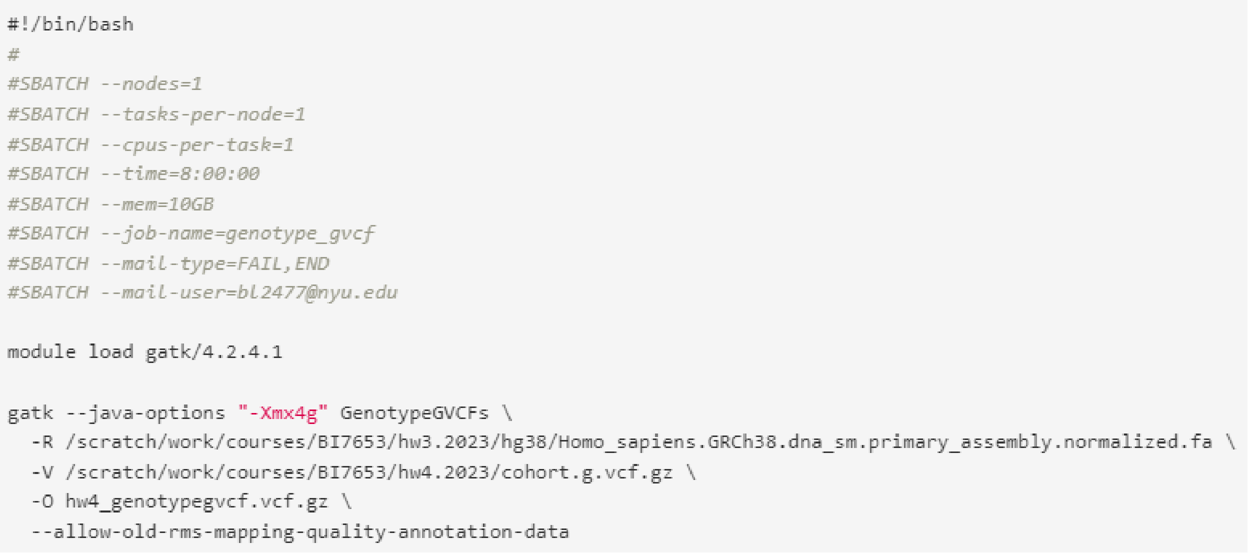
Khan Inan
Transcriptomics 7653
Assignment week 4

Q1.1 Please include contents of your job submission script [2 points].



```
#!/bin/bash
#
#SBATCH --nodes=1
#SBATCH --tasks-per-node=1
#SBATCH --cpus-per-task=1
#SBATCH --time=8:00:00
#SBATCH --mem=10GB
#SBATCH --job-name=genotype_gvcf
#SBATCH --mail-type=FAIL,END
#SBATCH --mail-user=bl2477@nyu.edu

module load gatk/4.2.4.1

gatk --java-options "-Xmx4g" GenotypeGVCFs \
  -R /scratch/work/courses/BI7653/hw3.2023/hg38/Homo_sapiens.GRCh38.dna_sm.primary_assembly.normalized.fa \
  -V /scratch/work/courses/BI7653/hw4.2023/cohort.g.vcf.gz \
  -O hw4_genotypegvcf.vcf.gz \
  --allow-old-rms-mapping-quality-annotation-data
```

Q1.2 When your script has completed, report the first 20 lines of the output gzipped vcf [1 point]:

(next 2 pages)

```
##fileformat=VCFv4.2
##ALT=<ID=NON_REF,Description="Represents any possible alternative allele at this location">
##FILTER=<ID=LowQual,Description="Low quality">
##FORMAT=<ID=AD,Number=R,Type=Integer,Description="Allelic depths for the ref and alt alleles in the order listed">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth (reads with MQ=255 or with bad mates are filtered)">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=MIN_DP,Number=1,Type=Integer,Description="Minimum DP observed within the GVCF block">
##FORMAT=<ID=PGT,Number=1,Type=String,Description="Physical phasing haplotype information, describing how the alternate alleles are phased in relation to one another">
##FORMAT=<ID=PID,Number=1,Type=String,Description="Physical phasing ID information, where each unique ID within a given sample (but not across samples) connects records within a phasing group">
##FORMAT=<ID=PL,Number=G,Type=Integer,Description="Normalized, Phred-scaled likelihoods for genotypes as defined in the VCF specification">
##FORMAT=<ID=RGQ,Number=1,Type=Integer,Description="Unconditional reference genotype confidence, encoded as a phred quality -10*log10 p(genotype call is wrong)">
##FORMAT=<ID=SB,Number=4,Type=Integer,Description="Per-sample component statistics which comprise the Fisher's Exact Test to detect strand bias.">
##GATKCommandLine=<ID=CombineGVCFs,CommandLine="CombineGVCFs  --output cohort.intervals.g.vcf.gz --variant /scratch/courses/BI7653/hw4.2019/gvcfs.list --intervals 1:1-5000000 --intervals 2:1-5000000 --intervals 3:1-5000000 --reference /scratch/courses/BI7653/hw3.2019/hg38/Homo_sapiens.GRCh38.dna_sm.primary_assembly.normalized.fa  --annotation-group StandardAnnotation --disable-tool-default-annotations false --convert-to-base-pair-resolution false --break-bands-at-multiples-of 0 --ignore-variants-starting-outside-interval false --interval-set-rule UNION --interval-padding 0 --interval-exclusion-padding 0 --interval-merging-rule ALL --read-validation-stringency SILENT --seconds-between-progress-updates 10.0 --disable-sequence-dictionary-validation false --create-output-bam-index true --create-output-bam-md5 false --create-output-variant-index true --create-output-variant-md5 false --lenient false --add-output-sam-program-record true --add-output-vcf-command-line true --cloud-prefetch-buffer 40 --cloud-index-prefetch-buffer -1 --disable-bam-index-caching false --help false --version false --showHidden false --verbosity INFO --QUIET false --use-jdk-deflater false --use-jdk-inflater false --gcs-max-retries 20 --disable-tool-default-read-filters false",Version=4.0.2.1,Date="September 25, 2019 5:45:02 PM EDT">
##GATKCommandLine=<ID=GenotypeGVCFs,CommandLine="GenotypeGVCFs --output hw4_genotypegvcf.vcf.gz --variant /scratch/work/courses/BI7653/hw4.2023/cohort.g.vcf.gz --reference /scratch/work/courses/BI7653/hw3.2023/hg38/Homo_sapiens.GRCh38.dna_sm.primary_assembly.normalized.fa --allow-old-rms-mapping-quality-annotation-data true --include-non-variant-sites false --merge-input-intervals false --input-is-somatic false --tumor-lod-to-emit 3.5 --allele-fraction-error 0.001 --keep-combined-raw-annotations false --use-posteriors-to-calculate-qual false --dont-use-dragstr-priors false --use-new-qual-calculator true --annotate-with-num-discovered-alleles false --heterozygosity 0.001 --indel-heterozygosity 1.25E-4 --heterozygosity-stdev 0.01 --standard-min-confidence-threshold-for-calling 30.0 --max-alternate-alleles 6 --max-genotype-count 1024 --sample-ploidy 2 --num-reference-samples-if-no-call 0 --genotype-assignment-method USE_PLS_TO_ASSIGN --call-genotypes false --genomicsdb-use-bcf-codec false --genomicsdb-shared-posixfs-optimizations false --genomicsdb-use-gcs-hdfs-connector false --only-output-calls-starting-in-intervals false --interval-set-rule UNION --interval-padding 0 --interval-exclusion-padding 0 --interval-merging-rule ALL --read-validation-stringency SILENT --seconds-between-progress-updates 10.0 --disable-sequence-dictionary-validation false --create-output-bam-index true --create-output-bam-md5 false --create-output-variant-index true --create-output-variant-md5 false --max-variants-per-shard 0 --lenient false --add-output-sam-program-record true --add-output-vcf-command-line true --cloud-prefetch-buffer 40 --cloud-index-prefetch-buffer -1 --disable-bam-index-caching false --sites-only-vcf-output false --help false --version false --showHidden false --verbosity INFO --QUIET false --use-jdk-deflater false --use-jdk-inflater f
```

```
alse --gcs-max-retries 20 --gcs-project-for-requester-pays  --disable-tool-default-read-filters false --disable-tool-default
-annotations false --enable-all-annotations false",Version="4.2.4.1",Date="February 21, 2023 9:18:43 PM EST">
##GATKCommandLine=<ID=HaplotypeCaller,CommandLine="HaplotypeCaller  --emit-ref-confidence GVCF --output NA19098.g.vcf --inpu
t NA19098.sorted.markdups.bam --reference /scratch/courses/BI7653/hw3.2019/hg38/Homo_sapiens.GRCh38.dna_sm.primary_assembly.
normalized.fa  --annotation-group StandardAnnotation --annotation-group StandardHCAnnotation --disable-tool-default-annotati
ons false --gvcf-gq-bands 1 --gvcf-gq-bands 2 --gvcf-gq-bands 3 --gvcf-gq-bands 4 --gvcf-gq-bands 5 --gvcf-gq-bands 6 --gvcf
-gq-bands 7 --gvcf-gq-bands 8 --gvcf-gq-bands 9 --gvcf-gq-bands 10 --gvcf-gq-bands 11 --gvcf-gq-bands 12 --gvcf-gq-bands 13
--gvcf-gq-bands 14 --gvcf-gq-bands 15 --gvcf-gq-bands 16 --gvcf-gq-bands 17 --gvcf-gq-bands 18 --gvcf-gq-bands 19 --gvcf-gq-
bands 20 --gvcf-gq-bands 21 --gvcf-gq-bands 22 --gvcf-gq-bands 23 --gvcf-gq-bands 24 --gvcf-gq-bands 25 --gvcf-gq-bands 26 -
-gvcf-gq-bands 27 --gvcf-gq-bands 28 --gvcf-gq-bands 29 --gvcf-gq-bands 30 --gvcf-gq-bands 31 --gvcf-gq-bands 32 --gvcf-gq-b
ands 33 --gvcf-gq-bands 34 --gvcf-gq-bands 35 --gvcf-gq-bands 36 --gvcf-gq-bands 37 --gvcf-gq-bands 38 --gvcf-gq-bands 39 --
gvcf-gq-bands 40 --gvcf-gq-bands 41 --gvcf-gq-bands 42 --gvcf-gq-bands 43 --gvcf-gq-bands 44 --gvcf-gq-bands 45 --gvcf-gq-ba
nds 46 --gvcf-gq-bands 47 --gvcf-gq-bands 48 --gvcf-gq-bands 49 --gvcf-gq-bands 50 --gvcf-gq-bands 51 --gvcf-gq-bands 52 --g
vcf-gq-bands 53 --gvcf-gq-bands 54 --gvcf-gq-bands 55 --gvcf-gq-bands 56 --gvcf-gq-bands 57 --gvcf-gq-bands 58 --gvcf-gq-ban
ds 59 --gvcf-gq-bands 60 --gvcf-gq-bands 70 --gvcf-gq-bands 80 --gvcf-gq-bands 90 --gvcf-gq-bands 99 --indel-size-to-elimina
te-in-ref-model 10 --use-alleles-trigger false --disable-optimizations false --just-determine-active-regions false --dont-ge
notype false --dont-trim-active-regions false --max-disc-ar-extension 25 --max-gga-ar-extension 300 --padding-around-indels
150 --padding-around-snps 20 --kmer-size 10 --kmer-size 25 --dont-increase-kmer-sizes-for-cycles false --allow-non-unique-km
ers-in-ref false --num-pruning-samples 1 --recover-dangling-heads false --do-not-recover-dangling-branches false --min-dangl
ing-branch-length 4 --consensus false --max-num-haplotypes-in-population 128 --error-correct-kmers false --min-pruning 2 --d
ebug-graph-transformations false --kmer-length-for-read-error-correction 25 --min-observations-for-kmer-to-be-solid 20 --lik
elihood-calculation-engine PairHMM --base-quality-score-threshold 18 --pair-hmm-gap-continuation-penalty 10 --pair-hmm-imple
mentation FASTEST_AVAILABLE --pcr-indel-model CONSERVATIVE --phred-scaled-global-read-mismapping-rate 45 --native-pair-hmm-t
hreads 4 --native-pair-hmm-use-double-precision false --debug false --use-filtered-reads-for-annotations false --bam-writer-
type CALLED_HAPLOTYPES --dont-use-soft-clipped-bases false --capture-assembly-failure-bam false --error-correct-reads false
--do-not-run-physical-phasing false --min-base-quality-score 10 --smith-waterman JAVA --use-new-qual-calculator false --anno
tate-with-num-discovered-alleles false --heterozygosity 0.001 --indel-heterozygosity 1.25E-4 --heterozygosity-stdev 0.01 --s
tandard-min-confidence-threshold-for-calling 10.0 --max-alternate-alleles 6 --max-genotype-count 1024 --sample-ploidy 2 --ge
notyping-mode DISCOVERY --contamination-fraction-to-filter 0.0 --output-mode EMIT_VARIANTS_ONLY --all-site-pls false --min-a
ssembly-region-size 50 --max-assembly-region-size 300 --assembly-region-padding 100 --max-reads-per-alignment-start 50 --act
ive-probability-threshold 0.002 --max-prob-propagation-distance 50 --interval-set-rule UNION --interval-padding 0 --interval
-exclusion-padding 0 --interval-merging-rule ALL --read-validation-stringency SILENT --seconds-between-progress-updates 10.0
--disable-sequence-dictionary-validation false --create-output-bam-index true --create-output-bam-md5 false --create-output-
variant-index true --create-output-variant-md5 false --lenient false --add-output-sam-program-record true --add-output-vcf-c
ommand-line true --cloud-prefetch-buffer 40 --cloud-index-prefetch-buffer -1 --disable-bam-index-caching false --help false
--version false --showHidden false --verbosity INFO --QUIET false --use-jdk-deflater false --use-jdk-inflater false --gcs-ma
x-retries 20 --disable-tool-default-read-filters false --minimum-mapping-quality 20",Version=4.0.2.1,Date="September 23, 201
9 9:55:52 PM EDT">
##INFO=<ID=AC,Number=A,Type=Integer,Description="Allele count in genotypes, for each ALT allele, in the same order as liste
d">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency, for each ALT allele, in the same order as listed">
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">
##INFO=<ID=BaseQRankSum,Number=1,Type=Float,Description="Z-score from Wilcoxon rank sum test of Alt Vs. Ref base qualities">
```

Q1.3 The GenotypeGVCFs command will produce a VCF file with both indels and snps. You can use a grep command that excludes header lines beginning with a “#” and extract only variant records from a VCF.

How many total variants are in the VCF file [1 point]?

There seems to be 91636 total variants

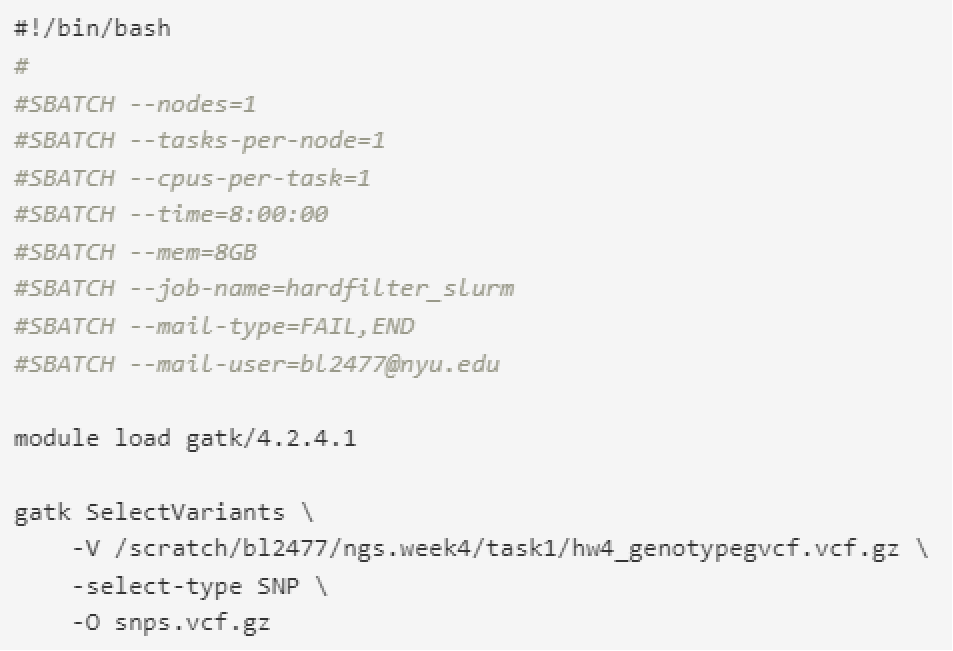
Q1.4a The HaplotypeCaller + CombineGVCFs + GenotypeGVCFs workflow addresses what is known as the n + 1 problem. What is the n + 1 problem?

Doing the BAM to VCF file conversion in one step is computationally intensive and inefficient. This results in the N+1 problem because if one sample was added to the data, you have to repeat the process from the beginning. HaplotypeCaller + CombineGVCFs + GenotypeGVCFs workflow solves this because it is a two step method that involves converting the bam files to a vcf file and then running the gVCFs as input for multiple sample SNP genotyping and calling.

Q1.4b If after completing your assignment your instructor provides you with an additional .gvcf file to include in your snp callset, which steps in the workflow would you need to re-execute to generate a VCF with all samples?

We have to combine the .gvcf files from each sample by running combineGVCFs to make a multiple sample .gvcf. After its done, you have to run genotypeGVCFs for the multisample calling and genotyping.

Q2.1 Paste the contents of your script into your answers file [1 point].



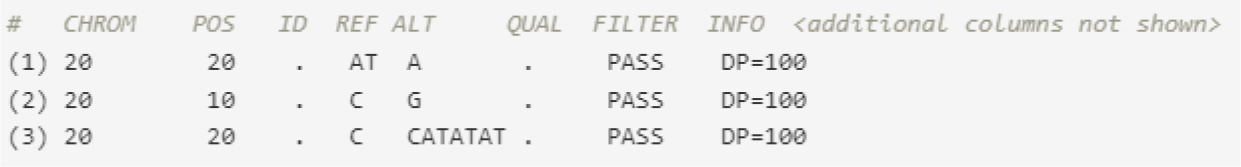
```
#!/bin/bash
#
#SBATCH --nodes=1
#SBATCH --tasks-per-node=1
#SBATCH --cpus-per-task=1
#SBATCH --time=8:00:00
#SBATCH --mem=8GB
#SBATCH --job-name=hardfilter_slurm
#SBATCH --mail-type=FAIL,END
#SBATCH --mail-user=bl2477@nyu.edu

module load gatk/4.2.4.1

gatk SelectVariants \
    -V /scratch/bl2477/ngs.week4/task1/hw4_genotypegvcf.vcf.gz \
    -select-type SNP \
    -O snps.vcf.gz
```

Q2.2 Although we are not working with Indel ("insertion/deletion") variants here, they are important in many contexts including studies of frameshift mutations in protein-coding genes. Please review the VCF format specification for how indels are specified in VCF format in section 5, p. 13:

https://samtools.github.io/hts-specs/VCFv4.3.pdf

For **each** VCF-encoded variant below, answer the following.

1. Is the variant a SNP or indel?
2. If it is an indel, is the reference or the alternate allele the deletion allele?
3. If it is an indel, how many bases are deleted relative to the insertion allele?
4. If it is an indel, for each allele, which base is found at the genomic position in the POS column [1 point]?



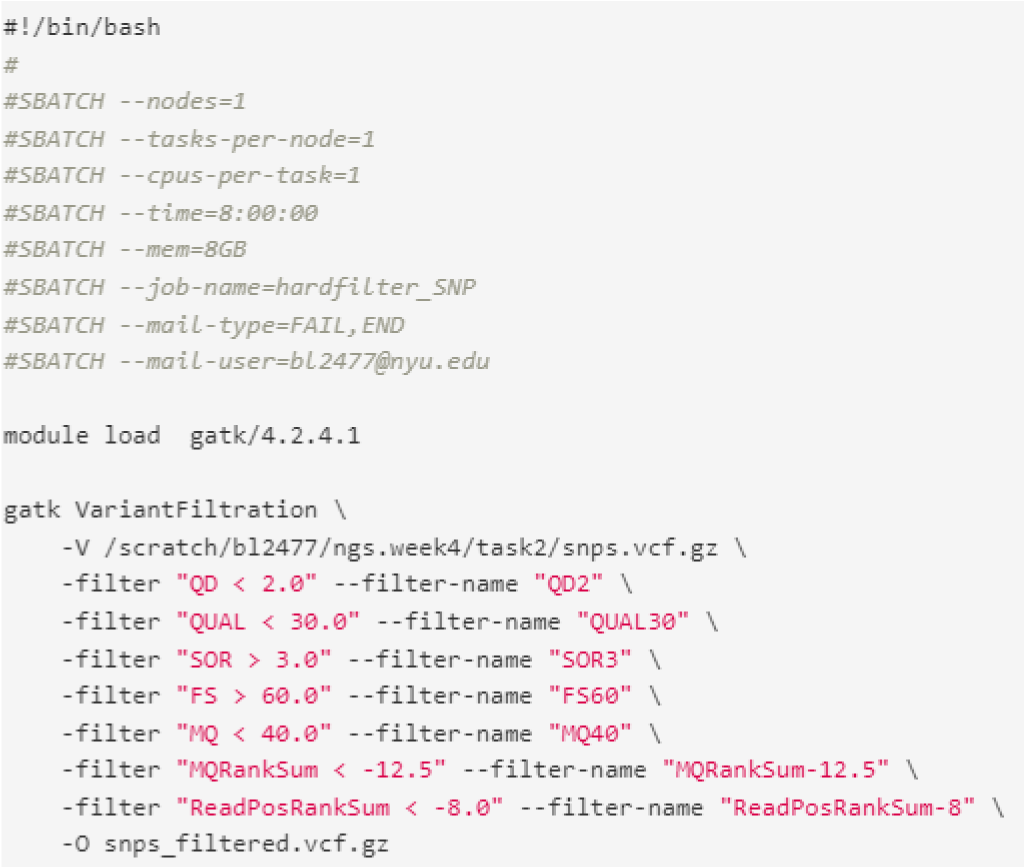
```
#   CHROM    POS   ID  REF ALT     QUAL  FILTER  INFO  <additional columns not shown>
(1) 20        20    .   AT  A        .     PASS    DP=100
(2) 20        10    .   C   G        .     PASS    DP=100
(3) 20        20    .   C   CATATAT  .     PASS    DP=100
```

The first record is an indel. There is a deletion allele for the alternate allele with the deletion of a single base (T at position 20). Base A is found at the reference allele position at position 20, at the same time, base A is also found at the alternate allele in position 20

The second record , the VCF is an SNP

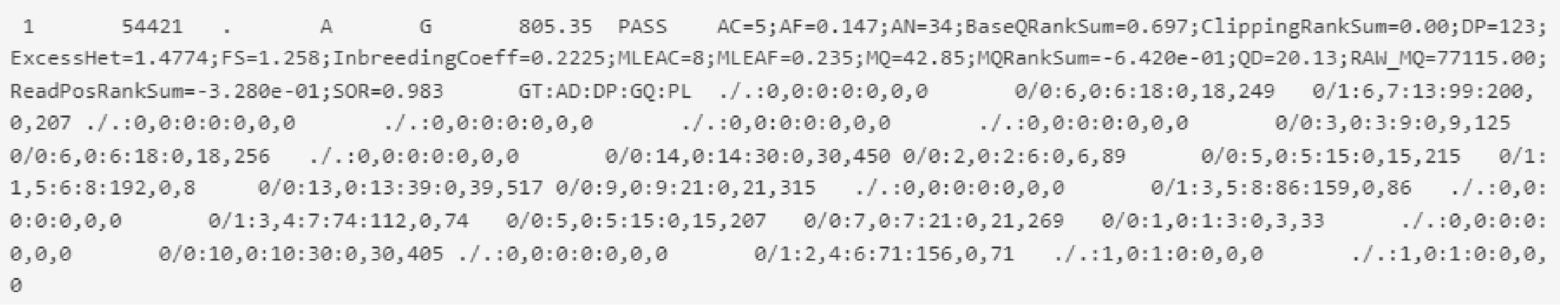
The third record is an indel. The reference allele is a deletion allele and its an insertion because the base C is being replace by ATATAT and C. 6 bases were deleted in relation to the insertion allele. At position 20 for both the reference allele and the alternate allele, the base C was found

Q3.1 Paste the contents of your script here [1 point].



```
#!/bin/bash
#
#SBATCH --nodes=1
#SBATCH --tasks-per-node=1
#SBATCH --cpus-per-task=1
#SBATCH --time=8:00:00
#SBATCH --mem=8GB
#SBATCH --job-name=hardfilter_SNP
#SBATCH --mail-type=FAIL,END
#SBATCH --mail-user=bl2477@nyu.edu

module load  gatk/4.2.4.1

gatk VariantFiltration \
    -V /scratch/bl2477/ngs.week4/task2/snps.vcf.gz \
    -filter "QD < 2.0" --filter-name "QD2" \
    -filter "QUAL < 30.0" --filter-name "QUAL30" \
    -filter "SOR > 3.0" --filter-name "SOR3" \
    -filter "FS > 60.0" --filter-name "FS60" \
    -filter "MQ < 40.0" --filter-name "MQ40" \
    -filter "MQRankSum < -12.5" --filter-name "MQRankSum-12.5" \
    -filter "ReadPosRankSum < -8.0" --filter-name "ReadPosRankSum-8" \
    -O snps_filtered.vcf.gz
```

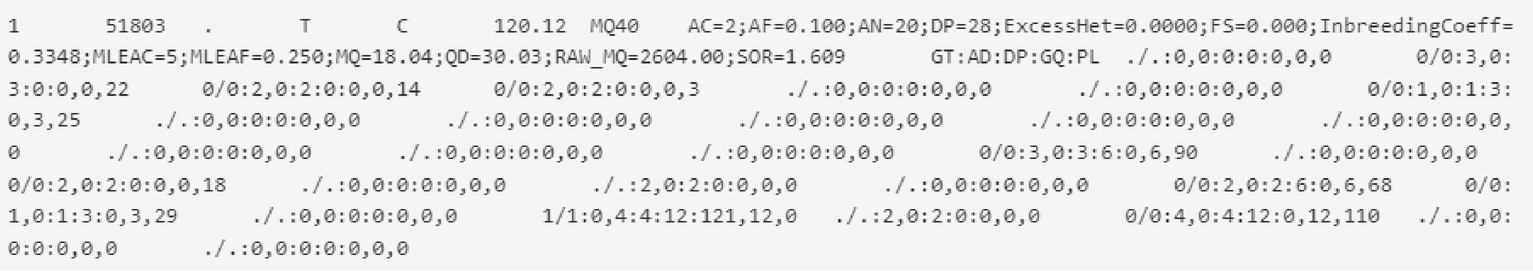
Q3.2a Report one record that passed the filtering criterion. What is the Depth of this variant across samples? What is the SNP quality?



```
 1       54421   .       A       G       805.35   PASS    AC=5;AF=0.147;AN=34;BaseQRankSum=0.697;ClippingRankSum=0.00;DP=123;
ExcessHet=1.4774;FS=1.258;InbreedingCoeff=0.2225;MLEAC=8;MLEAF=0.235;MQ=42.85;MQRankSum=-6.420e-01;QD=20.13;RAW_MQ=77115.00;
ReadPosRankSum=-3.280e-01;SOR=0.983      GT:AD:DP:GQ:PL  ./.:0,0:0:0:0,0,0       0/0:6,0:6:18:0,18,249   0/1:6,7:13:99:200,
0,207 ./.:0,0:0:0:0,0,0       ./.:0,0:0:0:0,0,0       ./.:0,0:0:0:0,0,0       ./.:0,0:0:0:0,0,0       0/0:3,0:3:9:0,9,125
0/0:6,0:6:18:0,18,256   ./.:0,0:0:0:0,0,0       0/0:14,0:14:30:0,30,450 0/0:2,0:2:6:0,6,89      0/0:5,0:5:15:0,15,215   0/1:
1,5:6:8:192,0,8     0/0:13,0:13:39:0,39,517 0/0:9,0:9:21:0,21,315   ./.:0,0:0:0:0,0,0       0/1:3,5:8:86:159,0,86   ./.:0,0:
0:0:0,0,0       0/1:3,4:7:74:112,0,74   0/0:5,0:5:15:0,15,207   0/0:7,0:7:21:0,21,269   0/0:1,0:1:3:0,3,33      ./.:0,0:0:0:
0,0,0       0/0:10,0:10:30:0,30,405 ./.:0,0:0:0:0,0,0       0/1:2,4:6:71:156,0,71   ./.:1,0:1:0:0,0,0       ./.:1,0:1:0:0,0,
0
```

The variant depth across the samples is 123. The quality of the SNP is 805.35

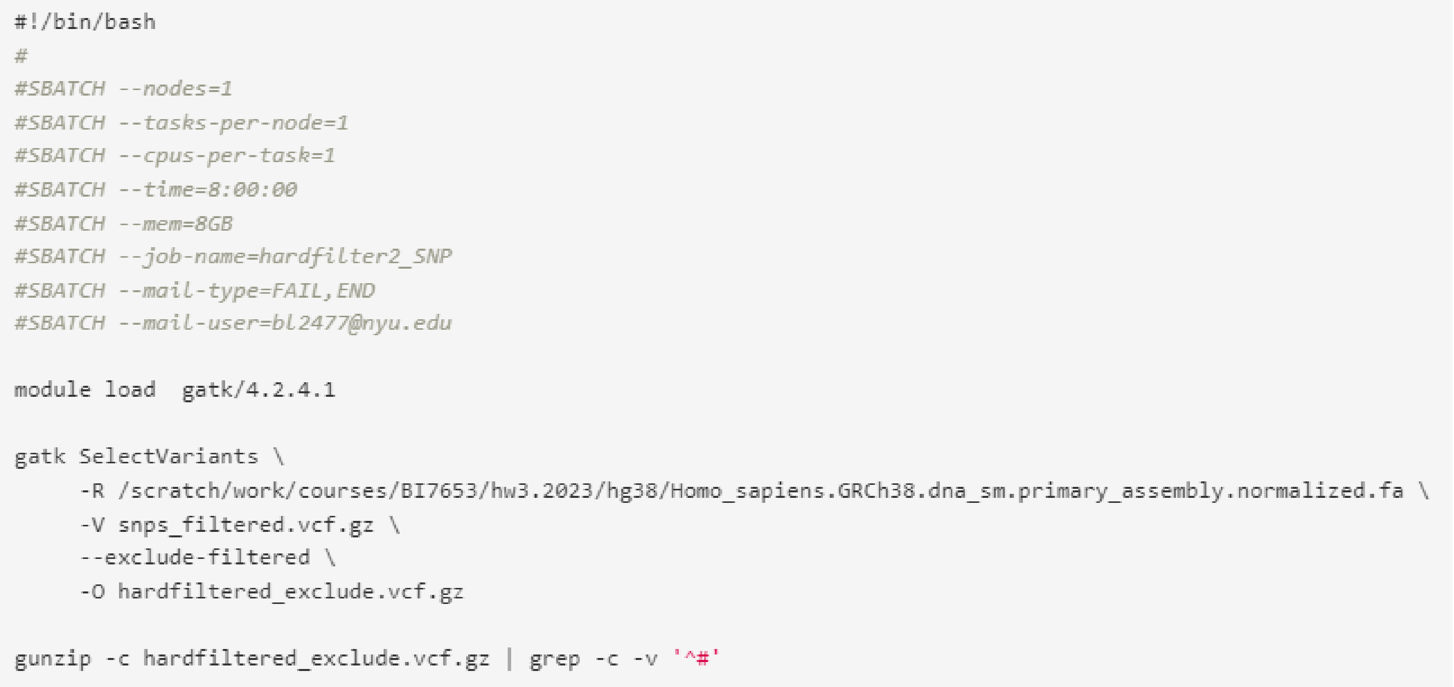
Q3.2b Report one record that failed one or more filters then answer which filters did it fail? What is the threshold of the filter(s) that it failed and what is the value(s) for the filter for the SNP in question?



```
1       51803   .       T       C       120.12  MQ40    AC=2;AF=0.100;AN=20;DP=28;ExcessHet=0.0000;FS=0.000;InbreedingCoeff=
0.3348;MLEAC=5;MLEAF=0.250;MQ=18.04;QD=30.03;RAW_MQ=2604.00;SOR=1.609       GT:AD:DP:GQ:PL  ./.:0,0:0:0:0,0,0       0/0:3,0:
3:0:0,0,22      0/0:2,0:2:0:0,0,14      0/0:2,0:2:0:0,0,3       ./.:0,0:0:0:0,0,0       ./.:0,0:0:0:0,0,0       0/0:1,0:1:3:
0,3,25      ./.:0,0:0:0:0,0,0       ./.:0,0:0:0:0,0,0       ./.:0,0:0:0:0,0,0       ./.:0,0:0:0:0,0,0       ./.:0,0:0:0:0,0,
0       ./.:0,0:0:0:0,0,0       ./.:0,0:0:0:0,0,0       ./.:0,0:0:0:0,0,0       0/0:3,0:3:6:0,6,90      ./.:0,0:0:0:0,0,0
0/0:2,0:2:0:0,0,18      ./.:0,0:0:0:0,0,0       ./.:2,0:2:0:0,0,0       ./.:0,0:0:0:0,0,0       0/0:2,0:2:6:0,6,68      0/0:
1,0:1:3:0,3,29      ./.:0,0:0:0:0,0,0       1/1:0,4:4:12:121,12,0   ./.:2,0:2:0:0,0,0       0/0:4,0:4:12:0,12,110   ./.:0,0:
0:0:0,0,0       ./.:0,0:0:0:0,0,0
```

One of the filters failed. There is a value that is less than the threshold value of 40. It seems to be MQ which is 18.04

Q3.3 Create a job submission script with the following command line to remove SNPs that failed the filter criteria from the VCF.



```
#!/bin/bash
#
#SBATCH --nodes=1
#SBATCH --tasks-per-node=1
#SBATCH --cpus-per-task=1
#SBATCH --time=8:00:00
#SBATCH --mem=8GB
#SBATCH --job-name=hardfilter2_SNP
#SBATCH --mail-type=FAIL,END
#SBATCH --mail-user=bl2477@nyu.edu

module load  gatk/4.2.4.1

gatk SelectVariants \
     -R /scratch/work/courses/BI7653/hw3.2023/hg38/Homo_sapiens.GRCh38.dna_sm.primary_assembly.normalized.fa \
     -V snps_filtered.vcf.gz \
     --exclude-filtered \
     -O hardfiltered_exclude.vcf.gz

gunzip -c hardfiltered_exclude.vcf.gz | grep -c -v '^#'
```

For the final filtered set, there are 74265 SNPs