homo_sapiens_assembly38
/the_dysk/BGI_exome/F18FTSEUET0180/HUMehbE/seq-5/V300016291_L01_550_1.fq.gz
/the_dysk/BGI_exome/F18FTSEUET0180/HUMehbE/seq-5/V300016291_L01_550_2.fq.gz
> /the_dysk/BGI_exome/F18FTSEUET0180/WGS/seq-5/tmp/seq5_1.sam

/nas/usr/local/bin/bwa mem -t 4 -R
"@RG\tID:V300016291_L01_551\tSM:seq5\tPL:BGI\tPI:380" -M
/the_dysk/BGI_exome/reference/Homo_sapiens_assembly38
/the_dysk/BGI_exome/F18FTSEUET0180/HUMehbE/seq-5/V300016291_L01_551_1.fq.gz
```

```
/the_dysk/BGI_exome/F18FTSEUET0180/HUMehbE/seq-5/V300016291_L01_551_2.fq.gz
> /the_dysk/BGI_exome/F18FTSEUET0180/WGS/seq-5/tmp/seq5_2.sam

/nas/usr/local/bin/bwa mem -t 4 -R
"@RG\tID:V300016291_L01_552\tSM:seq5\tPL:BGI\tPI:380" -M
/the_dysk/BGI_exome/reference/Homo_sapiens_assembly38
/the_dysk/BGI_exome/F18FTSEUET0180/HUMehbE/seq-5/V300016291_L01_552_1.fq.gz
/the_dysk/BGI_exome/F18FTSEUET0180/HUMehbE/seq-5/V300016291_L01_552_2.fq.gz
> /the_dysk/BGI_exome/F18FTSEUET0180/WGS/seq-5/tmp/seq5_3.sam

java -Djava.io.tmpdir=/nas/osotolongo/tmp/ -Xmx8g -jar
/nas/usr/local/bin/picard.jar MergeSamFiles
I=/the_dysk/BGI_exome/F18FTSEUET0180/WGS/seq-5/tmp/seq5_0.sam
I=/the_dysk/BGI_exome/F18FTSEUET0180/WGS/seq-5/tmp/seq5_1.sam
I=/the_dysk/BGI_exome/F18FTSEUET0180/WGS/seq-5/tmp/seq5_2.sam
I=/the_dysk/BGI_exome/F18FTSEUET0180/WGS/seq-5/tmp/seq5_3.sam
O=/the_dysk/BGI_exome/F18FTSEUET0180/WGS/seq-5/tmp/seq5.sam
java -Djava.io.tmpdir=/nas/osotolongo/tmp/ -Xmx8g -jar
/nas/usr/local/bin/picard.jar SortSam
I=/the_dysk/BGI_exome/F18FTSEUET0180/WGS/seq-5/tmp/seq5.sam
O=/the_dysk/BGI_exome/F18FTSEUET0180/WGS/seq-5/tmp/seq5_sorted.bam
SORT_ORDER=coordinate
/nas/software/samtools/bin/samtools index
/the_dysk/BGI_exome/F18FTSEUET0180/WGS/seq-5/tmp/seq5_sorted.bam

/nas/usr/local/bin/verifyBamID --vcf
/the_dysk/BGI_exome/reference/hapmap_3.3.hg38.vcf.gz --bam
/the_dysk/BGI_exome/F18FTSEUET0180/WGS/seq-5/tmp/seq5_sorted.bam --chip-none
--maxDepth 1000 --precise --verbose --ignoreRG --out
/the_dysk/BGI_exome/F18FTSEUET0180/WGS/seq-5/results/seq5_verifybam |& grep
-v "Skipping marker"

java -Djava.io.tmpdir=/nas/osotolongo/tmp/ -Xmx8g -jar
/nas/usr/local/bin/picard.jar ValidateSamFile IGNORE=MATE_NOT_FOUND
IGNORE=MISSING_READ_GROUP IGNORE=RECORD_MISSING_READ_GROUP
I=/the_dysk/BGI_exome/F18FTSEUET0180/WGS/seq-5/tmp/seq5_sorted.bam
java -Djava.io.tmpdir=/nas/osotolongo/tmp/ -Xmx8g -jar
/nas/usr/local/bin/picard.jar MarkDuplicates
I=/the_dysk/BGI_exome/F18FTSEUET0180/WGS/seq-5/tmp/seq5_sorted.bam
O=/the_dysk/BGI_exome/F18FTSEUET0180/WGS/seq-5/tmp/seq5_GATKready.bam
METRICS_FILE=/the_dysk/BGI_exome/F18FTSEUET0180/WGS/seq-5/tmp/seq5_metrics.t
xt QUIET=true MAX_RECORDS_IN_RAM=2000000 ASSUME_SORTED=TRUE
CREATE_INDEX=TRUE
java -jar /nas/usr/local/opt/gatk3/GenomeAnalysisTK.jar -T DepthOfCoverage -
R /the_dysk/BGI_exome/reference/Homo_sapiens_assembly38.fasta -nt 1 -ct 10 -
ct 15 -ct 20 -ct 30 --omitDepthOutputAtEachBase --omitIntervalStatistics --
omitLocusTable -L
/the_dysk/BGI_exome/reference/MGI_Exome_Capture_V5_bis2.bed -I
/the_dysk/BGI_exome/F18FTSEUET0180/WGS/seq-5/tmp/seq5_GATKready.bam -o
/the_dysk/BGI_exome/F18FTSEUET0180/WGS/seq-5/results/seq5_exome_coverage
singularity run --cleanenv -B /nas:/nas -B /the_dysk:/the_dysk
```

```
/usr/local/bin/gatk4.simg gatk --java-options "-
DGATK_STACKTRACE_ON_USER_EXCEPTION=true -Xmx16G" BaseRecalibrator -I
/the_dysk/BGI_exome/F18FTSEUET0180/WGS/seq-5/tmp/seq5_GATKready.bam -R
/the_dysk/BGI_exome/reference/Homo_sapiens_assembly38.fasta --known-sites
/the_dysk/BGI_exome/reference/Mills_and_1000G_gold_standard.indels.hg38.vcf.
gz --known-sites /the_dysk/BGI_exome/reference/dbsnp_146.hg38.vcf.gz --
known-sites
/the_dysk/BGI_exome/reference/1000G_phase1.snps.high_confidence.hg38.vcf.gz
-O /the_dysk/BGI_exome/F18FTSEUET0180/WGS/seq-5/tmp/seq5_recal_data.table1
singularity run --cleanenv -B /nas:/nas -B /the_dysk:/the_dysk
/usr/local/bin/gatk4.simg gatk --java-options "-
DGATK_STACKTRACE_ON_USER_EXCEPTION=true -Xmx16G" ApplyBQSR -R
/the_dysk/BGI_exome/reference/Homo_sapiens_assembly38.fasta -I
/the_dysk/BGI_exome/F18FTSEUET0180/WGS/seq-5/tmp/seq5_GATKready.bam -bqsr-
recal-file
/the_dysk/BGI_exome/F18FTSEUET0180/WGS/seq-5/tmp/seq5_recal_data.table1 -O
/the_dysk/BGI_exome/F18FTSEUET0180/WGS/seq-5/results/seq5_recal.bam
singularity run --cleanenv -B /nas:/nas -B /the_dysk:/the_dysk
/usr/local/bin/gatk4.simg gatk --java-options "-
DGATK_STACKTRACE_ON_USER_EXCEPTION=true -Xmx16G" AnalyzeCovariates -bqsr
/the_dysk/BGI_exome/F18FTSEUET0180/WGS/seq-5/tmp/seq5_recal_data.table1 --
plots
/the_dysk/BGI_exome/F18FTSEUET0180/WGS/seq-5/results/seq5_AnalyzeCovariates.
pdf
singularity run --cleanenv -B /nas:/nas -B /the_dysk:/the_dysk
/usr/local/bin/gatk4.simg gatk --java-options "-
DGATK_STACKTRACE_ON_USER_EXCEPTION=true -Xmx16G" BaseRecalibrator -I
/the_dysk/BGI_exome/F18FTSEUET0180/WGS/seq-5/results/seq5_recal.bam -R
/the_dysk/BGI_exome/reference/Homo_sapiens_assembly38.fasta --known-sites
/the_dysk/BGI_exome/reference/Mills_and_1000G_gold_standard.indels.hg38.vcf.
gz --known-sites /the_dysk/BGI_exome/reference/dbsnp_146.hg38.vcf.gz --
known-sites
/the_dysk/BGI_exome/reference/1000G_phase1.snps.high_confidence.hg38.vcf.gz
-O /the_dysk/BGI_exome/F18FTSEUET0180/WGS/seq-5/tmp/seq5_recal_data.table2
singularity run --cleanenv -B /nas:/nas -B /the_dysk:/the_dysk
/usr/local/bin/gatk4.simg gatk --java-options "-
DGATK_STACKTRACE_ON_USER_EXCEPTION=true -Xmx16G" AnalyzeCovariates -before
/the_dysk/BGI_exome/F18FTSEUET0180/WGS/seq-5/tmp/seq5_recal_data.table1 -
after
/the_dysk/BGI_exome/F18FTSEUET0180/WGS/seq-5/tmp/seq5_recal_data.table2 -
plots /the_dysk/BGI_exome/F18FTSEUET0180/WGS/seq-5/results/seq5_before-
after-plots.pdf
singularity run --cleanenv -B /nas:/nas -B /the_dysk:/the_dysk
/usr/local/bin/gatk4.simg gatk --java-options "-
DGATK_STACKTRACE_ON_USER_EXCEPTION=true -Xmx16G" HaplotypeCaller -R
/the_dysk/BGI_exome/reference/Homo_sapiens_assembly38.fasta -I
/the_dysk/BGI_exome/F18FTSEUET0180/WGS/seq-5/results/seq5_recal.bam -ERC
GVCF --dbsnp /the_dysk/BGI_exome/reference/dbsnp_146.hg38.vcf.gz -O
/the_dysk/BGI_exome/F18FTSEUET0180/WGS/seq-5/results/seq5_raw.snps.indels.g.
vcf.gz
singularity run --cleanenv -B /nas:/nas -B /the_dysk:/the_dysk
```

```
/usr/local/bin/gatk4.simg gatk --java-options "-
DGATK_STACKTRACE_ON_USER_EXCEPTION=true -Xmx16G -XX:+UseConcMarkSweepGC"
VariantEval -R /the_dysk/BGI_exome/reference/Homo_sapiens_assembly38.fasta -
L /the_dysk/BGI_exome/reference/MGI_Exome_Capture_V5_bis2.bed -D
/the_dysk/BGI_exome/reference/dbsnp_146.hg38.vcf.gz -O
/the_dysk/BGI_exome/F18FTSEUET0180/WGS/seq-5/results/seq5_eval.gatkreport --
eval
/the_dysk/BGI_exome/F18FTSEUET0180/WGS/seq-5/results/seq5_raw.snps.indels.g.
vcf.gz
```

Paralelizacion

Parsing

Troubleshooting