**The first pipeline**

First of all, I worked on a dataset SRR8115017.fastq, I have decided to work with bowtie2 for alignment because I read their documentation I found their indexing is more efficient than BWA in addition to it’s manual is more readable and informative, so before alignment step I made me mind to read the whole manual and understand almost their arguments.

**Downloading the ref.**

Wget ftp://[ftp.ensembl.org/pub/release-96/fasta/homo\_sapiens/dna/Homo\_sapiens.GRCh38.dna.chromosome.21.fa.gz](http://ftp.ensembl.org/pub/release-96/fasta/homo_sapiens/dna/Homo_sapiens.GRCh38.dna.chromosome.21.fa.gz)

**Building ref. index**

bowtie2-build ~/BaseRecalibration\_Benchmarking/Homo\_sapiens.GRCh38.dna.chromosome.21.fa index\_two\_bowtie2/Homo\_sapiens.fa

Before beginning in the alignment step, I remember from the assignment that when I run the haplotypecaller, It has an error that due to the sample name and something like that, so I decided to include in the SAM file the header informations that includes the read group, platform unit, library preparation and sample name, I know that I am working with just one sample so it has the same library preparation in addition to other parameters but I feel that their absence is the cause of haplotypecaller error.

R1="$HOME/BaseRecalibration\_Benchmarking/SRR8115017.fastq.gz”

RGID=$(cat $R1 | head -n1 | sed 's/:/\_/g' |cut -d "." -f1)

PU=$RGID.$LB

LB="SRR8115017\_same"

PL="Illumina"

**Alignment step:**

**bowtie2 -p 20 -q --no-unal -x index/Homo\_sapiens.fa -U SRR8115017.fastq.gz --rg-id $RGID --rg SM=$SM --rg PL=$PL --rg LB=$LB --rg PU=$PU 2> align\_stats.txt| samtools view -Sb -o test\_two.bam**

**-p** is the number of threads

**-q--no unal** suppress SAM records for reads that fails to alignment

However, after sorting by samtools and begin in the step of Markduplicate this error appeared

**Error in markduplicate:**

Exception in thread "main" htsjdk.samtools.SAMFormatException: Error parsing SAM header. Problem parsing @RG key:value pair. Line:

@RG ID:@SRR8115017 SM=SRR8115017 PL=Illumina LB=SRR8115017\_same PU=@SRR8115017.; File /home/nourelislam/BaseRecalibration\_Benchmarking/test\_two.bam; Line number 3

So I realized that the error is due to the formatting of the SAM header, and after many trials, I found the error in the “ = ” sign, so I’ve changed it to “ : ” and repeat the alignment step and the Markduplicate step then the runnin is going smoothly.

bowtie2 -p 20 -q --no-unal -x index\_two\_bowtie2/Homo\_sapiens.fa -U SRR8115017.fastq.gz --rg-id $RGID --rg SM:$SM --rg PL:$PL --rg LB:$LB --rg PU:$PU 2> align\_stats.txt| samtools view -Sb -o bowtie2.bam

**Sorting**:

samtools sort bowtie2.bam -o SRR8115017.sorted.bam

**Mark-duplicates:**

picard\_path=$CONDA\_PREFIX/share/picard-2.19.0-0

java -Xmx2g -jar $picard\_path/picard.jar MarkDuplicates INPUT=SRR8115017.sorted.bam OUTPUT=SRR8115017.dedup.bam METRICS\_FILE=SRR8115017.metrics.txt

**Indexing**

java -Xmx2g -jar $picard\_path/picard.jar BuildBamIndex VALIDATION\_STRINGENCY=LENIENT INPUT=SRR8115017.dedup.bam

java -Xmx2g -jar $picard\_path/picard.jar CreateSequenceDictionary R=Homo\_sapiens.GRCh38.dna.chromosome.21.fa O=Homo\_sapiens.GRCh38.dna.chromosome.21.dict

samtools faidx Homo\_sapiens.GRCh38.dna.chromosome.21.fa

**Downloading known variants:**

**wget ftp://ftp.ensembl.org/pub/release-96/variation/vcf/homo\_sapiens/homo\_sapiens-chr21.vcf.gz**

**Indexing:**

**gatk IndexFeatureFile -F Homo\_sapiens\_chr21.vcf**

**Error A USER ERROR has occurred: Error while trying to create index for /home/nourelislam/BaseRecalibration\_Benchmarking/Homo\_sapiens\_chr21.vcf. Error was: htsjdk.tribble.TribbleException: The provided VCF file is malformed at approximately line number 1291601: empty alleles are not permitted in VCF records**

After searching about this error which is a normal error in the vcf file, some suggested to completely remove the record having that malformation, so when I run the following code and remove the specific malformed record, it runs smoothly.

**head -1291601 Homo\_sapiens\_chr21.vcf | tail -1** Then remove this record

**Base recalibration:**

gatk --java-options "-Xmx2G" BaseRecalibrator -R Homo\_sapiens.GRCh38.dna.chromosome.21.fa -I SRR8115017.dedup.bam --known-sites Homo\_sapiens\_chr21.vcf -O SRR8115017.report

gatk --java-options "-Xmx2G" ApplyBQSR -R Homo\_sapiens.GRCh38.dna.chromosome.21.fa -I SRR8115017.dedup.bam -bqsr SRR8115017.report -O SRR8115017.bqsr.bam --add-output-sam-program-record --emit-original-quals

**Haplotypecaller:**

**Without base rec.**

gatk --java-options "-Xmx2G" HaplotypeCaller -R Homo\_sapiens.GRCh38.dna.chromosome.21.fa -I SRR8115017.dedup.bam --emit-ref-confidence GVCF --pcr-indel-model NONE -O SRR8115017.gvcf

**Annotation:**

gatk --java-options "-Xmx60G" GenotypeGVCFs -R Homo\_sapiens.GRCh38.dna.chromosome.21.fa -V SRR8115017.bqsr.gvcf --max-alternate-alleles 2 --dbsnp Homo\_sapiens\_chr21.vcf -O SRR8115017\_bqsr\_ann.vcf

**With base rec.**

gatk --java-options "-Xmx2G" HaplotypeCaller -R Homo\_sapiens.GRCh38.dna.chromosome.21.fa -I SRR8115017.bqsr.bam --emit-ref-confidence GVCF --pcr-indel-model NONE -O SRR8115017.bqsr.gvcf

**Annotation:**

gatk --java-options "-Xmx60G" GenotypeGVCFs -R Homo\_sapiens.GRCh38.dna.chromosome.21.fa -V SRR8115017.gvcf --max-alternate-alleles 2 --dbsnp Homo\_sapiens\_chr21.vcf -O SRR8115017\_ann.vcf