02.Preprocessing Pipelines

FASTQ Alignment

```
#!/bin/bash

#PBS -l select=1:ncpus=10:mem=100gb
#PBS -l walltime=48:00:00
#PBS -N FASTQ2BAM

# reference: https://www.htslib.org/workflow/fastq.html

# Load the modules required for the analysis
module load samtools
module load minimap2

# set directories
HOME="/Users/pollyhung/Desktop/QDNAseq/data"
FASTQ="$HOME/fastq"
SAMPLE_ID="$HOME/metadata/sample.txt"

# change directory
cd $FASTQ

## For HH_lung
# Loop through the files we wish to perform HsMetrics on
while IFS= read -r folder; do
# input
S1="$folder/S1.fq.gz"
S2="$folder/S2.fq.gz"
# reference
hg38_fasta=""
# others
temp_prefix="$folder/temp"
# intermediate outputs
minimap2_sam="$folder/minimap2.sam"
fixmate_bam="$folder/fixmate.bam"
sorted_bam="$folder/sorted.bam"
nodup_bam="$folder/nodup.bam"
# final outputs
```

```
sorted_nodup_bam="$folder/sorted_nodup.bam"
# Perform pipeline
minimap2 -t 8 -a -x sr "$hg38_fasta" "$S1" "$S2" -o "$minimap2_sam"
samtools fixmate -O bam,level=1 -m "$minimap2_sam" "$fixmate_bam"
samtools sort -l 1 -@8 -o "$sorted_bam" -T "$temp_prefix"
"$fixmate_bam"
samtools markdup -O bam,level=1 "$sorted_bam" "$nodup_bam"
samtools view -@8 "$nodup_bam" -o "$sorted_nodup_bam"
done < "$SAMPLE_ID"
```

DO NOT USE THIS

GATK Preprocess

```
#!/bin/bash
#PBS -l nodes=1:ppn=12
#PBS -l mem=50g
#PBS -l walltime=4:00:00
#PBS -m ae
#PBS -N sg1_2
#PBS -q medium

# library modules
echo "Start loading modules at $(date)"
module load bwa
module load Picard
module load samtools
echo "Finish loading modules at $(date)"

# sample_id
sample_id="H2_20PD"

# home
HOME="/home/polly_hung/sWGS"
REF="/home/polly_hung/reference"
DATA="$HOME/data"
RESULT="$HOME/output"
ERROR="$RESULT/error"
RECAL="$HOME/recalibration"

# reference
HG38_REF="$REF/hg38/Homo_sapiens.GRCh38.dna.primary_assembly.fa"
KNOWN_INDELS="$REF/vcf/Homo_sapiens_assembly38.known_indels.vcf.gz"
```

```
MILL_INDELS="$REF/vcf/Mills.indels.contig.adjusted.hg38.vcf.gz"
DBSNP="$REF/vcf/Homo_sapiens_assembly38.dbsnp138.vcf.gz"

# sample-based definition
SEQ_1="$DATA/$sample_id-1.fq.gz"
SEQ_2="$DATA/$sample_id-2.fq.gz"

error="$ERROR/$sample_id-bwa.err"
METRICS="$ERROR/$sample_id.metrics.txt"
RECAL_TABLE="$RECAL/$sample_id.recal_data.table"

SAM="$RESULT/sorted_dedup/$sample_id.sam"
SORTED_SAM="$RESULT/sorted_dedup/$sample_id.sorted.sam"
TAGGED_SAM="$RESULT/sorted_dedup/$sample_id.sorted.tagged.sam"
MARKED_BAM="$RESULT/sorted_dedup/$sample_id.sorted.marked.bam"
CALIBRATED_BAM="$RESULT/calibrated/$sample_id.sorted.marked.calibrate
d.bam"


cd "$DATA"
echo "Processing sample_id: $sample_id"

# Alignment to hg38 genome
bwa mem -M -t 12 \
"$HG38_REF" \
"$SEQ_1" "$SEQ_2" \
2> "$error" \
> "$SAM"


# Sorting SAM by coordinates
java -jar /software/Picard/3.2.0/picard.jar SortSam \
--INPUT "$SAM" \
--OUTPUT "$SORTED_SAM" \
--SORT_ORDER coordinate \
--VALIDATION_STRINGENCY SILENT


# Add read groups to the sorted sam
java -jar /software/Picard/3.2.0/picard.jar AddOrReplaceReadGroups \
--INPUT "$SORTED_SAM" \
--OUTPUT "$TAGGED_SAM" \
--RGLB "Short-Insert_Library" \
--RGPL "DNBSEQ" \
```

```
--RGPU "V350264599.2" \
--RGSM "$sample_id"


# Marking Duplicates
java -jar /software/Picard/3.2.0/picard.jar MarkDuplicates \
--INPUT "$TAGGED_SAM" \
--OUTPUT "$MARKED_BAM" \
--METRICS_FILE "$METRICS" \
--ASSUME_SORTED true \
--REMOVE_DUPLICATES true \
--VALIDATION_STRINGENCY SILENT


# Indexing the BAM file
cd "$RESULT"
samtools index "$MARKED_BAM"


echo "Start loading modules at $(date)"
module load miniconda3
module load GenomeAnalysisTK/4.2.0.0
source activate /software/GenomeAnalysisTK/4.2.0.0
echo "Finish loading modules at $(date)"


# base recalibrator
gatk BaseRecalibrator \
-I "$MARKED_BAM" \
-R "$HG38_REF" \
--known-sites "$KNOWN_INDELS" \
--known-sites "$MILL_INDELS" \
--known-sites "$DBSNP" \
-O "$RECAL_TABLE"

# apply recalibrator
gatk ApplyBQSR \
-R "$HG38_REF" \
-I "$MARKED_BAM" \
--bqsr-recal-file "$RECAL_TABLE" \
-O "$CALIBRATED_BAM"


# After indexing, remove the mid-process files
```

```
cd "$RESULT"
# rm "$SAM"
# rm "$SORTED_SAM"
# rm "$TAGGED_SAM"
```

Job submit records

```
qsub batch 1 data

qsub -v sample_id="FT190_NT" -N "FT190_NT" batch_1.sh
qsub -v sample_id="sg1_2" -N "sg1_2" batch_1.sh
qsub -v sample_id="sg1_3" -N "sg1_3" batch_1.sh
qsub -v sample_id="sg1_4" -N "sg1_4" batch_1.sh
qsub -v sample_id="sg1_5" -N "sg1_5" batch_1.sh

qsub batch 2 data

qsub -v sample_id="H2_20PD" -N "H2_20PD" batch_2.sh
qsub -v sample_id="H3_20PD" -N "H3_20PD" batch_2.sh
qsub -v sample_id="H6_20PD" -N "H6_20PD" batch_2.sh
qsub -v sample_id="H8_20PD" -N "H8_20PD" batch_2.sh
qsub -v sample_id="H9_20PD" -N "H9_20PD" batch_2.sh

qsub normal

qsub -v
SAMPLE_ID="/home/polly_hung/ccoc/samples/batch_by_type/normal/batch_1
.txt" -N "batch_1" 01_mutect2_normal_vcf.sh
qsub -v
SAMPLE_ID="/home/polly_hung/ccoc/samples/batch_by_type/normal/batch_2
.txt" -N "batch_2" 01_mutect2_normal_vcf.sh
qsub -v
SAMPLE_ID="/home/polly_hung/ccoc/samples/batch_by_type/normal/batch_3
.txt" -N "batch_3" 01_mutect2_normal_vcf.sh
qsub -v
SAMPLE_ID="/home/polly_hung/ccoc/samples/batch_by_type/normal/batch_4
.txt" -N "batch_4" 01_mutect2_normal_vcf.sh
qsub -v
SAMPLE_ID="/home/polly_hung/ccoc/samples/batch_by_type/normal/batch_5
.txt" -N "batch_5" 01_mutect2_normal_vcf.sh
qsub -v
SAMPLE_ID="/home/polly_hung/ccoc/samples/batch_by_type/normal/batch_6
.txt" -N "batch_6" 01_mutect2_normal_vcf.sh
```

GATK Re-calibration

```
#!/bin/bash
#PBS -l nodes=1:ppn=2
#PBS -l mem=10g
#PBS -l walltime=4:00:00
#PBS -m ae
#PBS -N FT190_NT
#PBS -q small

# library modules
echo "Start loading modules at $(date)"
module load miniconda3
module load GenomeAnalysisTK/4.2.0.0
source activate /software/GenomeAnalysisTK/4.2.0.0
echo "Finish loading modules at $(date)"

# HOME
HOME="/home/polly_hung/sWGS"
REF="/home/polly_hung/reference"

# Folders
DATA="$HOME/data"
RESULT="$HOME/output"
RECAL="$HOME/recalibration"
ALIGNMENT_SUMMARY="$HOME/alignment_summary"
VCF_REF="$REF/vcf/new_vcf"

# Files
HG38_REF="$REF/hg38/Homo_sapiens.GRCh38.dna.primary_assembly.fa"
KNOWN_INDELS="$VCF_REF/Homo_sapiens_assembly38.known_indels.vcf.gz"
MILL_INDELS="$VCF_REF/Mills.indels.contig.adjusted.hg38.vcf.gz"
DBSNP="$VCF_REF/Homo_sapiens_assembly38.dbsnp138.vcf.gz"

cd "$RESULT"
sample_id="FT190_NT"

# Define Variables
echo "Processing sample_id: $sample_id"
MARKED_BAM="$RESULT/$sample_id.sorted.marked.bam"
CALIBRATED_BAM="$RESULT/$sample_id.sorted.marked.calibrated.bam"
RECAL_TABLE="$RECAL/$sample_id.recal_data.table"
```

```
RECAL_TABLE_AFTER="$RECAL/$sample_id.recal_data_calibrated.table"
ASM_TABLE="$ALIGNMENT_SUMMARY/$sample_id.alignment_summary.txt"

# make the Marked bam file modifiable
chmod +x "$MARKED_BAM"

# base recalibrator
gatk BaseRecalibrator \
-I "$MARKED_BAM" \
-R "$HG38_REF" \
--known-sites "$KNOWN_INDELS" \
--known-sites "$MILL_INDELS" \
--known-sites "$DBSNP" \
-O "$RECAL_TABLE"

# apply recalibrator
gatk ApplyBQSR \
-R "$HG38_REF" \
-I "$MARKED_BAM" \
--bqsr-recal-file "$RECAL_TABLE" \
-O "$CALIBRATED_BAM"

# get base recalibrator score for the calibrated sample
gatk BaseRecalibrator \
-I "$CALIBRATED_BAM" \
-R "$HG38_REF" \
--known-sites "$KNOWN_INDELS" \
--known-sites "$MILL_INDELS" \
--known-sites "$DBSNP" \
-O "$RECAL_TABLE_AFTER"
```

qsub -v SAMPLE_ID="/home/polly_hung/sWGS/g1.txt" -N "g1" alignment.sh
qsub -v SAMPLE_ID="/home/polly_hung/sWGS/I.txt" -N "I" compiled.sh
qsub -v SAMPLE_ID="/home/polly_hung/sWGS/H.txt" -N "H" compiled.sh
qsub -v SAMPLE_ID="/home/polly_hung/sWGS/FT190.txt" -N "FT190" compiled.sh
qsub -v SAMPLE_ID="/home/polly_hung/sWGS/F.txt" -N "F" compiled.sh

qsub -v sample_id="/home/polly_hung/RNA-seq/CPOS_Data_20250210/batch1.txt" -N "batch1" salmon.sh
qsub -v sample_id="/home/polly_hung/RNA-seq/CPOS_Data_20250210/batch2.txt" -N "batch2" salmon.sh

qsub -v sample_id="/home/polly_hung/RNA-seq/CPOS_Data_20250210/batch3.txt" -N "batch3" salmon.sh