Workshop 2024 commands

1. Trimming

```
trimmomatic PE sequences/${Sample}_*_R1_*.fastq.gz
sequences/${Sample}_*_R2_*.fastq.gz -baseout output/${Sample}.fq.gz
ILLUMINACLIP:$adapters:2:30:10:2:keepBothReads LEADING:3 SLIDINGWINDOW:4:15
MINLEN:40
```

2. Paired end assembly

```
flash output/${Sample}_1P.fq.gz output/${Sample}_2P.fq.gz --cap-mismatch-quals -O
-M 250 -o output/${Sample}
```

3. Mapping to reference genome

```
bwa mem -R "@RG\tID:AML\tPL:ILLUMINA\tLB:LIB-MIPS\tSM:24NGS457-B1\tPI:200" -M -t
20 $genome output/${Sample}.extendedFrags.fastq > output/${Sample}.sam
```

4. Sam conversion

```
samtools view -b output/${Sample}.sam > output/${Sample}.bam
samtools sort output/${Sample}.bam > output/${Sample}.sorted.bam
samtools index output/${Sample}.sorted.bam > output/${Sample}.sorted.bam.bai
```

5. GATK Best Practices for data pre-processing

Details of this step can be found here : https://gatk.broadinstitute.org/hc/en-us/articles/360035535912-Data-pre-processing-for-variant-discovery

```
java -Xmx8G -jar $GATK38_path -T RealignerTargetCreator -R $genome -nt 10 -I
output/24NGS775-A1.sorted.bam --known $site1 -o output/${Sample}.intervals

java -Xmx8G -jar ${GATK38_path} -T IndelRealigner -R ${genome} -I output/24NGS775-
A1.sorted.bam -known ${site1} --targetIntervals output/24NGS775-A1.intervals -o
output/${Sample}.realigned.bam

java -Xmx8G -jar ${GATK38_path} -T BaseRecalibrator -R ${genome} -I
output/24NGS775-A1.realigned.bam -knownSites ${site2} -knownSites ${site3} -
maxCycle 600 -o output/${Sample}.recal_data.table${params.site3} -maxCycle 600 -o
```

```
${Sample}.recal_data.table

java -Xmx8G -jar ${GATK38_path} -T PrintReads -R ${genome} -I output/24NGS775-
A1.realigned.bam --BQSR output/24NGS775-A1.recal_data.table -o
output/${Sample}.final.bam
```

6. Coverage calculation

```
bedtools bamtobed -i output/${Sample}.sorted.bam > output/${Sample}.sorted.bed
bedtools coverage -counts -a $bedfile.bed -b output/${Sample}.sorted.bed >
output/${Sample}.counts.bed
```

7. Variant calling

```
VarDict -G $genome -f 0.01 -N ${Sample} -b output/${Sample}.sorted.bam -c 1 -S 2 -
E 3 -g 4 $bedfile.bed | sed '1d' | teststrandbias.R | var2vcf_valid.pl -N
${Sample} -E -f 0.01 > output/${Sample}_vardict.vcf
```

8. Variant annotation

```
convert2annovar.pl -format vcf4 output/${Sample}_vardict.vcf --outfile
output/${Sample}.avinput --withzyg --includeinfo

table_annovar.pl output/${Sample}.avinput --out output/${Sample}_final --remove --
protocol refGene,cosmic84,exac03 --operation g,f,f --buildver hg19 --nastring '-1'
--otherinfo --csvout $database
```

9. Format output

```
python3 $formatVardict_script_path output/${Sample}_final.hg19_multianno.csv
${Sample} output/
```