**The second pipeline**

After I noticed that there is no enough variants in the gvcf file that was created in the 1st pipeline, there was roughly one page of variants in the file, I thought that all of that due to the bad mapping quality with chr21 which was 33.83%, hence the mapping 33% so the variants in those mapped reads presents in small extent. In other words, there is no enough variants in those low mapped reads to generate a large variants in vcf file.

At this point, I made my mind to mapping against the whole human genome to have enough and strong mapping rate that will generate a large number of variants.

**Downloading the whole ref genome:**

https://www.gencodegenes.org/human/release\_30.html/Genome sequence (GRCh38.p12)

**Build index of bowtie2 for whole human genome**

bowtie2-build ~/BaseRecalibration\_Benchmarking/GRCh38.p12.genome.fa.gz index\_whole\_genome/whole\_genome.fa

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

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

PU=$RGID.$LB

LB="SRR8115017\_same"

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

**Sorting**:

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

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

**Mark-duplicates:**

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

**Indexing**

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

java -Xmx2g -jar $picard\_path/picard.jar CreateSequenceDictionary R=whole\_genome.fa O=whole\_genome.dict

samtools faidx whole\_genome.fa

**Downloading whole genome known variants:**

wget ftp://ftp.ensembl.org/pub/release-89/variation/vcf/homo\_sapiens/Homo\_sapiens.vcf.gz

**Haplotypecaller without base rec.**

gatk --java-options "-Xmx2G" HaplotypeCaller -R whole\_genome.fa -I SRR8115017\_whole.dedup.bam --emit-ref-confidence GVCF --pcr-indel-model NONE -O SRR8115017\_whole.gvcf

Error

.HaplotypeCaller - Shutting down engine [May 10, 2019 12:25:00 AM EET] org.broadinstitute.hellbender.tools.walkers.haplotypecaller.HaplotypeCaller done. Elapsed time: 2.57 minutes.

Runtime.totalMemory()=2145386496

Exception in thread "main" java.lang.OutOfMemoryError: Java heap space

As expected the gvcf file has more variants but it did not proceed till the end due to run out of memory.