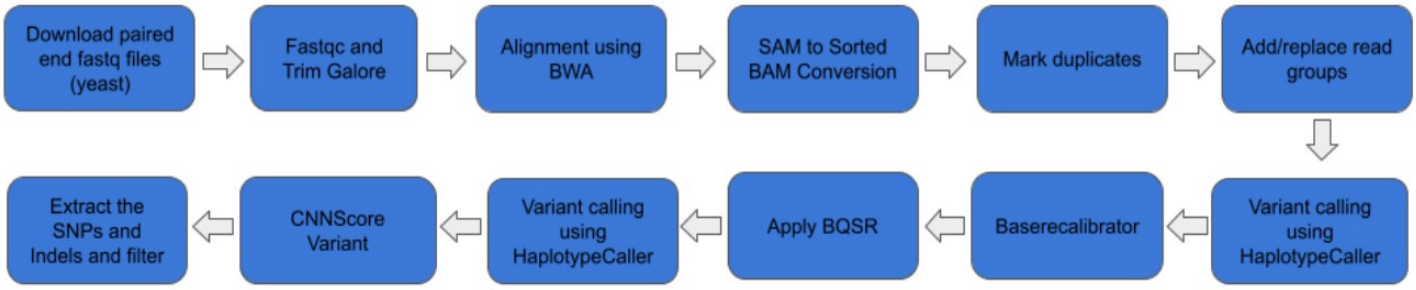
Assignment-7

Goal: Variant calling analysis using GATK followed by the filtration of variants (continuing from the Mark Duplicates step from Aug 11, 2025 lab).

Materials needed:

1. Fastq files - *Saccharomyces cerevisiae*; SRR18577795
2. Picard tool
3. Samtools
4. gatk

Workflow for GATK and variant annotation:



Exercise: Run a variant calling workflow for chromosome 20 using the UCSC reference genome, starting with a processed BAM file, and perform Base Quality Score Recalibration with known SNPs from different databases, variant calling with HaplotypeCaller, and subsequent extraction of SNPs and Indels.

Part A: Generate and Analyze Initial Variants

Steps:

samtools faidx hg38_chr20.fa

- FASTA files are plain text, so without an index, GATK (or samtools) would have to read line-by-line to find a chromosome/region.
- The .fai allows *random access*: HaplotypeCaller (or BQSR, etc.) can jump directly to chr20 (or any locus) instead of scanning the whole reference.

```
(base) ibab@IBAB-MScBDB2-Comp007:~/NGS/variant_calling$ samtools faidx hg38_chr20.fa
(base) ibab@IBAB-MScBDB2-Comp007:~/NGS/variant_calling$ ls
11aug_files  hg38_chr20.fa  hg38_chr20.fa.fai  hg38.fa  hg38.fa.fai  sratoolkit.3.2.1-ubuntu64  sratoolkit.current-ubuntu64.tar.gz
```

gatk CreateSequenceDictionary -R hg38_chr20.fa -O hg38_chr20.dict

- The .dict lists the names and lengths of contigs (e.g., chr1, chr2 … chr20).

- Picard/GATK tools require it to check the consistency between the BAM header and the reference genome.
- Without the .dict, GATK will complain: *"A sequence dictionary is required for the reference."*

```
(base) ibab@IBAB-MScBDB2-Comp007:~/NGS/variant_calling/18aug$ gatk CreateSequenceDictionary -R hg38_chr20.fa -O hg38_chr20.dict
Using GATK jar /home/ibab/miniconda3/share/gatk4-4.3.0.0-0/gatk-package-4.3.0.0-local.jar
Running:
    java -Dsamjdk.use_async_io_read_samtools=false -Dsamjdk.use_async_io_write_samtools=true -Dsamjdk.use_async_io_write_tribble=false -Dsamjdk.compression_level=2 -jar /home/ibab/miniconda3/share/gatk4-4.3.0.0-0/gatk-package-4.3.0.0-local.jar CreateSequenceDictionary -R hg38_chr20.fa -O hg38_chr20.dict
16:29:25.515 INFO  NativeLibraryLoader - Loading libgkl_compression.so from jar:file:/home/ibab/miniconda3/share/gatk4-4.3.0.0-0/gatk-package-4.3.0.0-local.jar!/com/intel/gkl/native/libgkl_compression.so
[Thu Aug 21 16:29:25 IST 2025] CreateSequenceDictionary --OUTPUT hg38_chr20.dict --REFERENCE hg38_chr20.fa --TRUNCATE_NAMES_AT_WHITESPACE true --NUM_SEQUENCES 2147483647 --VERBOSITY INFO --QUIET false --VALIDATION_STRINGENCY STRICT --COMPRESSION_LEVEL 2 --MAX_RECORDS_IN_RAM 500000 --CREATE_INDEX false --CREATE_MD5_FILE false --GA4GH_CLIENT_SECRETS client_secrets.json --help false --version false --showHidden false --USE_JDK_DEFLATER false --USE_JDK_INFLATER false
[Thu Aug 21 16:29:25 IST 2025] Executing as ibab@IBAB-MScBDB2-Comp007.ibab.ac.in on Linux 6.14.0-28-generic amd64; OpenJDK 64-Bit Server VM 11.0.1+13-LTS; Deflater: Intel; Inflater: Intel; Provider GCS is available; Picard version: Version:4.3.0.0
[Thu Aug 21 16:29:25 IST 2025] picard.sam.CreateSequenceDictionary done. Elapsed time: 0.01 minutes.
Runtime.totalMemory()=310378496
Tool returned:
0
(base) ibab@IBAB-MScBDB2-Comp007:~/NGS/variant_calling/18aug$ ls
hg38_chr20.dict  hg38_chr20.fa  hg38_chr20.fa.fai  hg38.fa  hg38.fa.fai
(base) ibab@IBAB-MScBDB2-Comp007:~/NGS/variant_calling/18aug$ ls -lh
total 3.2G
-rw-rw-r-- 1 ibab ibab  131 Aug 21 16:29 hg38_chr20.dict
-rw-rw-r-- 1 ibab ibab  63M Aug 21 14:56 hg38_chr20.fa
-rw-rw-r-- 1 ibab ibab   23 Aug 21 15:19 hg38_chr20.fa.fai
-rw-rw-r-- 1 ibab ibab 3.2G Oct 28  2022 hg38.fa
-rw-rw-r-- 1 ibab ibab  31K Aug 21 14:56 hg38.fa.fai
```

samtools index SRR15117878_grpadded.bam

- The BAM is binary and compressed; can't jump to chr20:10,000–20,000 unless an index exists.
- Variant callers (like HaplotypeCaller) often process one region at a time, so they need the .bai to fetch only the relevant reads quickly. Otherwise, forced to read the entire BAM into memory for every query.

```
(base) ibab@IBAB-MScBDB2-Comp007:~/NGS/variant_calling/18aug$ samtools index SRR15117878_grpadded.bam
(base) ibab@IBAB-MScBDB2-Comp007:~/NGS/variant_calling/18aug$ ls -lh
total 3.4G
-rw-rw-r-- 1 ibab ibab  131 Aug 21 16:29 hg38_chr20.dict
-rw-rw-r-- 1 ibab ibab  63M Aug 21 14:56 hg38_chr20.fa
-rw-rw-r-- 1 ibab ibab   23 Aug 21 15:19 hg38_chr20.fa.fai
-rw-rw-r-- 1 ibab ibab 3.2G Oct 28  2022 hg38.fa
-rw-rw-r-- 1 ibab ibab  31K Aug 21 14:56 hg38.fa.fai
-rw-rw-r-- 1 ibab ibab 162M Aug 18 17:35 SRR15117878_grpadded.bam
-rw-rw-r-- 1 ibab ibab  93K Aug 21 16:54 SRR15117878_grpadded.bam.bai
```

Part-A.a:

Use the VCF output format to save the data. After calling the final variants, extract both SNPs and Indels, and save them as separate files.

gatk HaplotypeCaller -I SRR15117878_grpadded.bam -R hg38_chr20.fa -O SRR15117878_raw_variants.vcf

The GATK's HaplotypeCaller perform the first round of variant calling on pre-processed BAM file (SRR15117878_grpadded.bam) using the reference genome for chromosome 20 (hg38_chr20.fa). The tool scans through the aligned reads, compares them to the reference, and identifies sites where the sample differs from the reference (SNPs and indels), writing the results into a VCF file (SRR15117878_raw_variants.vcf). This step is crucial because it produces an initial set of variants: for human samples, these raw calls may be used directly for downstream analysis, while for organisms without established variant databases (like yeast), they serve as "known sites" for the Base Quality Score Recalibration (BQSR) step, improving the accuracy of subsequent variant discovery.

```
(base) ibab@IBAB-MScBDB2-Comp007:~/NGS/variant_calling/18aug$ gatk HaplotypeCaller -I SRR15117878_grpadded.bam -R hg38_chr20.fa -O  SRR15117878_raw_variants.vcf
Using GATK jar /home/ibab/miniconda3/share/gatk4-4.3.0.0-0/gatk-package-4.3.0.0-local.jar
Running:
    java -Dsamjdk.use_async_io_read_samtools=false -Dsamjdk.use_async_io_write_samtools=true -Dsamjdk.use_async_io_write_tribble=false -Dsamjdk.compression_level=2 -jar /home/ibab/miniconda3/share/gatk4-4.3.0.0-0/gatk-package-4.3.0.0-local.jar HaplotypeCaller -I SRR15117878_grpadded.bam -R hg38_chr20.fa -O SRR15117878_raw_variants.vcf
17:12:09.914 INFO  NativeLibraryLoader - Loading libgkl_compression.so from jar:file:/home/ibab/miniconda3/share/gatk4-4.3.0.0-0/gatk-package-4.3.0.0-local.jar!/com/intel/gkl/native/libgkl_compression.so
17:12:09.988 INFO  HaplotypeCaller - ------------------------------------------------------------
17:12:09.988 INFO  HaplotypeCaller - The Genome Analysis Toolkit (GATK) v4.3.0.0
17:12:09.988 INFO  HaplotypeCaller - For support and documentation go to https://software.broadinstitute.org/gatk/
17:12:09.988 INFO  HaplotypeCaller - Executing as ibab@IBAB-MScBDB2-Comp007.ibab.ac.in on Linux v6.14.0-28-generic amd64
17:12:09.988 INFO  HaplotypeCaller - Java runtime: OpenJDK 64-Bit Server VM v11.0.1+13-LTS
17:12:09.988 INFO  HaplotypeCaller - Start Date/Time: August 21, 2025 at 5:12:09 PM IST
17:12:09.988 INFO  HaplotypeCaller - ------------------------------------------------------------
17:12:09.988 INFO  HaplotypeCaller - ------------------------------------------------------------
17:12:09.989 INFO  HaplotypeCaller - HTSJDK Version: 3.0.1
17:12:09.989 INFO  HaplotypeCaller - Picard Version: 2.27.5
17:12:09.989 INFO  HaplotypeCaller - Built for Spark Version: 2.4.5
17:12:09.989 INFO  HaplotypeCaller - HTSJDK Defaults.COMPRESSION_LEVEL : 2
17:12:09.989 INFO  HaplotypeCaller - HTSJDK Defaults.USE_ASYNC_IO_READ_FOR_SAMTOOLS : false
17:12:09.989 INFO  HaplotypeCaller - HTSJDK Defaults.USE_ASYNC_IO_WRITE_FOR_SAMTOOLS : true
```

less -S SRR15117878_raw_variants.vcf

```
##fileformat=VCFv4.2
##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=PL,Number=G,Type=Integer,Description="Normalized, Phred-scaled likelihoods for genotypes as defined in the VCF specification">
##GATKCommandLine=<ID=HaplotypeCaller,CommandLine="HaplotypeCaller --output SRR15117878_raw_variants.vcf --input SRR15117878_grpadded.bam --reference>
##INFO=<ID=AC,Number=A,Type=Integer,Description="Allele count in genotypes, for each ALT allele, in the same order as listed">
##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">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth; some reads may have been filtered">
##INFO=<ID=ExcessHet,Number=1,Type=Float,Description="Phred-scaled p-value for exact test of excess heterozygosity">
##INFO=<ID=FS,Number=1,Type=Float,Description="Phred-scaled p-value using Fisher's exact test to detect strand bias">
##INFO=<ID=InbreedingCoeff,Number=1,Type=Float,Description="Inbreeding coefficient as estimated from the genotype likelihoods per-sample when compare>
##INFO=<ID=MLEAC,Number=A,Type=Integer,Description="Maximum likelihood expectation (MLE) for the allele counts (not necessarily the same as the AC), >
##INFO=<ID=MLEAF,Number=A,Type=Float,Description="Maximum likelihood expectation (MLE) for the allele frequency (not necessarily the same as the AF),>
##INFO=<ID=MQ,Number=1,Type=Float,Description="RMS Mapping Quality">
##INFO=<ID=MQRankSum,Number=1,Type=Float,Description="Z-score From Wilcoxon rank sum test of Alt vs. Ref read mapping qualities">
##INFO=<ID=QD,Number=1,Type=Float,Description="Variant Confidence/Quality by Depth">
##INFO=<ID=ReadPosRankSum,Number=1,Type=Float,Description="Z-score from Wilcoxon rank sum test of Alt vs. Ref read position bias">
##INFO=<ID=SOR,Number=1,Type=Float,Description="Symmetric Odds Ratio of 2x2 contingency table to detect strand bias">
##contig=<ID=chr20,length=64444167>
##source=HaplotypeCaller
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  sample_name
chr20   65453   .       C       T       78.32   .       AC=2;AF=1.00;AN=2;DP=2;ExcessHet=0.0000;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=25.00;QD=25.36;SOR=2.>
chr20   65455   .       C       T       78.32   .       AC=2;AF=1.00;AN=2;DP=2;ExcessHet=0.0000;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=25.00;QD=28.73;SOR=2.>
chr20   65460   .       C       A       78.32   .       AC=2;AF=1.00;AN=2;DP=2;ExcessHet=0.0000;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=25.00;QD=30.97;SOR=2.>
chr20   65482   .       T       G       78.32   .       AC=2;AF=1.00;AN=2;DP=2;ExcessHet=0.0000;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=25.00;QD=27.24;SOR=2.>
chr20   65486   .       T       A       78.32   .       AC=2;AF=1.00;AN=2;DP=2;ExcessHet=0.0000;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=25.00;QD=28.20;SOR=2.>
chr20   67012   .       A       T       35.48   .       AC=2;AF=1.00;AN=2;DP=1;ExcessHet=0.0000;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=43.00;QD=25.00;SOR=1.>
chr20   67020   .       C       T       35.48   .       AC=2;AF=1.00;AN=2;DP=1;ExcessHet=0.0000;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=43.00;QD=29.56;SOR=1.>
chr20   86236   .       C       T       78.32   .       AC=2;AF=1.00;AN=2;DP=2;ExcessHet=0.0000;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=25.08;QD=30.62;SOR=0.>
:
```

gatk BaseRecalibrator -I SRR15117878_grpadded.bam -R hg38_chr20.fa –known-sites SRR15117878_raw_variants.vcf -O recalibration.table

The GATK's BaseRecalibrator build a recalibration model of the base quality scores in the BAM file. It takes the aligned reads (SRR15117878_grpadded.bam) and the reference genome (hg38_chr20.fa), then compares the bases in the reads to a set of known variant sites (SRR15117878_raw_variants.vcf). By doing this, it identifies systematic errors made by the sequencer (for example, certain machines might consistently overestimate quality at the ends of reads or for specific sequence contexts). The result is a recalibration table of co-variation patterns to model base quality score errors, improving the accuracy and reliability of variant calls.

```
(base) ibab@IBAB-MScBDB2-Comp007:~/NGS/variant_calling/18aug$ gatk BaseRecalibrator -I SRR15117878_grpadded.bam -R hg38_chr20.fa --known-sites SRR1511
7878_raw_variants.vcf -O recalibration.table
Using GATK jar /home/ibab/miniconda3/share/gatk4-4.3.0.0-0/gatk-package-4.3.0.0-local.jar
Running:
    java -Dsamjdk.use_async_io_read_samtools=false -Dsamjdk.use_async_io_write_samtools=true -Dsamjdk.use_async_io_write_tribble=false -Dsamjdk.compre
ssion_level=2 -jar /home/ibab/miniconda3/share/gatk4-4.3.0.0-0/gatk-package-4.3.0.0-local.jar BaseRecalibrator -I SRR15117878_grpadded.bam -R hg38_chr
20.fa --known-sites SRR15117878_raw_variants.vcf -O recalibration.table
19:11:59.700 INFO  NativeLibraryLoader - Loading libgkl_compression.so from jar:file:/home/ibab/miniconda3/share/gatk4-4.3.0.0-0/gatk-package-4.3.0.0-
local.jar!/com/intel/gkl/native/libgkl_compression.so
19:11:59.765 INFO  BaseRecalibrator - ------------------------------------------------------------
19:11:59.765 INFO  BaseRecalibrator - The Genome Analysis Toolkit (GATK) v4.3.0.0
19:11:59.765 INFO  BaseRecalibrator - For support and documentation go to https://software.broadinstitute.org/gatk/
19:11:59.765 INFO  BaseRecalibrator - Executing as ibab@IBAB-MScBDB2-Comp007.ibab.ac.in on Linux v6.14.0-28-generic amd64
19:11:59.765 INFO  BaseRecalibrator - Java runtime: OpenJDK 64-Bit Server VM v11.0.1+13-LTS
19:11:59.765 INFO  BaseRecalibrator - Start Date/Time: August 21, 2025 at 7:11:59 PM IST
19:11:59.765 INFO  BaseRecalibrator - ------------------------------------------------------------
19:11:59.765 INFO  BaseRecalibrator - ------------------------------------------------------------
19:11:59.766 INFO  BaseRecalibrator - HTSJDK Version: 3.0.1
19:11:59.766 INFO  BaseRecalibrator - Picard Version: 2.27.5
19:11:59.766 INFO  BaseRecalibrator - Built for Spark Version: 2.4.5
19:11:59.766 INFO  BaseRecalibrator - HTSJDK Defaults.COMPRESSION_LEVEL : 2
19:11:59.766 INFO  BaseRecalibrator - HTSJDK Defaults.USE_ASYNC_IO_READ_FOR_SAMTOOLS : false
19:11:59.766 INFO  BaseRecalibrator - HTSJDK Defaults.USE_ASYNC_IO_WRITE_FOR_SAMTOOLS : true
```

less recalibration.table

```
#:GATKReport.v1.1:5
#:GATKTable:2:17:%s:%s:;
#:GATKTable:Arguments:Recalibration argument collection values used in this run
Argument                    Value
binary_tag_name             null
covariate                   ReadGroupCovariate,QualityScoreCovariate,ContextCovariate,CycleCovariate
default_platform            null
deletions_default_quality   45
force_platform              null
indels_context_size         3
insertions_default_quality  45
low_quality_tail            2
maximum_cycle_value         500
mismatches_context_size     2
mismatches_default_quality  -1
no_standard_covs            false
quantizing_levels           16
recalibration_report        null
run_without_dbsnp           false
solid_nocall_strategy       THROW_EXCEPTION
solid_recal_mode            SET_Q_ZERO

#:GATKTable:3:94:%d:%d:%d:;
#:GATKTable:Quantized:Quality quantization map
QualityScore  Count    QuantizedScore
           0        0              93
           1        0              93
           2        0              93
           3        0              93
           4        0              93
           5        0              93
           6        0              93
           7        0              93
           8        0              93
:
```

```
#:GATKTable:6:1:%s:%s:%.4f:%.4f:%d:%.2f:;
#:GATKTable:RecalTable0:
ReadGroup  EventType  EmpiricalQuality  EstimatedQReported  Observations  Errors
unit1      M                  14.0000             24.4080      11029098  429145.00

#:GATKTable:6:3:%s:%d:%s:%.4f:%d:%.2f:;
#:GATKTable:RecalTable1:
ReadGroup  QualityScore  EventType  EmpiricalQuality  Observations  Errors
unit1               11  M                  12.0000        445829   28754.00
unit1               25  M                  14.0000        825341   35191.00
unit1               37  M                  18.0000       9757928  365200.00

#:GATKTable:8:926:%s:%d:%s:%s:%s:%.4f:%d:%.2f:;
#:GATKTable:RecalTable2:
ReadGroup  QualityScore  CovariateValue  CovariateName  EventType  EmpiricalQuality  Observations  Errors
unit1               11  -1              Cycle          M                  18.0000          8250     12.00
unit1               11  -10             Cycle          M                  14.0000          7900    240.00
unit1               11  -100            Cycle          M                  10.0000           420     58.00
unit1               11  -101            Cycle          M                  10.0000           418     60.00
unit1               11  -102            Cycle          M                  10.0000           449     60.00
unit1               11  -103            Cycle          M                  11.0000           414     42.00
unit1               11  -104            Cycle          M                  10.0000           415     59.00
unit1               11  -105            Cycle          M                  10.0000           379     51.00
unit1               11  -106            Cycle          M                  10.0000           375     52.00
unit1               11  -107            Cycle          M                  10.0000           380     47.00
unit1               11  -108            Cycle          M                  10.0000           378     56.00
unit1               11  -109            Cycle          M                  10.0000           341     45.00
unit1               11  -11             Cycle          M                  14.0000          7733    251.00
unit1               11  -110            Cycle          M                  10.0000           332     49.00
unit1               11  -111            Cycle          M                  10.0000           313     42.00
unit1               11  -112            Cycle          M                  11.0000           329     35.00
unit1               11  -113            Cycle          M                  11.0000           316     35.00
unit1               11  -114            Cycle          M                  10.0000           306     40.00
unit1               11  -115            Cycle          M                  11.0000           280     25.00
:
```

gatk ApplyBQSR -I SRR15117878_grpadded.bam -R hg38_chr20.fa -bqsr recalibration.table -O SRR15117878_recalibrated_reads.bam

The GATK's ApplyBQSR, takes the recalibration table generated in the previous step and applies those corrections to the sequencing reads. The input is the aligned BAM file (SRR15117878_grpadded.bam), the reference genome (hg38_chr20.fa), and the recalibration table (recalibration.table). During this process, the original base quality scores assigned by the sequencer are adjusted according to the error patterns identified earlier—for example, systematic over- or under-estimation of quality at certain sequence contexts or read positions. The output is a new BAM file (SRR15117878_recalibrated_reads.bam) where the base quality scores now more accurately reflect the true probability of errors. This step is essential because many variant callers, including GATK's HaplotypeCaller, rely heavily on base quality scores to distinguish between true variants and sequencing errors. By applying BQSR, you improve the accuracy and confidence of downstream variant calling, leading to fewer false positives and false negatives.

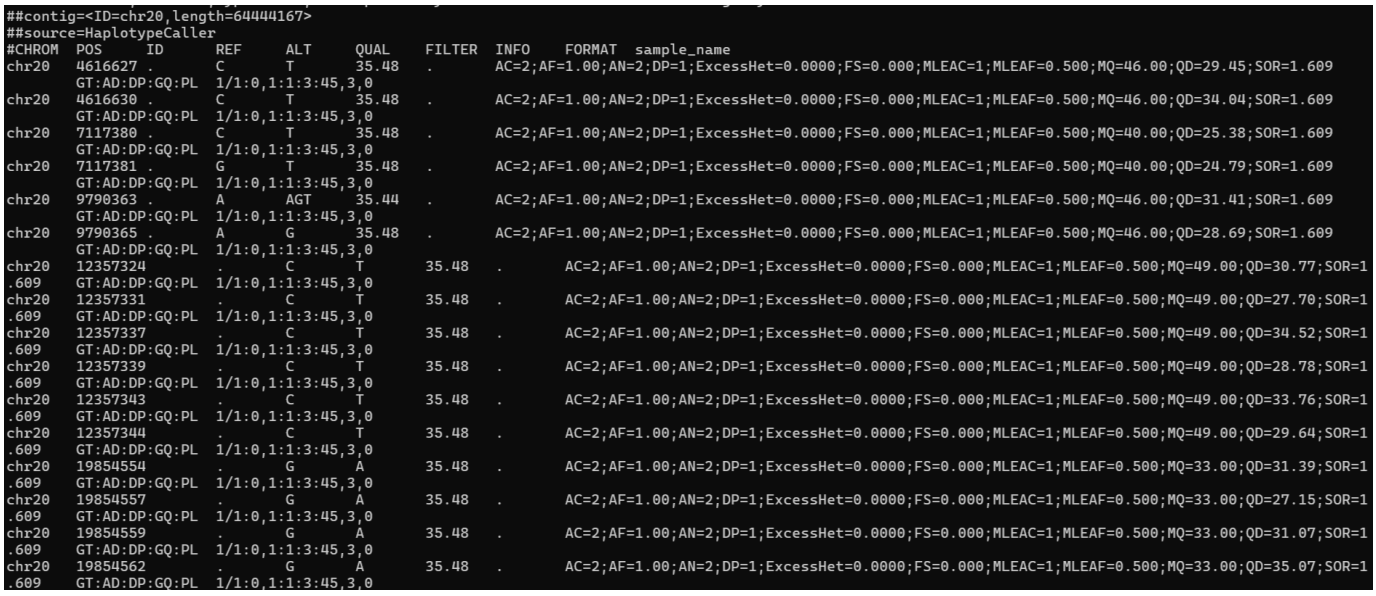
```
(base) ibab@IBAB-MScBDB2-Comp007:~/NGS/variant_calling/18aug$ gatk ApplyBQSR -I SRR15117878_grpadded.bam -R hg38_chr20.fa -bqsr recalibration.table -O
 SRR15117878_recalibrated_reads.bam
Using GATK jar /home/ibab/miniconda3/share/gatk4-4.3.0.0-0/gatk-package-4.3.0.0-local.jar
Running:
    java -Dsamjdk.use_async_io_read_samtools=false -Dsamjdk.use_async_io_write_samtools=true -Dsamjdk.use_async_io_write_tribble=false -Dsamjdk.compre
ssion_level=2 -jar /home/ibab/miniconda3/share/gatk4-4.3.0.0-0/gatk-package-4.3.0.0-local.jar ApplyBQSR -I SRR15117878_grpadded.bam -R hg38_chr20.fa -
bqsr recalibration.table -O SRR15117878_recalibrated_reads.bam
19:35:16.719 INFO  NativeLibraryLoader - Loading libgkl_compression.so from jar:file:/home/ibab/miniconda3/share/gatk4-4.3.0.0-0/gatk-package-4.3.0.0-
local.jar!/com/intel/gkl/native/libgkl_compression.so
19:35:16.783 INFO  ApplyBQSR - ------------------------------------------------------------
19:35:16.783 INFO  ApplyBQSR - The Genome Analysis Toolkit (GATK) v4.3.0.0
19:35:16.783 INFO  ApplyBQSR - For support and documentation go to https://software.broadinstitute.org/gatk/
19:35:16.783 INFO  ApplyBQSR - Executing as ibab@IBAB-MScBDB2-Comp007.ibab.ac.in on Linux v6.14.0-28-generic amd64
19:35:16.783 INFO  ApplyBQSR - Java runtime: OpenJDK 64-Bit Server VM v11.0.1+13-LTS
19:35:16.783 INFO  ApplyBQSR - Start Date/Time: August 21, 2025 at 7:35:16 PM IST
19:35:16.783 INFO  ApplyBQSR - ------------------------------------------------------------
19:35:16.783 INFO  ApplyBQSR - ------------------------------------------------------------
19:35:16.784 INFO  ApplyBQSR - HTSJDK Version: 3.0.1
19:35:16.784 INFO  ApplyBQSR - Picard Version: 2.27.5
19:35:16.784 INFO  ApplyBQSR - Built for Spark Version: 2.4.5
19:35:16.784 INFO  ApplyBQSR - HTSJDK Defaults.COMPRESSION_LEVEL : 2
```

samtools view SRR15117878_recalibrated_reads.bam

gatk HaplotypeCaller -R hg38_chr20.fa -I SRR15117878_recalibrated_reads.bam -O SRR15117878_final.vcf

This command runs GATK HaplotypeCaller to perform variant calling on the processed and recalibrated reads. Even though we have done an initial variant calling earlier (the “raw variants” step), that first run was only to generate a set of known sites for BQSR or for preliminary analysis. After BQSR, the base quality scores in BAM have been corrected, which improves the accuracy of variant detection. Therefore, we need to run HaplotypeCaller again on the recalibrated BAM (SRR15117878_recalibrated_reads.bam) against the reference genome (hg38_chr20.fa) to produce the final, high-confidence VCF (SRR15117878_final.vcf). This final VCF contains SNPs and INDELs with corrected base quality information, making it suitable for downstream analysis, filtering, or splitting into SNPs and INDELs. Essentially, the second HaplotypeCaller run is the one that gives the definitive set of variants.

Less SRR15117878_final.vcf



gatk CNNScoreVariants -R hg38_chr20.fa -V SRR15117878_final.vcf –tensor-type READ_TENSOR -O SRR15117878_CNN_filtered.vcf

or

gatk CNNScoreVariants -R hg38_chr20.fa -V SRR15117878_final.vcf -I SRR15117878_recalibrated_reads.bam –tensor-type READ_TENSOR -O SRR15117878_CNN_filtered.vcf

CNNScore is a GATK parameter/tool that uses a Convolutional Neural Network (CNN) to filter out false positive variant calls from a raw variant list. The key difference is that the first command uses the -I flag to explicitly provide the read alignments from the BAM file, while the second command does not. This allows the first command to generate a more accurate CNN score by analyzing the detailed read-level data, resulting in more reliable variant filtering in the output file.

```
------------------------------------------------------------------------------------

**********************************************************************

A USER ERROR has occurred: CNNScoreVariants is no longer included in GATK as of version 4.6.1.0. Please use the replacement tool NVScoreVariants instead, wh
ich produces virtually identical results

**********************************************************************
Set the system property GATK_STACKTRACE_ON_USER_EXCEPTION (--java-options '-DGATK_STACKTRACE_ON_USER_EXCEPTION=true') to print the stack trace.
```

```
ibab@LAPTOP-BVSTVK8Q:~/NGS/18aug_lab$ /home/ibab/NGS/Packages/gatk-4.6.2.0/gatk NVScoreVariants -R hg38_chr20.fa -V SRR15117878_final.vcf -O SRR15117878_CNN
_filtered.vcf
Using GATK jar /home/ibab/NGS/Packages/gatk-4.6.2.0/gatk-package-4.6.2.0-local.jar
Running:
    java -Dsamjdk.use_async_io_read_samtools=false -Dsamjdk.use_async_io_write_samtools=true -Dsamjdk.use_async_io_write_tribble=false -Dsamjdk.compression_
level=2 -jar /home/ibab/NGS/Packages/gatk-4.6.2.0/gatk-package-4.6.2.0-local.jar NVScoreVariants -R hg38_chr20.fa -V SRR15117878_final.vcf -O SRR15117878_CN
N_filtered.vcf
05:51:22.318 INFO  NativeLibraryLoader - Loading libgkl_compression.so from jar:file:/home/ibab/NGS/Packages/gatk-4.6.2.0/gatk-package-4.6.2.0-local.jar!/co
m/intel/gkl/native/libgkl_compression.so
05:51:22.742 INFO  NVScoreVariants - ------------------------------------------------------------
05:51:22.748 INFO  NVScoreVariants - The Genome Analysis Toolkit (GATK) v4.6.2.0
05:51:22.748 INFO  NVScoreVariants - For support and documentation go to https://software.broadinstitute.org/gatk/
05:51:22.749 INFO  NVScoreVariants - Executing as ibab@LAPTOP-BVSTVK8Q on Linux v6.6.87.2-microsoft-standard-WSL2 amd64
05:51:22.749 INFO  NVScoreVariants - Java runtime: OpenJDK 64-Bit Server VM v17.0.16+8-Ubuntu-0ubuntu122.04.1
05:51:22.749 INFO  NVScoreVariants - Start Date/Time: August 22, 2025 at 5:51:22 AM IST
05:51:22.750 INFO  NVScoreVariants - ------------------------------------------------------------
05:51:22.750 INFO  NVScoreVariants - ------------------------------------------------------------
05:51:22.751 INFO  NVScoreVariants - HTSJDK Version: 4.2.0
05:51:22.751 INFO  NVScoreVariants - Picard Version: 3.4.0
05:51:22.751 INFO  NVScoreVariants - Built for Spark Version: 3.5.0
05:51:22.759 INFO  NVScoreVariants - HTSJDK Defaults.COMPRESSION_LEVEL : 2
05:51:22.761 INFO  NVScoreVariants - HTSJDK Defaults.USE_ASYNC_IO_READ_FOR_SAMTOOLS : false
05:51:22.762 INFO  NVScoreVariants - HTSJDK Defaults.USE_ASYNC_IO_WRITE_FOR_SAMTOOLS : true
05:51:22.763 INFO  NVScoreVariants - HTSJDK Defaults.USE_ASYNC_IO_WRITE_FOR_TRIBBLE : false
05:51:22.763 INFO  NVScoreVariants - Deflater: IntelDeflater
05:51:22.763 INFO  NVScoreVariants - Inflater: IntelInflater
05:51:22.764 INFO  NVScoreVariants - GCS max retries/reopens: 20
05:51:22.764 INFO  NVScoreVariants - Requester pays: disabled
05:51:22.764 WARN  NVScoreVariants -

   !!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!

   Warning: NVScoreVariants is a BETA tool and is not yet ready for use in production

   !!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
```

gatk SelectVariants -R hg38_chr20.fa -V SRR15117878_final.vcf –select-type-to-include SNP -O SRR15117878_snps.vcf

```
ibab@LAPTOP-BVSTVK8Q:~/NGS/18aug_lab$ /home/ibab/NGS/Packages/gatk-4.6.2.0/gatk SelectVariants -R hg38_chr20.fa -V SRR15117878_final.vcf --select-type-to-in
clude SNP -O SRR15117878_snps.vcf
Using GATK jar /home/ibab/NGS/Packages/gatk-4.6.2.0/gatk-package-4.6.2.0-local.jar
Running:
    java -Dsamjdk.use_async_io_read_samtools=false -Dsamjdk.use_async_io_write_samtools=true -Dsamjdk.use_async_io_write_tribble=false -Dsamjdk.compression_
level=2 -jar /home/ibab/NGS/Packages/gatk-4.6.2.0/gatk-package-4.6.2.0-local.jar SelectVariants -R hg38_chr20.fa -V SRR15117878_final.vcf --select-type-to-i
nclude SNP -O SRR15117878_snps.vcf
06:18:29.419 INFO  NativeLibraryLoader - Loading libgkl_compression.so from jar:file:/home/ibab/NGS/Packages/gatk-4.6.2.0/gatk-package-4.6.2.0-local.jar!/co
m/intel/gkl/native/libgkl_compression.so
06:18:30.411 INFO  SelectVariants - ------------------------------------------------------------
06:18:30.419 INFO  SelectVariants - The Genome Analysis Toolkit (GATK) v4.6.2.0
06:18:30.420 INFO  SelectVariants - For support and documentation go to https://software.broadinstitute.org/gatk/
06:18:30.421 INFO  SelectVariants - Executing as ibab@LAPTOP-BVSTVK8Q on Linux v6.6.87.2-microsoft-standard-WSL2 amd64
06:18:30.421 INFO  SelectVariants - Java runtime: OpenJDK 64-Bit Server VM v17.0.16+8-Ubuntu-0ubuntu122.04.1
06:18:30.422 INFO  SelectVariants - Start Date/Time: August 22, 2025 at 6:18:29 AM IST
06:18:30.422 INFO  SelectVariants - ------------------------------------------------------------
06:18:30.423 INFO  SelectVariants - ------------------------------------------------------------
06:18:30.424 INFO  SelectVariants - HTSJDK Version: 4.2.0
06:18:30.424 INFO  SelectVariants - Picard Version: 3.4.0
06:18:30.424 INFO  SelectVariants - Built for Spark Version: 3.5.0
06:18:30.427 INFO  SelectVariants - HTSJDK Defaults.COMPRESSION_LEVEL : 2
06:18:30.431 INFO  SelectVariants - HTSJDK Defaults.USE_ASYNC_IO_READ_FOR_SAMTOOLS : false
```

```
##source=SelectVariants
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  sample_name
chr20   4616627 .       C       T       35.48   .       AC=2;AF=1.00;AN=2;DP=1;ExcessHet=0.0000;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=46.00;QD=29.45;SOR=1.609
        GT:AD:DP:GQ:PL  1/1:0,1:1:3:45,3,0
chr20   4616630 .       C       T       35.48   .       AC=2;AF=1.00;AN=2;DP=1;ExcessHet=0.0000;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=46.00;QD=34.04;SOR=1.609
        GT:AD:DP:GQ:PL  1/1:0,1:1:3:45,3,0
chr20   7117380 .       C       T       35.48   .       AC=2;AF=1.00;AN=2;DP=1;ExcessHet=0.0000;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=40.00;QD=25.38;SOR=1.609
        GT:AD:DP:GQ:PL  1/1:0,1:1:3:45,3,0
chr20   7117381 .       G       T       35.48   .       AC=2;AF=1.00;AN=2;DP=1;ExcessHet=0.0000;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=40.00;QD=24.79;SOR=1.609
        GT:AD:DP:GQ:PL  1/1:0,1:1:3:45,3,0
chr20   9790365 .       A       G       35.48   .       AC=2;AF=1.00;AN=2;DP=1;ExcessHet=0.0000;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=46.00;QD=28.69;SOR=1.609
        GT:AD:DP:GQ:PL  1/1:0,1:1:3:45,3,0
chr20   12357324        .       C       T       35.48   .       AC=2;AF=1.00;AN=2;DP=1;ExcessHet=0.0000;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=49.00;QD=30.77;SOR=1
.609    GT:AD:DP:GQ:PL  1/1:0,1:1:3:45,3,0
chr20   12357331        .       C       T       35.48   .       AC=2;AF=1.00;AN=2;DP=1;ExcessHet=0.0000;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=49.00;QD=27.70;SOR=1
.609    GT:AD:DP:GQ:PL  1/1:0,1:1:3:45,3,0
chr20   12357337        .       C       T       35.48   .       AC=2;AF=1.00;AN=2;DP=1;ExcessHet=0.0000;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=49.00;QD=34.52;SOR=1
.609    GT:AD:DP:GQ:PL  1/1:0,1:1:3:45,3,0
chr20   12357339        .       C       T       35.48   .       AC=2;AF=1.00;AN=2;DP=1;ExcessHet=0.0000;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=49.00;QD=28.78;SOR=1
.609    GT:AD:DP:GQ:PL  1/1:0,1:1:3:45,3,0
chr20   12357343        .       C       T       35.48   .       AC=2;AF=1.00;AN=2;DP=1;ExcessHet=0.0000;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=49.00;QD=33.76;SOR=1
.609    GT:AD:DP:GQ:PL  1/1:0,1:1:3:45,3,0
chr20   12357344        .       C       T       35.48   .       AC=2;AF=1.00;AN=2;DP=1;ExcessHet=0.0000;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=49.00;QD=29.64;SOR=1
.609    GT:AD:DP:GQ:PL  1/1:0,1:1:3:45,3,0
chr20   19854554        .       G       A       35.48   .       AC=2;AF=1.00;AN=2;DP=1;ExcessHet=0.0000;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=33.00;QD=31.39;SOR=1
.609    GT:AD:DP:GQ:PL  1/1:0,1:1:3:45,3,0
chr20   19854557        .       G       A       35.48   .       AC=2;AF=1.00;AN=2;DP=1;ExcessHet=0.0000;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=33.00;QD=27.15;SOR=1
.609    GT:AD:DP:GQ:PL  1/1:0,1:1:3:45,3,0
chr20   19854559        .       G       A       35.48   .       AC=2;AF=1.00;AN=2;DP=1;ExcessHet=0.0000;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=33.00;QD=31.07;SOR=1
.609    GT:AD:DP:GQ:PL  1/1:0,1:1:3:45,3,0
chr20   19854562        .       G       A       35.48   .       AC=2;AF=1.00;AN=2;DP=1;ExcessHet=0.0000;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=33.00;QD=35.07;SOR=1
.609    GT:AD:DP:GQ:PL  1/1:0,1:1:3:45,3,0
```

gatk SelectVariants -R hg38_chr20.fa -V SRR15117878_final.vcf –select-type-to-include INDEL -O SRR15117878_indels.vcf

```
ibab@LAPTOP-BVSTVK8Q:~/NGS/18aug_lab$ /home/ibab/NGS/Packages/gatk-4.6.2.0/gatk SelectVariants -R hg38_chr20.fa -V SRR15117878_final.vcf --select-type-to-in
clude INDEL -O SRR15117878_indels.vcf
Using GATK jar /home/ibab/NGS/Packages/gatk-4.6.2.0/gatk-package-4.6.2.0-local.jar
Running:
    java -Dsamjdk.use_async_io_read_samtools=false -Dsamjdk.use_async_io_write_samtools=true -Dsamjdk.use_async_io_write_tribble=false -Dsamjdk.compression_
level=2 -jar /home/ibab/NGS/Packages/gatk-4.6.2.0/gatk-package-4.6.2.0-local.jar SelectVariants -R hg38_chr20.fa -V SRR15117878_final.vcf --select-type-to-i
nclude INDEL -O SRR15117878_indels.vcf
06:23:45.638 INFO  NativeLibraryLoader - Loading libgkl_compression.so from jar:file:/home/ibab/NGS/Packages/gatk-4.6.2.0/gatk-package-4.6.2.0-local.jar!/co
m/intel/gkl/native/libgkl_compression.so
06:23:46.106 INFO  SelectVariants - ------------------------------------------------------------
06:23:46.112 INFO  SelectVariants - The Genome Analysis Toolkit (GATK) v4.6.2.0
06:23:46.112 INFO  SelectVariants - For support and documentation go to https://software.broadinstitute.org/gatk/
06:23:46.113 INFO  SelectVariants - Executing as ibab@LAPTOP-BVSTVK8Q on Linux v6.6.87.2-microsoft-standard-WSL2 amd64
06:23:46.114 INFO  SelectVariants - Java runtime: OpenJDK 64-Bit Server VM v17.0.16+8-Ubuntu-0ubuntu122.04.1
06:23:46.114 INFO  SelectVariants - Start Date/Time: August 22, 2025 at 6:23:45 AM IST
06:23:46.114 INFO  SelectVariants - ------------------------------------------------------------
06:23:46.115 INFO  SelectVariants - ------------------------------------------------------------
06:23:46.116 INFO  SelectVariants - HTSJDK Version: 4.2.0
```

```
##source=HaplotypeCaller
##source=SelectVariants
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  sample_name
chr20   9790363 .       A       AGT     35.44   .       AC=2;AF=1.00;AN=2;DP=1;ExcessHet=0.0000;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=46.00;QD=31.41;SOR=1.609
        GT:AD:DP:GQ:PL  1/1:0,1:1:3:45,3,0
(END)
```

Part-A.b:

Use the Gvcf output format to save the data. After calling the final variants, extract both the SNPs and Indels, and save them as separate files.

Follow the same steps (initial call for known-sites if needed, BQSR, ApplyBQSR). The difference is: HaplotypeCaller in Gvcf mode (-ERC GVCF), then genotype the Gvcf to VCF, then filter and split.

gatk HaplotypeCaller -R hg38_chr20.fa -I SRR15117878_recalibrated_reads.bam -ERC GVCF -O SRR15117878_gvcf_final.g.vcf

It executes GATK HaplotypeCaller on the input BAM file (SRR15117878_recalibrated_reads.bam), which contains sequencing reads aligned to the reference genome hg38_chr20.fa and recalibrated for base quality. The -ERC GVCF option stands for “Emit Reference Confidence” in GVCF mode.

By using this mode, HaplotypeCaller does not just report variant sites but also captures information about homozygous reference sites (positions where the sample matches the reference) along with the confidence of that observation. This produces a Genomic VCF (GVCF), which differs from a standard VCF in that it contains a complete record of the genome, including both variant and non-variant regions, organized in blocks to reduce file size. GVCFs are particularly useful for joint genotyping across multiple samples, because they allow combining information from several GVCFs to accurately determine variants in a cohort without losing data from positions that are reference in some samples.

less -S SRR15117878_gvcf_final.g.vcf

The -S option in ‘less’ tells it to chop long lines rather than wrapping them, so each line stays on a single row and we can scroll horizontally to see content beyond the screen width.

```
ibab@LAPTOP-BVSTVK8Q:~/NGS/18aug_lab$ /home/ibab/NGS/Packages/gatk-4.6.2.0/gatk HaplotypeCaller -R hg38_chr20.fa -I SRR15117878_recalibrated_reads.bam -ERC
GVCF -O SRR15117878_gvcf_final.g.vcf
Using GATK jar /home/ibab/NGS/Packages/gatk-4.6.2.0/gatk-package-4.6.2.0-local.jar
Running:
    java -Dsamjdk.use_async_io_read_samtools=false -Dsamjdk.use_async_io_write_samtools=true -Dsamjdk.use_async_io_write_tribble=false -Dsamjdk.compression_
level=2 -jar /home/ibab/NGS/Packages/gatk-4.6.2.0/gatk-package-4.6.2.0-local.jar HaplotypeCaller -R hg38_chr20.fa -I SRR15117878_recalibrated_reads.bam -ERC
 GVCF -O SRR15117878_gvcf_final.g.vcf
06:35:49.158 INFO  NativeLibraryLoader - Loading libgkl_compression.so from jar:file:/home/ibab/NGS/Packages/gatk-4.6.2.0/gatk-package-4.6.2.0-local.jar!/co
m/intel/gkl/native/libgkl_compression.so
06:35:49.604 INFO  HaplotypeCaller - ------------------------------------------------------------
06:35:49.609 INFO  HaplotypeCaller - The Genome Analysis Toolkit (GATK) v4.6.2.0
06:35:49.610 INFO  HaplotypeCaller - For support and documentation go to https://software.broadinstitute.org/gatk/
06:35:49.610 INFO  HaplotypeCaller - Executing as ibab@LAPTOP-BVSTVK8Q on Linux v6.6.87.2-microsoft-standard-WSL2 amd64
06:35:49.610 INFO  HaplotypeCaller - Java runtime: OpenJDK 64-Bit Server VM v17.0.16+8-Ubuntu-0ubuntu122.04.1
06:35:49.610 INFO  HaplotypeCaller - Start Date/Time: August 22, 2025 at 6:35:49 AM IST
06:35:49.610 INFO  HaplotypeCaller - ------------------------------------------------------------
06:35:49.610 INFO  HaplotypeCaller - ------------------------------------------------------------
06:35:49.611 INFO  HaplotypeCaller - HTSJDK Version: 4.2.0
06:35:49.612 INFO  HaplotypeCaller - Picard Version: 3.4.0
06:35:49.612 INFO  HaplotypeCaller - Built for Spark Version: 3.5.0
06:35:49.614 INFO  HaplotypeCaller - HTSJDK Defaults.COMPRESSION_LEVEL : 2
06:35:49.615 INFO  HaplotypeCaller - HTSJDK Defaults.USE_ASYNC_IO_READ_FOR_SAMTOOLS : false
06:35:49.615 INFO  HaplotypeCaller - HTSJDK Defaults.USE_ASYNC_IO_WRITE_FOR_SAMTOOLS : true
```

```
##fileformat=VCFv4.2
##ALT=<ID=NON_REF,Description="Represents any possible alternative allele not already represented at this location by REF and ALT">
##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 o>
##FORMAT=<ID=PID,Number=1,Type=String,Description="Physical phasing ID information, where each unique ID within a given sample (but not across samples) con>
##FORMAT=<ID=PL,Number=G,Type=Integer,Description="Normalized, Phred-scaled likelihoods for genotypes as defined in the VCF specification">
##FORMAT=<ID=PS,Number=1,Type=Integer,Description="Phasing set (typically the position of the first variant in the set)">
##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=HaplotypeCaller,CommandLine="HaplotypeCaller --emit-ref-confidence GVCF --output SRR15117878_gvcf_final.g.vcf --input SRR15117878_rec>
##GVCFBlock0-1=minGQ=0(inclusive),maxGQ=1(exclusive)
##GVCFBlock1-2=minGQ=1(inclusive),maxGQ=2(exclusive)
##GVCFBlock10-11=minGQ=10(inclusive),maxGQ=11(exclusive)
##GVCFBlock11-12=minGQ=11(inclusive),maxGQ=12(exclusive)
##GVCFBlock12-13=minGQ=12(inclusive),maxGQ=13(exclusive)
##GVCFBlock13-14=minGQ=13(inclusive),maxGQ=14(exclusive)
##GVCFBlock14-15=minGQ=14(inclusive),maxGQ=15(exclusive)
##GVCFBlock15-16=minGQ=15(inclusive),maxGQ=16(exclusive)
##GVCFBlock16-17=minGQ=16(inclusive),maxGQ=17(exclusive)
```

```
##INFO=<ID=BaseQRankSum,Number=1,Type=Float,Description="Z-score from Wilcoxon rank sum test of Alt Vs. Ref base qualities">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth; some reads may have been filtered">
##INFO=<ID=END,Number=1,Type=Integer,Description="Stop position of the interval">
##INFO=<ID=ExcessHet,Number=1,Type=Float,Description="Phred-scaled p-value for exact test of excess heterozygosity">
##INFO=<ID=InbreedingCoeff,Number=1,Type=Float,Description="Inbreeding coefficient as estimated from the genotype likelihoods per-sample when compared agai>
##INFO=<ID=MLEAC,Number=A,Type=Integer,Description="Maximum likelihood expectation (MLE) for the allele counts (not necessarily the same as the AC), for ea>
##INFO=<ID=MLEAF,Number=A,Type=Float,Description="Maximum likelihood expectation (MLE) for the allele frequency (not necessarily the same as the AF), for e>
##INFO=<ID=MQRankSum,Number=1,Type=Float,Description="Z-score From Wilcoxon rank sum test of Alt vs. Ref read mapping qualities">
##INFO=<ID=RAW_MQandDP,Number=2,Type=Integer,Description="Raw data (sum of squared MQ and total depth) for improved RMS Mapping Quality calculation. Incomp>
##INFO=<ID=ReadPosRankSum,Number=1,Type=Float,Description="Z-score from Wilcoxon rank sum test of Alt vs. Ref read position bias">
##contig=<ID=chr20,length=64444167>
##source=HaplotypeCaller
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  sample_name
chr20   1       .       N       <NON_REF>       .       .       END=60516       GT:DP:GQ:MIN_DP:PL      0/0:0:0:0:0,0,0
chr20   60517   .       T       <NON_REF>       .       .       END=60525       GT:DP:GQ:MIN_DP:PL      0/0:1:3:1:0,3,15
chr20   60526   .       C       <NON_REF>       .       .       END=60526       GT:DP:GQ:MIN_DP:PL      0/0:1:0:1:0,0,0
chr20   60527   .       T       <NON_REF>       .       .       END=60546       GT:DP:GQ:MIN_DP:PL      0/0:1:3:1:0,3,15
chr20   60547   .       T       <NON_REF>       .       .       END=60720       GT:DP:GQ:MIN_DP:PL      0/0:0:0:0:0,0,0
chr20   60721   .       T       <NON_REF>       .       .       END=60739       GT:DP:GQ:MIN_DP:PL      0/0:1:3:1:0,3,11
chr20   60740   .       G       <NON_REF>       .       .       END=60740       GT:DP:GQ:MIN_DP:PL      0/0:1:0:1:0,0,0
chr20   60741   .       T       <NON_REF>       .       .       END=60749       GT:DP:GQ:MIN_DP:PL      0/0:1:3:1:0,3,12
chr20   60750   .       T       <NON_REF>       .       .       END=60750       GT:DP:GQ:MIN_DP:PL      0/0:1:0:1:0,0,0
chr20   60751   .       T       <NON_REF>       .       .       END=60753       GT:DP:GQ:MIN_DP:PL      0/0:1:3:1:0,3,15
:
```

gatk GenotypeGVCFs -R hg38_chr20.fa -V SRR15117878_gvcf_final.g.vcf -O SRR15117878_final_from_gvcf.vcf

This runs GATK GenotypeGVCFs, which takes a GVCF file (SRR15117878_gvcf_final.g.vcf) and produces a standard VCF file (SRR15117878_final_from_gvcf.vcf) containing called variants with genotypes. The purpose is that a GVCF (produced by HaplotypeCaller with -ERC GVCF) includes information about both variant and non-variant sites, but it does not assign final genotypes. GenotypeGVCFs evaluates the confidence scores across the genome (or across multiple samples if combined) and outputs a standard VCF with finalized genotype calls, suitable for downstream analysis like filtering or annotation.

- If we have only one sample, we can genotype directly from its GVCF.
- If we have multiple samples, we first merge them using CombineGVCFs or GenomicsDBImport to create a joint dataset, then run GenotypeGVCFs to call variants across all samples together, ensuring consistency and capturing variants that might be present at low frequency in some samples.

```
ibab@LAPTOP-BVSTVK8Q:~/NGS/18aug_lab$ /home/ibab/NGS/Packages/gatk-4.6.2.0/gatk GenotypeGVCFs -R hg38_chr20.fa -V SRR15117878_gvcf_final.g.vcf -O SRR1511787
8_final_from_gvcf.vcf
Using GATK jar /home/ibab/NGS/Packages/gatk-4.6.2.0/gatk-package-4.6.2.0-local.jar
Running:
    java -Dsamjdk.use_async_io_read_samtools=false -Dsamjdk.use_async_io_write_samtools=true -Dsamjdk.use_async_io_write_tribble=false -Dsamjdk.compression_
level=2 -jar /home/ibab/NGS/Packages/gatk-4.6.2.0/gatk-package-4.6.2.0-local.jar GenotypeGVCFs -R hg38_chr20.fa -V SRR15117878_gvcf_final.g.vcf -O SRR151178
78_final_from_gvcf.vcf
06:50:41.054 INFO  NativeLibraryLoader - Loading libgkl_compression.so from jar:file:/home/ibab/NGS/Packages/gatk-4.6.2.0/gatk-package-4.6.2.0-local.jar!/co
m/intel/gkl/native/libgkl_compression.so
06:50:41.514 INFO  GenotypeGVCFs - ------------------------------------------------------------
06:50:41.519 INFO  GenotypeGVCFs - The Genome Analysis Toolkit (GATK) v4.6.2.0
06:50:41.519 INFO  GenotypeGVCFs - For support and documentation go to https://software.broadinstitute.org/gatk/
06:50:41.519 INFO  GenotypeGVCFs - Executing as ibab@LAPTOP-BVSTVK8Q on Linux v6.6.87.2-microsoft-standard-WSL2 amd64
06:50:41.519 INFO  GenotypeGVCFs - Java runtime: OpenJDK 64-Bit Server VM v17.0.16+8-Ubuntu-0ubuntu122.04.1
06:50:41.519 INFO  GenotypeGVCFs - Start Date/Time: August 22, 2025 at 6:50:40 AM IST
06:50:41.519 INFO  GenotypeGVCFs - ------------------------------------------------------------
06:50:41.520 INFO  GenotypeGVCFs - ------------------------------------------------------------
06:50:41.520 INFO  GenotypeGVCFs - HTSJDK Version: 4.2.0
06:50:41.521 INFO  GenotypeGVCFs - Picard Version: 3.4.0
06:50:41.521 INFO  GenotypeGVCFs - Built for Spark Version: 3.5.0
06:50:41.523 INFO  GenotypeGVCFs - HTSJDK Defaults.COMPRESSION_LEVEL : 2
06:50:41.524 INFO  GenotypeGVCFs - HTSJDK Defaults.USE_ASYNC_IO_READ_FOR_SAMTOOLS : false
```

less -S SRR15117878_final_from_gvcf.vcf

```
##fileformat=VCFv4.2
##ALT=<ID=NON_REF,Description="Represents any possible alternative allele not already represented at this location by REF and ALT">
##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 o>
##FORMAT=<ID=PID,Number=1,Type=String,Description="Physical phasing ID information, where each unique ID within a given sample (but not across samples) con>
##FORMAT=<ID=PL,Number=G,Type=Integer,Description="Normalized, Phred-scaled likelihoods for genotypes as defined in the VCF specification">
##FORMAT=<ID=PS,Number=1,Type=Integer,Description="Phasing set (typically the position of the first variant in the set)">
##FORMAT=<ID=RGQ,Number=1,Type=Integer,Description="Unconditional reference genotype confidence, encoded as a phred quality -10*log10 p(genotype call is wr>
##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=GenotypeGVCFs,CommandLine="GenotypeGVCFs --output SRR15117878_final_from_gvcf.vcf --variant SRR15117878_gvcf_final.g.vcf --reference >
##GATKCommandLine=<ID=HaplotypeCaller,CommandLine="HaplotypeCaller --emit-ref-confidence GVCF --output SRR15117878_gvcf_final.g.vcf --input SRR15117878_rec>
##INFO=<ID=AC,Number=A,Type=Integer,Description="Allele count in genotypes, for each ALT allele, in the same order as listed">
##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">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth; some reads may have been filtered">
##INFO=<ID=END,Number=1,Type=Integer,Description="Stop position of the interval">
##INFO=<ID=ExcessHet,Number=1,Type=Float,Description="Phred-scaled p-value for exact test of excess heterozygosity">
##INFO=<ID=FS,Number=1,Type=Float,Description="Phred-scaled p-value using Fisher's exact test to detect strand bias">
##INFO=<ID=InbreedingCoeff,Number=1,Type=Float,Description="Inbreeding coefficient as estimated from the genotype likelihoods per-sample when compared agai>
##INFO=<ID=MLEAC,Number=A,Type=Integer,Description="Maximum likelihood expectation (MLE) for the allele counts (not necessarily the same as the AC), for ea>
##INFO=<ID=MLEAF,Number=A,Type=Float,Description="Maximum likelihood expectation (MLE) for the allele frequency (not necessarily the same as the AF), for e>
##INFO=<ID=MQ,Number=1,Type=Float,Description="RMS Mapping Quality">
##INFO=<ID=MQRankSum,Number=1,Type=Float,Description="Z-score From Wilcoxon rank sum test of Alt vs. Ref read mapping qualities">
##INFO=<ID=QD,Number=1,Type=Float,Description="Variant Confidence/Quality by Depth">
##INFO=<ID=RAW_MQandDP,Number=2,Type=Integer,Description="Raw data (sum of squared MQ and total depth) for improved RMS Mapping Quality calculation. Incomp>
##INFO=<ID=ReadPosRankSum,Number=1,Type=Float,Description="Z-score from Wilcoxon rank sum test of Alt vs. Ref read position bias">
##INFO=<ID=SOR,Number=1,Type=Float,Description="Symmetric Odds Ratio of 2x2 contingency table to detect strand bias">
##contig=<ID=chr20,length=64444167>
##source=GenotypeGVCFs
##source=HaplotypeCaller
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  sample_name
chr20   4616627 .       C       T       35.48   .       AC=2;AF=1.00;AN=2;DP=1;ExcessHet=0.0000;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=46.00;QD=25.36;SOR=1.609   >
chr20   4616630 .       C       T       35.48   .       AC=2;AF=1.00;AN=2;DP=1;ExcessHet=0.0000;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=46.00;QD=28.73;SOR=1.609   >
chr20   7117380 .       C       T       35.48   .       AC=2;AF=1.00;AN=2;DP=1;ExcessHet=0.0000;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=40.00;QD=30.97;SOR=1.609   >
```

gatk SelectVariants -R hg38_chr20.fa -V SRR15117878_final_from_gvcf.vcf –select-type-to-include SNP -O SRR15117878_snps_from_gvcf.vcf

```
ibab@LAPTOP-BVSTVK8Q:~/NGS/18aug_lab$ /home/ibab/NGS/Packages/gatk-4.6.2.0/gatk SelectVariants -R hg38_chr20.fa -V SRR15117878_final_from_gvcf.vcf --select-type-to-include SNP -O SRR15117878_snps_from_gvcf.vcf
Using GATK jar /home/ibab/NGS/Packages/gatk-4.6.2.0/gatk-package-4.6.2.0-local.jar
Running:
    java -Dsamjdk.use_async_io_read_samtools=false -Dsamjdk.use_async_io_write_samtools=true -Dsamjdk.use_async_io_write_tribble=false -Dsamjdk.compression_level=2 -jar /home/ibab/NGS/Packages/gatk-4.6.2.0/gatk-package-4.6.2.0-local.jar SelectVariants -R hg38_chr20.fa -V SRR15117878_final_from_gvcf.vcf --select-type-to-include SNP -O SRR15117878_snps_from_gvcf.vcf
07:02:11.798 INFO  NativeLibraryLoader - Loading libgkl_compression.so from jar:file:/home/ibab/NGS/Packages/gatk-4.6.2.0/gatk-package-4.6.2.0-local.jar!/com/intel/gkl/native/libgkl_compression.so
07:02:12.270 INFO  SelectVariants - ------------------------------------------------------------
07:02:12.277 INFO  SelectVariants - The Genome Analysis Toolkit (GATK) v4.6.2.0
07:02:12.277 INFO  SelectVariants - For support and documentation go to https://software.broadinstitute.org/gatk/
07:02:12.277 INFO  SelectVariants - Executing as ibab@LAPTOP-BVSTVK8Q on Linux v6.6.87.2-microsoft-standard-WSL2 amd64
07:02:12.278 INFO  SelectVariants - Java runtime: OpenJDK 64-Bit Server VM v17.0.16+8-Ubuntu-0ubuntu122.04.1
07:02:12.278 INFO  SelectVariants - Start Date/Time: August 22, 2025 at 7:02:11 AM IST
07:02:12.278 INFO  SelectVariants - ------------------------------------------------------------
07:02:12.278 INFO  SelectVariants - ------------------------------------------------------------
07:02:12.280 INFO  SelectVariants - HTSJDK Version: 4.2.0
07:02:12.280 INFO  SelectVariants - Picard Version: 3.4.0
07:02:12.280 INFO  SelectVariants - Built for Spark Version: 3.5.0
07:02:12.283 INFO  SelectVariants - HTSJDK Defaults.COMPRESSION_LEVEL : 2
07:02:12.284 INFO  SelectVariants - HTSJDK Defaults.USE_ASYNC_IO_READ_FOR_SAMTOOLS : false
07:02:12.284 INFO  SelectVariants - HTSJDK Defaults.USE_ASYNC_IO_WRITE_FOR_SAMTOOLS : true
```

less SRR15117878_snps_from_gvcf.vcf

```
##contig=<ID=chr20,length=64444167,assembly=hg38_chr20.fa>
##reference=file:///home/ibab/NGS/18aug_lab/hg38_chr20.fa
##source=GenotypeGVCFs
##source=SelectVariants
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  sample_name
chr20   4616627 .       C       T       35.48   .       AC=2;AF=1.00;AN=2;DP=1;ExcessHet=0.0000;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=46.00;QD=25.36;SOR=1.609
        GT:AD:DP:GQ:PGT:PID:PL:PS       1|1:0,1:1:3:1|1:4616627_C_T:45,3,0:4616627
chr20   4616630 .       C       T       35.48   .       AC=2;AF=1.00;AN=2;DP=1;ExcessHet=0.0000;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=46.00;QD=28.73;SOR=1.609
        GT:AD:DP:GQ:PGT:PID:PL:PS       1|1:0,1:1:3:1|1:4616627_C_T:45,3,0:4616627
chr20   7117380 .       C       T       35.48   .       AC=2;AF=1.00;AN=2;DP=1;ExcessHet=0.0000;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=40.00;QD=30.97;SOR=1.609
        GT:AD:DP:GQ:PGT:PID:PL:PS       1|1:0,1:1:3:1|1:7117380_C_T:45,3,0:7117380
chr20   7117381 .       G       T       35.48   .       AC=2;AF=1.00;AN=2;DP=1;ExcessHet=0.0000;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=40.00;QD=27.24;SOR=1.609
        GT:AD:DP:GQ:PGT:PID:PL:PS       1|1:0,1:1:3:1|1:7117380_C_T:45,3,0:7117380
chr20   9790365 .       A       G       35.48   .       AC=2;AF=1.00;AN=2;DP=1;ExcessHet=0.0000;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=46.00;QD=25.00;SOR=1.609
        GT:AD:DP:GQ:PGT:PID:PL:PS       1|1:0,1:1:3:1|1:9790363_A_AGT:45,3,0:9790363
chr20   12357324        .       C       T       35.48   .       AC=2;AF=1.00;AN=2;DP=1;ExcessHet=0.0000;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=49.00;QD=29.56;SOR=1
.609    GT:AD:DP:GQ:PGT:PID:PL:PS       1|1:0,1:1:3:1|1:12357324_C_T:45,3,0:12357324
chr20   12357331        .       C       T       35.48   .       AC=2;AF=1.00;AN=2;DP=1;ExcessHet=0.0000;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=49.00;QD=30.62;SOR=1
.609    GT:AD:DP:GQ:PGT:PID:PL:PS       1|1:0,1:1:3:1|1:12357324_C_T:45,3,0:12357324
chr20   12357337        .       C       T       35.48   .       AC=2;AF=1.00;AN=2;DP=1;ExcessHet=0.0000;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=49.00;QD=28.17;SOR=1
.609    GT:AD:DP:GQ:PGT:PID:PL:PS       1|1:0,1:1:3:1|1:12357324_C_T:45,3,0:12357324
chr20   12357339        .       C       T       35.48   .       AC=2;AF=1.00;AN=2;DP=1;ExcessHet=0.0000;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=49.00;QD=26.80;SOR=1
.609    GT:AD:DP:GQ:PGT:PID:PL:PS       1|1:0,1:1:3:1|1:12357324_C_T:45,3,0:12357324
chr20   12357343        .       C       T       35.48   .       AC=2;AF=1.00;AN=2;DP=1;ExcessHet=0.0000;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=49.00;QD=26.00;SOR=1
.609    GT:AD:DP:GQ:PGT:PID:PL:PS       1|1:0,1:1:3:1|1:12357324_C_T:45,3,0:12357324
chr20   12357344        .       C       T       35.48   .       AC=2;AF=1.00;AN=2;DP=1;ExcessHet=0.0000;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=49.00;QD=30.02;SOR=1
.609    GT:AD:DP:GQ:PGT:PID:PL:PS       1|1:0,1:1:3:1|1:12357324_C_T:45,3,0:12357324
chr20   19854554        .       G       A       35.48   .       AC=2;AF=1.00;AN=2;DP=1;ExcessHet=0.0000;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=33.00;QD=31.98;SOR=1
.609    GT:AD:DP:GQ:PGT:PID:PL:PS       1|1:0,1:1:3:1|1:19854554_G_A:45,3,0:19854554
chr20   19854557        .       G       A       35.48   .       AC=2;AF=1.00;AN=2;DP=1;ExcessHet=0.0000;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=33.00;QD=27.51;SOR=1
.609    GT:AD:DP:GQ:PGT:PID:PL:PS       1|1:0,1:1:3:1|1:19854554_G_A:45,3,0:19854554
chr20   19854559        .       G       A       35.48   .       AC=2;AF=1.00;AN=2;DP=1;ExcessHet=0.0000;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=33.00;QD=29.11;SOR=1
.609    GT:AD:DP:GQ:PGT:PID:PL:PS       1|1:0,1:1:3:1|1:19854554_G_A:45,3,0:19854554
chr20   19854562        .       G       A       35.48   .       AC=2;AF=1.00;AN=2;DP=1;ExcessHet=0.0000;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=33.00;QD=28.08;SOR=1
.609    GT:AD:DP:GQ:PGT:PID:PL:PS       1|1:0,1:1:3:1|1:19854554_G_A:45,3,0:19854554
(END)
```

gatk SelectVariants -R hg38_chr20.fa -V SRR15117878_final_from_gvcf.vcf –select -type-to-include INDEL -O SRR15117878_indels_from_gvcf.vcf

```
ibab@LAPTOP-BVSTVK8Q:~/NGS/18aug_lab$  /home/ibab/NGS/Packages/gatk-4.6.2.0/gatk SelectVariants -R hg38_chr20.fa -V SRR15117878_final_from_gvcf.vcf --select-type-to-include INDEL -O SRR15117878_indels_from_gvcf.vcf
Using GATK jar /home/ibab/NGS/Packages/gatk-4.6.2.0/gatk-package-4.6.2.0-local.jar
Running:
    java -Dsamjdk.use_async_io_read_samtools=false -Dsamjdk.use_async_io_write_samtools=true -Dsamjdk.use_async_io_write_tribble=false -Dsamjdk.compression_level=2 -jar /home/ibab/NGS/Packages/gatk-4.6.2.0/gatk-package-4.6.2.0-local.jar SelectVariants -R hg38_chr20.fa -V SRR15117878_final_from_gvcf.vcf --select-type-to-include INDEL -O SRR15117878_indels_from_gvcf.vcf
07:08:23.231 INFO  NativeLibraryLoader - Loading libgkl_compression.so from jar:file:/home/ibab/NGS/Packages/gatk-4.6.2.0/gatk-package-4.6.2.0-local.jar!/com/intel/gkl/native/libgkl_compression.so
07:08:23.711 INFO  SelectVariants - ------------------------------------------------------------
07:08:23.717 INFO  SelectVariants - The Genome Analysis Toolkit (GATK) v4.6.2.0
07:08:23.717 INFO  SelectVariants - For support and documentation go to https://software.broadinstitute.org/gatk/
07:08:23.718 INFO  SelectVariants - Executing as ibab@LAPTOP-BVSTVK8Q on Linux v6.6.87.2-microsoft-standard-WSL2 amd64
07:08:23.718 INFO  SelectVariants - Java runtime: OpenJDK 64-Bit Server VM v17.0.16+8-Ubuntu-0ubuntu122.04.1
07:08:23.718 INFO  SelectVariants - Start Date/Time: August 22, 2025 at 7:08:23 AM IST
07:08:23.719 INFO  SelectVariants - ------------------------------------------------------------
07:08:23.719 INFO  SelectVariants - ------------------------------------------------------------
07:08:23.720 INFO  SelectVariants - HTSJDK Version: 4.2.0
07:08:23.721 INFO  SelectVariants - Picard Version: 3.4.0
07:08:23.721 INFO  SelectVariants - Built for Spark Version: 3.5.0
07:08:23.724 INFO  SelectVariants - HTSJDK Defaults.COMPRESSION_LEVEL : 2
```

```
##source=GenotypeGVCFs
##source=SelectVariants
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  sample_name
chr20   9790363 .       A       AGT     35.44   .       AC=2;AF=1.00;AN=2;DP=1;ExcessHet=0.0000;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=46.00;QD=28.20;SOR=1.609
        GT:AD:DP:GQ:PGT:PID:PL:PS       1|1:0,1:1:3:1|1:9790363_A_AGT:45,3,0:9790363
(END)
```

Conclusion:

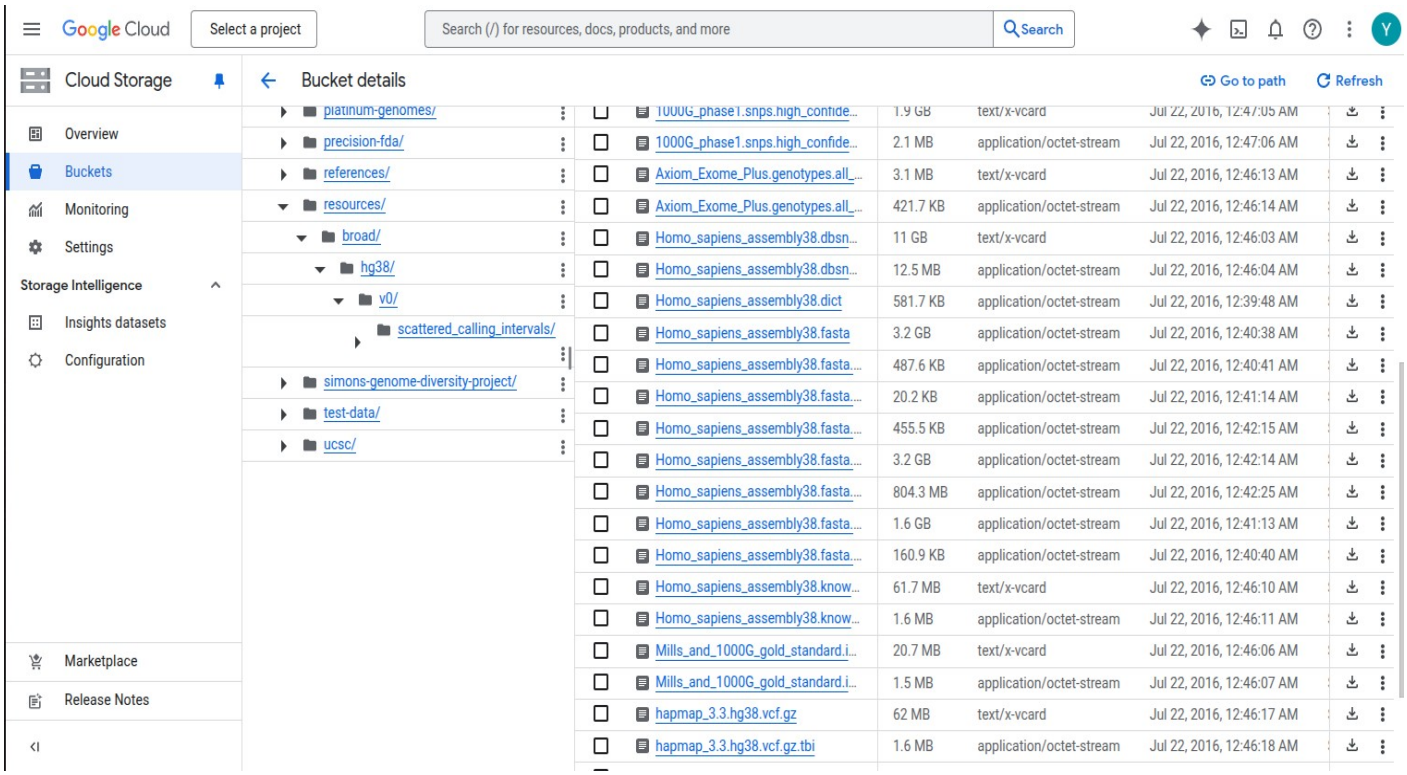
For a **single sample**, the **SNPs and indels** in the GVCF and the final VCF **are essentially the same**. The GVCF contains additional information about non-variant (homozygous reference) sites, whereas the VCF produced by GenotypeGVCFs lists only the called variants. Therefore, when comparing just SNPs or indels for one sample, we won't see any difference; the main advantage of GVCF arises when performing joint genotyping across multiple samples.

Part B: Perform Variant Calling with Known Sites

***Using the 1000 Genomes Project, dbSNP, and HapMap data as known sites for Base Quality Score Recalibration.**

Steps:

- Visit - https://console.cloud.google.com/storage/browser/genomics-public-data/resources/broad/hg38/v0;tab=objects?prefix=&forceOnObjectsSortingFiltering=false
- Download the underlined .vcf.gz files along with their .idx or .tbi (indexed) files – either manually or by using 'wget' command.




```
(base) ibab@IBAB-MScBDB2-Comp007:~/NGS/variant_calling/variant_calling_known$ ls -lh
total 13G
-rw-rw-r-- 1 ibab ibab  131 Aug 21 16:29 hg38_chr20.dict
-rw-rw-r-- 1 ibab ibab  63M Aug 21 14:56 hg38_chr20.fa
-rw-rw-r-- 1 ibab ibab   23 Aug 21 15:19 hg38_chr20.fa.fai
-rw-rw-r-- 1 ibab ibab 1.8G Aug 22 13:53 resources_broad_hg38_v0_1000G_phase1.snps.high_confidence.hg38.vcf.gz
-rw-rw-r-- 1 ibab ibab 2.1M Aug 22 13:50 resources_broad_hg38_v0_1000G_phase1.snps.high_confidence.hg38.vcf.gz.tbi
-rw-rw-r-- 1 ibab ibab  60M Aug 22 13:54 resources_broad_hg38_v0_hapmap_3.3.hg38.vcf.gz
-rw-rw-r-- 1 ibab ibab 1.5M Aug 22 13:54 resources_broad_hg38_v0_hapmap_3.3.hg38.vcf.gz.tbi
-rw-rw-r-- 1 ibab ibab  11G Aug 22 14:07 resources_broad_hg38_v0_Homo_sapiens_assembly38.dbsnp138.vcf
-rw-rw-r-- 1 ibab ibab  12M Aug 22 13:51 resources_broad_hg38_v0_Homo_sapiens_assembly38.dbsnp138.vcf.idx
-rw-rw-r-- 1 ibab ibab  20M Aug 22 18:40 resources_broad_hg38_v0_Mills_and_1000G_gold_standard.indels.hg38.vcf.gz
-rw-rw-r-- 1 ibab ibab 1.5M Aug 22 18:40 resources_broad_hg38_v0_Mills_and_1000G_gold_standard.indels.hg38.vcf.gz.tbi
-rw-rw-r-- 1 ibab ibab 162M Aug 18 17:35 SRR15117878_grpadded.bam
-rw-rw-r-- 1 ibab ibab  93K Aug 21 16:54 SRR15117878_grpadded.bam.bai
```

- Since, dbsnp138.vcf is not in compressed format, so we need to bgzip it along with indexing (tabix).

```
(base) ibab@IBAB-MScBDB2-Comp007:~/NGS/variant_calling/variant_calling_known$ bgzip -c resources_broad_hg38_v0_Homo_sapiens_assembly38.dbsnp138.vcf > resources_broad_hg38_v0_Homo_sapiens_assembly38.dbsnp138.vcf.gz
(base) ibab@IBAB-MScBDB2-Comp007:~/NGS/variant_calling/variant_calling_known$ tabix -p vcf resources_broad_hg38_v0_Homo_sapiens_assembly38.dbsnp138.vcf.gz
```

- There are two main styles of chromosome names in reference + VCFs: UCSC-style - chr1, chr2, ..., chr20, chrX, chrY, chrM, and Ensembl/NCBI-style - 1, 2, ..., 20, X, Y, MT. If the reference is chr20, all known-sites VCFs must also use chr20, or if the reference is 20, all must use 20. If not consistent then it will throw ERROR: Contig '20' not found in reference dictionary.

```
(base) ibab@IBAB-MScBDB2-Comp007:~/NGS/variant_calling/variant_calling_known$ zgrep -v "^#" resources_broad_hg38_v0_Homo_sapiens_assembly38.dbsnp138.vcf.gz | head -2
chr1    10019   rs376643643     TA      T       .       PASS    OTHERKG;R5;RS=376643643;RSPOS=10020;SAO=0;SSR=0;VC=DIV;VP=0x0500000200010000002000200;WGT=1;dbSNPBuildID=138
chr1    10109   rs376007522     A       T       .       PASS    OTHERKG;R5;RS=376007522;RSPOS=10109;SAO=0;SSR=0;VC=SNV;VP=0x0500000200010000002000100;WGT=1;dbSNPBuildID=138
(base) ibab@IBAB-MScBDB2-Comp007:~/NGS/variant_calling/variant_calling_known$ grep ">" hg38_chr20.fa | head
>chr20
```

```
(base) ibab@IBAB-MScBDB2-Comp007:~/NGS/variant_calling/variant_calling_known$ zgrep -v "^#" resources_broad_hg38_v0_hapmap_3.3.hg38.vcf.gz | head -2
chr1    55299   rs10399749      C       .       .       PASS    AN=510
chr1    55394   rs2949420       T       .       .       PASS    AN=178
(base) ibab@IBAB-MScBDB2-Comp007:~/NGS/variant_calling/variant_calling_known$ zgrep -v "^#" resources_broad_hg38_v0_1000G_phase1.snps.high_confidence.hg38.vcf.gz | head -2
chr1    51479   rs116400033     T       A       11726.81        PASS    AC=229;AF=0.3253;AN=704;BaseQRankSum=-6.949;DB;DP=1570;Dels=0.00;FS=3.130;HRun=0;HaplotypeScore=0.1377;InbreedingCoeff=0.2907;MQ=34.37;MQ0=174;MQRankSum=1.476;QD=16.08;ReadPosRankSum=-0.202;SB=-4317.78;VQSLOD=5.1635;pop=EUR.admix
chr1    55367   .       G       A       207.20  PASS    AC=2;AF=0.00117;AN=1714;BaseQRankSum=2.243;DP=4926;Dels=0.00;FS=3.005;HRun=0;HaplotypeScore=0.1382;InbreedingCoeff=-0.0188;MQ=45.57;MQ0=365;MQRankSum=0.185;QD=21.22;ReadPosRankSum=0.136;SB=-111.01;VQSLOD=6.3979;pop=ALL
(base) ibab@IBAB-MScBDB2-Comp007:~/NGS/variant_calling/variant_calling_known$ zgrep -v "^#" resources_broad_hg38_v0_Mills_and_1000G_gold_standard.indels.hg38.vcf.gz | head -2
chr1    55249   .       C       CTATGG  6160.83 PASS    set=Intersect1000GMinusBI
chr1    87114   .       CT      C       666.86  PASS    set=Intersect1000GAll
```

- Now we have known variants, there is no need to do Haplotype Calling, so we can proceed with base quality score recalibration (BQSR).

```
(base) ibab@IBAB-MScBDB2-Comp007:~/NGS/variant_calling/variant_calling_known$ gatk BaseRecalibrator \
  -R hg38_chr20.fa \
  -I SRR15117878_grpadded.bam \
  --known-sites resources_broad_hg38_v0_1000G_phase1.snps.high_confidence.hg38.vcf.gz \
  --known-sites resources_broad_hg38_v0_hapmap_3.3.hg38.vcf.gz \
  --known-sites resources_broad_hg38_v0_Homo_sapiens_assembly38.dbsnp138.vcf.gz \
  --known-sites resources_broad_hg38_v0_Mills_and_1000G_gold_standard.indels.hg38.vcf.gz \
  -L chr20 \
  -O SRR15117878_recal_data.table
Using GATK jar /home/ibab/miniconda3/share/gatk4-4.3.0.0-0/gatk-package-4.3.0.0-local.jar
Running:
    java -Dsamjdk.use_async_io_read_samtools=false -Dsamjdk.use_async_io_write_samtools=true -Dsamjdk.use_async_io_write_tribble=false -Dsamjdk.compression_level=2 -jar /home/ibab/miniconda3/share/gatk4-4.3.0.0-0/gatk-package-4.3.0.0-local.jar BaseRecalibrator -R hg38_chr20.fa -I SRR15117878_grpadded.bam --known-sites resources_broad_hg38_v0_1000G_phase1.snps.high_confidence.hg38.vcf.gz --known-sites resources_broad_hg38_v0_hapmap_3.3.hg38.vcf.gz --known-sites resources_broad_hg38_v0_Homo_sapiens_assembly38.dbsnp138.vcf.gz --known-sites resources_broad_hg38_v0_Mills_and_1000G_gold_standard.indels.hg38.vcf.gz -L chr20 -O SRR15117878_recal_data.table
19:29:41.172 INFO  NativeLibraryLoader - Loading libgkl_compression.so from jar:file:/home/ibab/miniconda3/share/gatk4-4.3.0.0-0/gatk-package-4.3.0.0-local.jar!/com/intel/gkl/native/libgkl_compression.so
```

This analyzes the sequencing reads in the BAM file (SRR15117878_grpadded.bam) against a set of known, trusted variant databases and records where the base qualities differ from expectations. Here, the reference genome is provided with -R hg38_chr20.fa to align coordinates, and -I specifies the BAM file to recalibrate. The --known-sites options give high-confidence variant datasets (dbSNP, 1000 Genomes SNPs, HapMap SNPs, and Mills + 1000G gold-standard indels) that should

not be treated as sequencing errors, allowing the model to learn real sequencing error patterns. The -L chr20 option restricts recalibration to chromosome 20 only, which saves computation since the BAM contains only chr20 reads. The output is a recalibration table (SRR15117878_recal_data.table), which is not the corrected BAM itself but a model of error patterns that will be applied in the next step (ApplyBQSR) to adjust the base quality scores.

```
#:GATKReport.v1.1:5
#:GATKTable:2:17:%s:%s:;
#:GATKTable:Arguments:Recalibration argument collection values used in this run
Argument                    Value
binary_tag_name             null
covariate                   ReadGroupCovariate,QualityScoreCovariate,ContextCovariate,CycleCovariate
default_platform            null
deletions_default_quality   45
force_platform              null
indels_context_size         3
insertions_default_quality  45
low_quality_tail            2
maximum_cycle_value         500
mismatches_context_size     2
mismatches_default_quality  -1
no_standard_covs            false
quantizing_levels           16
recalibration_report        null
run_without_dbsnp           false
solid_nocall_strategy       THROW_EXCEPTION
solid_recal_mode            SET_Q_ZERO

#:GATKTable:3:94:%d:%d:%d:;
#:GATKTable:Quantized:Quality quantization map
QualityScore  Count    QuantizedScore
           0       0              93
           1       0              93
           2       0              93
           3       0              93
           4       0              93
           5       0              93
           6       0              93
           7       0              93
           8       0              93
:
```

```
#:GATKTable:6:1:%s:%s:%.4f:%.4f:%d:%.2f:;
#:GATKTable:RecalTable0:
ReadGroup  EventType  EmpiricalQuality  EstimatedQReported  Observations  Errors
unit1      M                 14.0000             24.4729      10315006  428453.00

#:GATKTable:6:3:%s:%d:%s:%.4f:%d:%.2f:;
#:GATKTable:RecalTable1:
ReadGroup  QualityScore  EventType  EmpiricalQuality  Observations  Errors
unit1               11  M                  12.0000        410116   27671.00
unit1               25  M                  13.0000        768036   34986.00
unit1               37  M                  18.0000       9136854  365796.00

#:GATKTable:8:926:%s:%d:%s:%s:%s:%.4f:%d:%.2f:;
#:GATKTable:RecalTable2:
ReadGroup  QualityScore  CovariateValue  CovariateName  EventType  EmpiricalQuality  Observations  Errors
unit1               11  -1              Cycle          M                  18.0000          7713     11.00
unit1               11  -10             Cycle          M                  14.0000          7298    226.00
unit1               11  -100            Cycle          M                  10.0000           367     50.00
unit1               11  -101            Cycle          M                  10.0000           349     52.00
unit1               11  -102            Cycle          M                  10.0000           372     53.00
unit1               11  -103            Cycle          M                  10.0000           357     41.00
```

```
(base) ibab@IBAB-MScBDB2-Comp007:~/NGS/variant_calling/variant_calling_known$ gatk ApplyBQSR -I SRR15117878_grpadded.bam -R hg38_chr20.fa -bqsr SRR151
17878_recal_data.table -L chr20 -O SRR15117878_recalibrated_reads.bam
Using GATK jar /home/ibab/miniconda3/share/gatk4-4.3.0.0-0/gatk-package-4.3.0.0-local.jar
Running:
    java -Dsamjdk.use_async_io_read_samtools=false -Dsamjdk.use_async_io_write_samtools=true -Dsamjdk.use_async_io_write_tribble=false -Dsamjdk.compre
ssion_level=2 -jar /home/ibab/miniconda3/share/gatk4-4.3.0.0-0/gatk-package-4.3.0.0-local.jar ApplyBQSR -I SRR15117878_grpadded.bam -R hg38_chr20.fa -
bqsr SRR15117878_recal_data.table -L chr20 -O SRR15117878_recalibrated_reads.bam
19:43:28.406 INFO  NativeLibraryLoader - Loading libgkl_compression.so from jar:file:/home/ibab/miniconda3/share/gatk4-4.3.0.0-0/gatk-package-4.3.0.0-
local.jar!/com/intel/gkl/native/libgkl_compression.so
19:43:28.471 INFO  ApplyBQSR - ------------------------------------------------------------
19:43:28.471 INFO  ApplyBQSR - The Genome Analysis Toolkit (GATK) v4.3.0.0
19:43:28.471 INFO  ApplyBQSR - For support and documentation go to https://software.broadinstitute.org/gatk/
19:43:28.471 INFO  ApplyBQSR - Executing as ibab@IBAB-MScBDB2-Comp007.ibab.ac.in on Linux v6.14.0-28-generic amd64
19:43:28.471 INFO  ApplyBQSR - Java runtime: OpenJDK 64-Bit Server VM v11.0.1+13-LTS
19:43:28.471 INFO  ApplyBQSR - Start Date/Time: August 22, 2025 at 7:43:28 PM IST
19:43:28.471 INFO  ApplyBQSR - ------------------------------------------------------------
```

Part-B.a:

Use the VCF output format to save the data. After calling the final variants, extract both SNPs and Indels, and save them as separate files.

```
(base) ibab@IBAB-MScBDB2-Comp007:~/NGS/variant_calling/variant_calling_known$ gatk HaplotypeCaller \
  -R hg38_chr20.fa \
  -I SRR15117878_recalibrated_reads.bam \
  -O SRR15117878_final_variants.vcf.gz \
  -L chr20
Using GATK jar /home/ibab/miniconda3/share/gatk4-4.3.0.0-0/gatk-package-4.3.0.0-local.jar
Running:
    java -Dsamjdk.use_async_io_read_samtools=false -Dsamjdk.use_async_io_write_samtools=true -Dsamjdk.use_async_io_write_tribble=false -Dsamjdk.compre
ssion_level=2 -jar /home/ibab/miniconda3/share/gatk4-4.3.0.0-0/gatk-package-4.3.0.0-local.jar HaplotypeCaller -R hg38_chr20.fa -I SRR15117878_recalibr
ated_reads.bam -O SRR15117878_final_variants.vcf.gz -L chr20
20:07:36.639 INFO  NativeLibraryLoader - Loading libgkl_compression.so from jar:file:/home/ibab/miniconda3/share/gatk4-4.3.0.0-0/gatk-package-4.3.0.0-
local.jar!/com/intel/gkl/native/libgkl_compression.so
20:07:36.704 INFO  HaplotypeCaller - ------------------------------------------------------------
20:07:36.704 INFO  HaplotypeCaller - The Genome Analysis Toolkit (GATK) v4.3.0.0
20:07:36.704 INFO  HaplotypeCaller - For support and documentation go to https://software.broadinstitute.org/gatk/
20:07:36.704 INFO  HaplotypeCaller - Executing as ibab@IBAB-MScBDB2-Comp007.ibab.ac.in on Linux v6.14.0-28-generic amd64
20:07:36.704 INFO  HaplotypeCaller - Java runtime: OpenJDK 64-Bit Server VM v11.0.1+13-LTS
20:07:36.704 INFO  HaplotypeCaller - Start Date/Time: August 22, 2025 at 8:07:36 PM IST
```

```
(base) ibab@IBAB-MScBDB2-Comp007:~/NGS/variant_calling/variant_calling_known$ gatk SelectVariants \
  -R hg38_chr20.fa \
  -V SRR15117878_final_variants.vcf.gz \
  --select-type-to-include SNP \
  -O SRR15117878_snps.vcf.gz
Using GATK jar /home/ibab/miniconda3/share/gatk4-4.3.0.0-0/gatk-package-4.3.0.0-local.jar
Running:
    java -Dsamjdk.use_async_io_read_samtools=false -Dsamjdk.use_async_io_write_samtools=true -Dsamjdk.use_async_io_write_tribble=false -Dsamjdk.compre
ssion_level=2 -jar /home/ibab/miniconda3/share/gatk4-4.3.0.0-0/gatk-package-4.3.0.0-local.jar SelectVariants -R hg38_chr20.fa -V SRR15117878_final_var
iants.vcf.gz --select-type-to-include SNP -O SRR15117878_snps.vcf.gz
20:14:06.544 INFO  NativeLibraryLoader - Loading libgkl_compression.so from jar:file:/home/ibab/miniconda3/share/gatk4-4.3.0.0-0/gatk-package-4.3.0.0-
local.jar!/com/intel/gkl/native/libgkl_compression.so
20:14:06.608 INFO  SelectVariants - ------------------------------------------------------------
20:14:06.608 INFO  SelectVariants - The Genome Analysis Toolkit (GATK) v4.3.0.0
20:14:06.609 INFO  SelectVariants - For support and documentation go to https://software.broadinstitute.org/gatk/
```

```
##source=HaplotypeCaller
##source=SelectVariants
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  sample_name
chr20   4616627 .       C       T       35.48   .       AC=2;AF=1.00;AN=2;DP=1;ExcessHet=0.0000;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=46.00;QD=29.45;SOR=1.6
09      GT:AD:DP:GQ:PL  1/1:0,1:1:3:45,3,0
chr20   7117380 .       C       T       35.48   .       AC=2;AF=1.00;AN=2;DP=1;ExcessHet=0.0000;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=40.00;QD=34.04;SOR=1.6
09      GT:AD:DP:GQ:PL  1/1:0,1:1:3:45,3,0
chr20   7117381 .       G       T       35.48   .       AC=2;AF=1.00;AN=2;DP=1;ExcessHet=0.0000;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=40.00;QD=25.38;SOR=1.6
09      GT:AD:DP:GQ:PL  1/1:0,1:1:3:45,3,0
chr20   9790365 .       A       G       35.48   .       AC=2;AF=1.00;AN=2;DP=1;ExcessHet=0.0000;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=46.00;QD=31.41;SOR=1.6
09      GT:AD:DP:GQ:PL  1/1:0,1:1:3:45,3,0
chr20   16521665        .       G       A       35.48   .       AC=2;AF=1.00;AN=2;DP=1;ExcessHet=0.0000;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=52.00;QD=28.69
;SOR=1.609      GT:AD:DP:GQ:PL  1/1:0,1:1:3:45,3,0
chr20   16521667        .       G       A       35.48   .       AC=2;AF=1.00;AN=2;DP=3;ExcessHet=0.0000;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=48.40;QD=32.64
;SOR=1.609      GT:AD:DP:GQ:PL  1/1:0,1:1:3:45,3,0
chr20   19854554        .       G       A       35.48   .       AC=2;AF=1.00;AN=2;DP=1;ExcessHet=0.0000;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=33.00;QD=32.93
;SOR=1.609      GT:AD:DP:GQ:PL  1/1:0,1:1:3:45,3,0
chr20   19854557        .       G       A       35.48   .       AC=2;AF=1.00;AN=2;DP=1;ExcessHet=0.0000;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=33.00;QD=31.88
;SOR=1.609      GT:AD:DP:GQ:PL  1/1:0,1:1:3:45,3,0
chr20   19854559        .       G       A       35.48   .       AC=2;AF=1.00;AN=2;DP=1;ExcessHet=0.0000;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=33.00;QD=25.09
;SOR=1.609      GT:AD:DP:GQ:PL  1/1:0,1:1:3:45,3,0
chr20   19854562        .       G       A       35.48   .       AC=2;AF=1.00;AN=2;DP=1;ExcessHet=0.0000;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=33.00;QD=28.36
;SOR=1.609      GT:AD:DP:GQ:PL  1/1:0,1:1:3:45,3,0
chr20   34992258        .       A       T       35.48   .       AC=2;AF=1.00;AN=2;DP=1;ExcessHet=0.0000;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=26.00;QD=32.08
;SOR=1.609      GT:AD:DP:GQ:PL  1/1:0,1:1:3:45,3,0
chr20   34992260        .       C       T       35.48   .       AC=2;AF=1.00;AN=2;DP=1;ExcessHet=0.0000;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=26.00;QD=27.85
;SOR=1.609      GT:AD:DP:GQ:PL  1/1:0,1:1:3:45,3,0
chr20   34992261        .       C       T       35.48   .       AC=2;AF=1.00;AN=2;DP=1;ExcessHet=0.0000;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=26.00;QD=33.72
;SOR=1.609      GT:AD:DP:GQ:PL  1/1:0,1:1:3:45,3,0
(END)
```

```
(base) ibab@IBAB-MScBDB2-Comp007:~/NGS/variant_calling/variant_calling_known$ gatk SelectVariants -R hg38_chr20.fa -V SRR15117878_final_variants.vcf.g
z --select-type-to-include INDEL -O SRR15117878_indels.vcf.gz
Using GATK jar /home/ibab/miniconda3/share/gatk4-4.3.0.0-0/gatk-package-4.3.0.0-local.jar
Running:
    java -Dsamjdk.use_async_io_read_samtools=false -Dsamjdk.use_async_io_write_samtools=true -Dsamjdk.use_async_io_write_tribble=false -Dsamjdk.compre
ssion_level=2 -jar /home/ibab/miniconda3/share/gatk4-4.3.0.0-0/gatk-package-4.3.0.0-local.jar SelectVariants -R hg38_chr20.fa -V SRR15117878_final_var
iants.vcf.gz --select-type-to-include INDEL -O SRR15117878_indels.vcf.gz
20:19:58.308 INFO  NativeLibraryLoader - Loading libgkl_compression.so from jar:file:/home/ibab/miniconda3/share/gatk4-4.3.0.0-0/gatk-package-4.3.0.0-
local.jar!/com/intel/gkl/native/libgkl_compression.so
20:19:58.374 INFO  SelectVariants - ------------------------------------------------------------
20:19:58.375 INFO  SelectVariants - The Genome Analysis Toolkit (GATK) v4.3.0.0
20:19:58.375 INFO  SelectVariants - For support and documentation go to https://software.broadinstitute.org/gatk/
20:19:58.375 INFO  SelectVariants - Executing as ibab@IBAB-MScBDB2-Comp007.ibab.ac.in on Linux v6.14.0-28-generic amd64
20:19:58.375 INFO  SelectVariants - Java runtime: OpenJDK 64-Bit Server VM v11.0.1+13-LTS
20:19:58.375 INFO  SelectVariants - Start Date/Time: August 22, 2025 at 8:19:58 PM IST
```

```
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  sample_name
chr20   9790363 .       A       AGT     35.44   .       AC=2;AF=1.00;AN=2;DP=1;ExcessHet=0.0000;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=46.00;QD=24.79;SOR=1.6
09      GT:AD:DP:GQ:PL  1/1:0,1:1:3:45,3,0
(END)
```

Part-B.b:

Use the Gvcf output format to save the data. After calling the final variants, extract both the SNPs and Indels, and save them as separate files.

```
(base) ibab@IBAB-MScBDB2-Comp007:~/NGS/variant_calling/variant_calling_known$ gatk HaplotypeCaller \
  -R hg38_chr20.fa \
  -I SRR15117878_recalibrated_reads.bam \
  -O SRR15117878_final_gvcf_variants.g.vcf.gz \
  -ERC GVCF \
  -L chr20
Using GATK jar /home/ibab/miniconda3/share/gatk4-4.3.0.0-0/gatk-package-4.3.0.0-local.jar
Running:
    java -Dsamjdk.use_async_io_read_samtools=false -Dsamjdk.use_async_io_write_samtools=true -Dsamjdk.use_async_io_write_tribble=false -Dsamjdk.compre
ssion_level=2 -jar /home/ibab/miniconda3/share/gatk4-4.3.0.0-0/gatk-package-4.3.0.0-local.jar HaplotypeCaller -R hg38_chr20.fa -I SRR15117878_recalibr
ated_reads.bam -O SRR15117878_final_gvcf_variants.g.vcf.gz -ERC GVCF -L chr20
20:27:43.999 INFO  NativeLibraryLoader - Loading libgkl_compression.so from jar:file:/home/ibab/miniconda3/share/gatk4-4.3.0.0-0/gatk-package-4.3.0.0-
local.jar!/com/intel/gkl/native/libgkl_compression.so
```

A raw gVCF still contains reference confidence blocks, so we need to do GenotypeGVCFs first before splitting.

```
(base) ibab@IBAB-MScBDB2-Comp007:~/NGS/variant_calling/variant_calling_known$ gatk GenotypeGVCFs \
  -R hg38_chr20.fa \
  -V SRR15117878_final_gvcf_variants.g.vcf.gz \
  -O SRR15117878_genotyped.vcf.gz
Using GATK jar /home/ibab/miniconda3/share/gatk4-4.3.0.0-0/gatk-package-4.3.0.0-local.jar
Running:
    java -Dsamjdk.use_async_io_read_samtools=false -Dsamjdk.use_async_io_write_samtools=true -Dsamjdk.use_async_io_write_tribble=false -Dsamjdk.compre
ssion_level=2 -jar /home/ibab/miniconda3/share/gatk4-4.3.0.0-0/gatk-package-4.3.0.0-local.jar GenotypeGVCFs -R hg38_chr20.fa -V SRR15117878_final_gvcf
_variants.g.vcf.gz -O SRR15117878_genotyped.vcf.gz
20:33:41.116 INFO  NativeLibraryLoader - Loading libgkl_compression.so from jar:file:/home/ibab/miniconda3/share/gatk4-4.3.0.0-0/gatk-package-4.3.0.0-
local.jar!/com/intel/gkl/native/libgkl_compression.so
20:33:41.182 INFO  GenotypeGVCFs - ------------------------------------------------------------
20:33:41.183 INFO  GenotypeGVCFs - The Genome Analysis Toolkit (GATK) v4.3.0.0
20:33:41.183 INFO  GenotypeGVCFs - For support and documentation go to https://software.broadinstitute.org/gatk/
```

```
(base) ibab@IBAB-MScBDB2-Comp007:~/NGS/variant_calling/variant_calling_known$ gatk SelectVariants \
  -R hg38_chr20.fa \
  -V SRR15117878_genotyped.vcf.gz \
  --select-type-to-include SNP \
  -O SRR15117878_genotyped_snps.vcf.gz
Using GATK jar /home/ibab/miniconda3/share/gatk4-4.3.0.0-0/gatk-package-4.3.0.0-local.jar
Running:
    java -Dsamjdk.use_async_io_read_samtools=false -Dsamjdk.use_async_io_write_samtools=true -Dsamjdk.use_async_io_write_tribble=false -Dsamjdk.compre
ssion_level=2 -jar /home/ibab/miniconda3/share/gatk4-4.3.0.0-0/gatk-package-4.3.0.0-local.jar SelectVariants -R hg38_chr20.fa -V SRR15117878_genotyped
.vcf.gz --select-type-to-include SNP -O SRR15117878_genotyped_snps.vcf.gz
20:35:14.512 INFO  NativeLibraryLoader - Loading libgkl_compression.so from jar:file:/home/ibab/miniconda3/share/gatk4-4.3.0.0-0/gatk-package-4.3.0.0-
local.jar!/com/intel/gkl/native/libgkl_compression.so
20:35:14.577 INFO  SelectVariants - ------------------------------------------------------------
20:35:14.577 INFO  SelectVariants - The Genome Analysis Toolkit (GATK) v4.3.0.0
20:35:14.577 INFO  SelectVariants - For support and documentation go to https://software.broadinstitute.org/gatk/
20:35:14.577 INFO  SelectVariants - Executing as ibab@IBAB-MScBDB2-Comp007.ibab.ac.in on Linux v6.14.0-28-generic amd64
```

```
##source=GenotypeGVCFs
##source=SelectVariants
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  sample_name
chr20   4616627 .       C       T       35.48   .       AC=2;AF=1.00;AN=2;DP=1;ExcessHet=0.0000;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=46.00;QD=25.36;SOR=1.
chr20   7117380 .       C       T       35.48   .       AC=2;AF=1.00;AN=2;DP=1;ExcessHet=0.0000;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=40.00;QD=28.73;SOR=1.
chr20   7117381 .       G       T       35.48   .       AC=2;AF=1.00;AN=2;DP=1;ExcessHet=0.0000;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=40.00;QD=30.97;SOR=1.
chr20   9790365 .       A       G       35.48   .       AC=2;AF=1.00;AN=2;DP=1;ExcessHet=0.0000;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=46.00;QD=28.20;SOR=1.
chr20   16521665        .       G       A       35.48   .       AC=2;AF=1.00;AN=2;DP=1;ExcessHet=0.0000;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=52.00;QD=25.0
chr20   16521667        .       G       A       35.48   .       AC=2;AF=1.00;AN=2;DP=3;ExcessHet=0.0000;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=48.40;QD=29.5
chr20   19854554        .       G       A       35.48   .       AC=2;AF=1.00;AN=2;DP=1;ExcessHet=0.0000;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=33.00;QD=30.6
chr20   19854557        .       G       A       35.48   .       AC=2;AF=1.00;AN=2;DP=1;ExcessHet=0.0000;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=33.00;QD=28.1
chr20   19854559        .       G       A       35.48   .       AC=2;AF=1.00;AN=2;DP=1;ExcessHet=0.0000;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=33.00;QD=26.8
chr20   19854562        .       G       A       35.48   .       AC=2;AF=1.00;AN=2;DP=1;ExcessHet=0.0000;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=33.00;QD=26.0
chr20   34992258        .       A       T       35.48   .       AC=2;AF=1.00;AN=2;DP=1;ExcessHet=0.0000;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=26.00;QD=30.0
chr20   34992260        .       C       T       35.48   .       AC=2;AF=1.00;AN=2;DP=1;ExcessHet=0.0000;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=26.00;QD=31.9
chr20   34992261        .       C       T       35.48   .       AC=2;AF=1.00;AN=2;DP=1;ExcessHet=0.0000;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=26.00;QD=27.5
(END)
```

```
(base) ibab@IBAB-MScBDB2-Comp007:~/NGS/variant_calling/variant_calling_known$ gatk SelectVariants \
  -R hg38_chr20.fa \
  -V SRR15117878_genotyped.vcf.gz \
  --select-type-to-include INDEL \
  -O SRR15117878_genotyped_indels.vcf.gz
Using GATK jar /home/ibab/miniconda3/share/gatk4-4.3.0.0-0/gatk-package-4.3.0.0-local.jar
Running:
    java -Dsamjdk.use_async_io_read_samtools=false -Dsamjdk.use_async_io_write_samtools=true -Dsamjdk.use_async_io_write_tribble=false -Dsamjdk.compression_level=2 -jar /home/ibab/miniconda3/share/gatk4-4.3.0.0-0/gatk-package-4.3.0.0-local.jar SelectVariants -R hg38_chr20.fa -V SRR15117878_genotyped.vcf.gz --select-type-to-include INDEL -O SRR15117878_genotyped_indels.vcf.gz
20:38:11.615 INFO  NativeLibraryLoader - Loading libgkl_compression.so from jar:file:/home/ibab/miniconda3/share/gatk4-4.3.0.0-0/gatk-package-4.3.0.0-local.jar!/com/intel/gkl/native/libgkl_compression.so
20:38:11.689 INFO  SelectVariants - ------------------------------------------------------------
20:38:11.690 INFO  SelectVariants - The Genome Analysis Toolkit (GATK) v4.3.0.0
20:38:11.690 INFO  SelectVariants - For support and documentation go to https://software.broadinstitute.org/gatk/
20:38:11.690 INFO  SelectVariants - Executing as ibab@IBAB-MScBDB2-Comp007.ibab.ac.in on Linux v6.14.0-28-generic amd64
20:38:11.690 INFO  SelectVariants - Java runtime: OpenJDK 64-Bit Server VM v11.0.1+13-LTS
```

```
##source=GenotypeGVCFs
##source=SelectVariants
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  sample_name
chr20   9790363 .       A       AGT     35.44   .       AC=2;AF=1.00;AN=2;DP=1;ExcessHet=0.0000;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=46.00;QD=27.24;SOR=1.609     GT:AD:DP:GQ:PGT:PID:PL:PS       1|1:0,1:1:3:1|1:9790363_A_AGT:45,3,0:9790363
(END)
```

Conclusion:

Variant calling was successfully performed in both VCF and gVCF modes using GATK's HaplotypeCaller after Base Quality Score Recalibration (BQSR). The variants were then separated into SNPs and Indels with GATK SelectVariants. When comparing the results, both approaches yielded the same set of variants for the region analyzed (chr20), confirming that the difference between VCF and gVCF modes lies mainly in how the data is stored:

- VCF mode directly reports only the variant positions.
- gVCF mode includes both variant calls and reference confidence blocks, but after genotyping, it produces the same SNPs and Indels as standard VCF mode.

Thus, for a single sample, both pipelines converge to the same biological results, but gVCF provides additional flexibility for future multi-sample joint genotyping.

Part-C:

What are the key differences between VCF and gVCF?

Feature	VCF (Variant Call Format)	gVCF (Genomic VCF / GVCF)
Purpose	Records only the positions in the genome where a sample has a variant (SNP or INDEL) relative to the reference.	Records all positions in the genome, including both variant and non-variant sites, along with confidence information.
Content	Contains only called variants. Non-variant positions are omitted.	Contains variants and non-variant blocks, summarizing contiguous regions that match the reference. Non-variant blocks include coverage and genotype likelihoods.
Output use	Suitable for single-sample analysis or when only the final variant list is needed.	Designed for multi-sample joint genotyping, allowing accurate calling of variants across multiple samples.
File size	Smaller, because only variant	Larger, because it includes information for all

	positions are recorded.	positions (variant + non-variant).
Downstream workflow	Can be directly used for filtering, annotation, or variant analysis.	Needs to be processed with GenotypeGVCFs (or similar tools) to convert to a standard VCF before filtering or annotation.
Handling missing sites	Missing positions are not represented; a position absent in a VCF could be truly reference or just not called.	All positions are represented; preserves genotype likelihoods for every base, which prevents ambiguity in joint genotyping.
Multi-sample support	Difficult to merge multiple single-sample VCFs accurately; sites missing in one sample may cause inconsistencies.	Designed for merging: gVCFs from multiple samples can be combined using CombineGVCFs or GenomicsDBImport, then jointly genotyped to produce a consistent multi-sample VCF.
Confidence information	Only at variant sites (QUAL score, filter, annotations).	Confidence info (genotype likelihoods, coverage) is stored for all sites, allowing better downstream filtering and evaluation.
HaplotypeCaller command	Default output: -O sample.vcf	Must specify -ERC GVCF to generate gVCF output: -O sample.g.vcf.gz
Use in pipelines	Quick, simple workflows where only the variants of one sample are needed.	Essential for large-scale studies, population genomics, or clinical pipelines requiring joint genotyping across many samples.
Representation of non-variant regions	Not represented.	Represented as contiguous blocks (<NON_REF>), storing reference confidence and read support.
Flexibility for later analysis	Limited: combining multiple VCFs can be tricky and may lose information.	Highly flexible: enables accurate multi-sample variant calling without losing reference information.

Mode	Steps	Explanation / Reasoning
VCF	**BAM preprocessing (MarkDuplicates + AddOrReplaceReadGroups)**	• Clean BAM: sort, add read groups, mark duplicates. Necessary to remove PCR artifacts and ensure proper read group info. • Independent of output type because both VCF and gVCF workflows require the same high-quality BAM.
gVCF	**BAM preprocessing (same)**	Same as VCF mode as BAM preparation is universal and required before any variant calling.
VCF	**Initial variant calling (raw VCF, optional for humans)**	• Generate preliminary VCF to be used as known-sites for BQSR. • Provides known sites for recalibration if no external databases are available (e.g., yeast) • Step is the same for both workflows because its purpose is only to provide input for BQSR, not the final variant output.
gVCF	**Initial variant calling (same)**	Same as VCF mode. Only used to provide known-sites for BQSR; final output format does not matter.
VCF	**BaseRecalibrator + ApplyBQSR**	• Correct systematic errors in base quality scores using known-sites. • Improves variant calling accuracy; BAM-level correction, independent of whether the final output will be VCF or gVCF.

gVCF	**BaseRecalibrator + ApplyBQSR (same)**	Same as VCF mode. Recalibration depends on BAM and known-sites; output type is irrelevant.
VCF	**Final variant calling (HaplotypeCaller, VCF output)**	• Calls variants only at positions that differ from reference; produces final.vcf. • Diverges here from gVCF.as it does not outputs reference blocks.
gVCF	**Final variant calling (HaplotypeCaller, gVCF output)**	Calls variants in -ERC GVCF mode - produces final.g.vcf.gz containing both variant & reference blocks (required for joint genotyping or multi-sample workflows).
VCF	**CNNScoreVariants**	• Filter variants using deep learning scoring. • Evaluates variants using BAM reads to improve confidence. • Step is the same because scoring uses the BAM and final VCF, independent of whether it came from VCF or genotyped gVCF.
gVCF	**CNNScoreVariants**	Uses BAM to assess variant likelihood; works identically for genotyped gVCF VCF.
VCF	**SelectVariants (split SNPs/INDELs)**	• Separate final VCF into SNPs (*_snps.vcf) and INDELs (*_indels.vcf), and prepare outputs for downstream analysis. • Step is the same because extraction operates on the final VCF; the origin (VCF vs gVCF) does not affect extraction.
gVCF	**SelectVariants (split SNPs/INDELs)**	Same as VCF mode; extraction only requires the final VCF. SNP/INDEL sets are comparable whether derived from direct VCF or genotyped gVCF.