245HSBD016

Yash Priya Baid

NGS Data Analysis Practical – Test 1

Part-B: Variant Calling

Choose Yeast (hard filtering) OR Human (VQSR filtering).

Tasks:

1. **Perform read QC and trimming.**
2. **Map reads to the reference genome.**
3. **Call variants (GATK HaplotypeCaller or equivalent).**
4. **Apply filtering:**
   - **Yeast: Hard filtering (specify the thresholds used).**
   - **Human: VQSR filtering (specify the tranche chosen).**
5. **Annotate the filtered variants.**

Step-0: Yeast's paired-end data (SRR22300007) was downloaded in the 1_Raw_data sub-directory before beginning the test. Following sub-directories were created in 'variant_calling' directory:

mkdir 1_Raw_data 2_fastqc 3_trimming 4_Reference 5_Alignment

Since, reference genome was provided a bit later, so 4_Reference sub-directory was created later, and thus ref. genome (GCF_000146045.2_R64_genomic) was downloaded in it.

Step-1: FASTQC

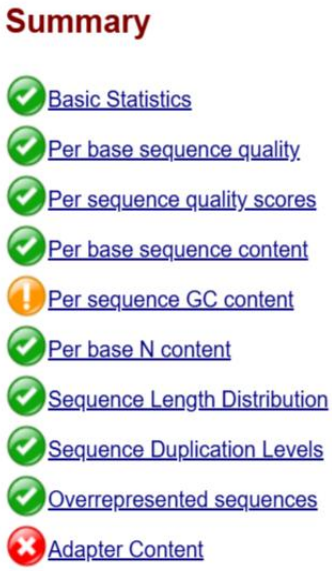
cd 2_fastqc

fastqc -o . /home/ibab/NGS/variant_calling/1_Raw_data/*.gz

```
ibab@LAPTOP-BVSTVK8Q:~/NGS/variant_calling$ cd 2_fastqc/
ibab@LAPTOP-BVSTVK8Q:~/NGS/variant_calling/2_fastqc$ fastqc -o . /home/ibab/NGS/variant_calling/1_Raw_data/*.gz
Started analysis of SRR22300007_1.fastq.gz
Approx 5% complete for SRR22300007_1.fastq.gz
Approx 10% complete for SRR22300007_1.fastq.gz
Approx 15% complete for SRR22300007_1.fastq.gz
Approx 20% complete for SRR22300007_1.fastq.gz
Approx 25% complete for SRR22300007_1.fastq.gz
Approx 30% complete for SRR22300007_1.fastq.gz
Approx 35% complete for SRR22300007_1.fastq.gz
```

Before trimming – file size

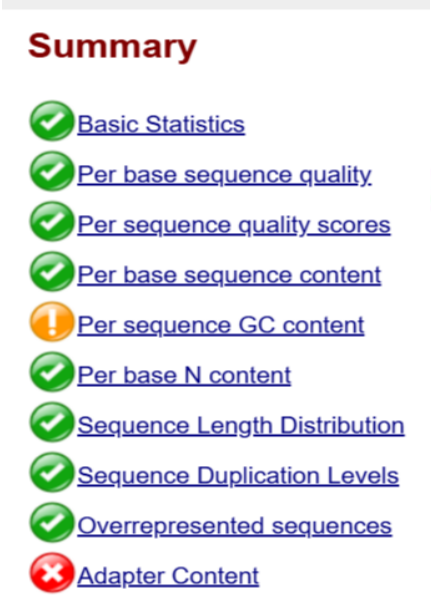
```
ibab@LAPTOP-BVSTVK8Q:~/NGS/variant_calling/2_fastqc$ ls -lh
total 1.8M
-rw-r--r-- 1 ibab ibab 565K Aug 28 06:14 SRR22300007_1_fastqc.html
-rw-r--r-- 1 ibab ibab 313K Aug 28 06:14 SRR22300007_1_fastqc.zip
-rw-r--r-- 1 ibab ibab 568K Aug 28 06:15 SRR22300007_2_fastqc.html
-rw-r--r-- 1 ibab ibab 315K Aug 28 06:15 SRR22300007_2_fastqc.zip
```



Before trimming – SRR22300007_1

Basic Statistics

Measure	Value
Filename	SRR22300007_1.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	4038492
Sequences flagged as poor quality	0
Sequence length	150
%GC	38



Before trimming – SRR22300007_2

Basic Statistics

Measure	Value
Filename	SRR22300007_2.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	4038492
Sequences flagged as poor quality	0
Sequence length	150
%GC	38

Step-2: Trimming (default parameters) along with FASTQC

cd 3_trimming

/home/ibab/NGS/Packages/TrimGalore/trim_galore --paired /home/ibab/NGS/variant_calling/1_Raw_data/SRR22300007_1.fastq.gz /home/ibab/NGS/variant_calling/1_Raw_data/SRR22300007_2.fastq.gz -q 25 --stringency 5 --fastqc -o .

```
ibab@LAPTOP-BVSTVK8Q:~/NGS/variant_calling/3_trimming$ /home/ibab/NGS/Packages/TrimGalore/trim_galore --paired /home/ibab/NGS/variant_calling/1_Raw_data/SRR
22300007_1.fastq.gz /home/ibab/NGS/variant_calling/1_Raw_data/SRR22300007_2.fastq.gz -q 25 --stringency 5 --fastqc -o .
Multicore support not enabled. Proceeding with single-core trimming.
Path to Cutadapt set as: 'cutadapt' (default)
Cutadapt seems to be working fine (tested command 'cutadapt --version')
Cutadapt version: 3.5
single-core operation.
Proceeding with 'gzip' for decompression
To decrease CPU usage of decompression, please install 'igzip' and run again

No quality encoding type selected. Assuming that the data provided uses Sanger encoded Phred scores (default)

Output will be written into the directory: /home/ibab/NGS/variant_calling/3_trimming/


AUTO-DETECTING ADAPTER TYPE
===========================
Attempting to auto-detect adapter type from the first 1 million sequences of the first file (>> /home/ibab/NGS/variant_calling/1_Raw_data/SRR22300007_1.fast
q.gz <<)
```

After trimming – file size

```
ibab@LAPTOP-BVSTVK8Q:~/NGS/variant_calling/3_trimming$ ls -lh
total 590M
-rw-r--r-- 1 ibab ibab 4.8K Aug 28 06:25 SRR22300007_1.fastq.gz_trimming_report.txt
-rw-r--r-- 1 ibab ibab 287M Aug 28 06:32 SRR22300007_1_val_1.fq.gz
-rw-r--r-- 1 ibab ibab 576K Aug 28 06:33 SRR22300007_1_val_1_fastqc.html
-rw-r--r-- 1 ibab ibab 283K Aug 28 06:33 SRR22300007_1_val_1_fastqc.zip
-rw-r--r-- 1 ibab ibab 4.9K Aug 28 06:32 SRR22300007_2.fastq.gz_trimming_report.txt
-rw-r--r-- 1 ibab ibab 302M Aug 28 06:32 SRR22300007_2_val_2.fq.gz
-rw-r--r-- 1 ibab ibab 583K Aug 28 06:33 SRR22300007_2_val_2_fastqc.html
-rw-r--r-- 1 ibab ibab 292K Aug 28 06:33 SRR22300007_2_val_2_fastqc.zip
```

Summary

- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content

After trimming – SRR22300007_1

Basic Statistics

Measure	Value
Filename	SRR22300007_1_val_1.fq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	3990936
Sequences flagged as poor quality	0
Sequence length	20-150
%GC	37

Summary

- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content

After trimming – SRR22300007_2

Basic Statistics

Measure	Value
Filename	SRR22300007_2_val_2.fq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	3990936
Sequences flagged as poor quality	0
Sequence length	20-150
%GC	37

FASTQC reports – before and after trimming (.html) have been attached.

Summary Reports:

```
ibab@LAPTOP-BVSTVK8Q:~/NGS/variant_calling/3_trimming$ cat SRR22300007_1.fastq.gz_trimming_report.txt

SUMMARISING RUN PARAMETERS
==========================
Input filename: /home/ibab/NGS/variant_calling/1_Raw_data/SRR22300007_1.fastq.gz
Trimming mode: paired-end
Trim Galore version: 0.6.10
Cutadapt version: 3.5
Number of cores used for trimming: 1
Quality Phred score cutoff: 25
Quality encoding type selected: ASCII+33
Using Illumina adapter for trimming (count: 173300). Second best hit was Nextera (count: 1)
Adapter sequence: 'AGATCGGAAGAGC' (Illumina TruSeq, Sanger iPCR; auto-detected)
Maximum trimming error rate: 0.1 (default)
Minimum required adapter overlap (stringency): 5 bp
Minimum required sequence length for both reads before a sequence pair gets removed: 20 bp
Running FastQC on the data once trimming has completed
Output file will be GZIP compressed
```

```
=== Summary ===

Total reads processed:              4,038,492
Reads with adapters:                  810,719 (20.1%)
Reads written (passing filters):    4,038,492 (100.0%)

Total basepairs processed:   605,773,800 bp
Quality-trimmed:               7,344,495 bp (1.2%)
Total written (filtered):    568,233,766 bp (93.8%)

=== Adapter 1 ===

Sequence: AGATCGGAAGAGC; Type: regular 3'; Length: 13; Trimmed: 810719 times

Minimum overlap: 5
No. of allowed errors:
1-9 bp: 0; 10-13 bp: 1

Bases preceding removed adapters:
  A: 27.7%
  C: 18.4%
  G: 19.1%
  T: 34.6%
  none/other: 0.2%

Overview of removed sequences
length  count   expect  max.err error counts
5       16006   3943.8  0       16006
6       15485   986.0   0       15485
7       16171   246.5   0       16171
8       13863   61.6    0       13863
9       16610   15.4    0       16509 101
10      15838   3.9     1       15508 330
11      11259   1.0     1       10961 298
12      15368   0.2     1       15023 345
13      14816   0.1     1       14442 374
14      14244   0.1     1       13906 338
15      13688   0.1     1       13350 338
16      12932   0.1     1       12583 349
17      13791   0.1     1       13417 374
```

SRR22300007_1

```
=== Summary ===

Total reads processed:              4,038,492
Reads with adapters:                  793,520 (19.6%)
Reads written (passing filters):    4,038,492 (100.0%)

Total basepairs processed:   605,773,800 bp
Quality-trimmed:              22,996,098 bp (3.8%)
Total written (filtered):    552,929,129 bp (91.3%)

=== Adapter 1 ===

Sequence: AGATCGGAAGAGC; Type: regular 3'; Length: 13; Trimmed: 793520 times

Minimum overlap: 5
No. of allowed errors:
1-9 bp: 0; 10-13 bp: 1

Bases preceding removed adapters:
  A: 27.5%
  C: 18.6%
  G: 19.1%
  T: 34.5%
  none/other: 0.2%

Overview of removed sequences
length  count   expect  max.err error counts
5       16233   3943.8  0       16233
6       15377   986.0   0       15377
7       17242   246.5   0       17242
8       11927   61.6    0       11927
9       17001   15.4    0       16922 79
10      14327   3.9     1       13990 337
11      10906   1.0     1       10649 257
12      15572   0.2     1       15210 362
13      12913   0.1     1       12552 361
14      25047   0.1     1       24559 488
```

SRR22300007_2

Rest of the removed sequences can be seen from the attached trimming_report.txt.

Step-3: Index the reference genome

bwa index GCF_000146045.2_R64_genomic.fna

```
ibab@LAPTOP-BVSTVK8Q:~/NGS/variant_calling/4_Reference$ ls
GCF_000146045.2_R64_genomic.fna      GCF_000146045.2_R64_genomic.fna.ann  GCF_000146045.2_R64_genomic.fna.pac
GCF_000146045.2_R64_genomic.fna.amb  GCF_000146045.2_R64_genomic.fna.bwt  GCF_000146045.2_R64_genomic.fna.sa
```

Step-4: Map the validated reads to the reference genome

cd 5_Alignment

bwa mem /home/ibab/NGS/variant_calling/4_Reference/GCF_000146045.2_R64_genomic.fna /home/ibab/NGS/variant_calling/3_trimming/SRR22300007_1_val_1.fq.gz /home/ibab/NGS/variant_calling/3_trimming/SRR22300007_2_val_2.fq.gz -o SRR22300007.sam

```
ibab@LAPTOP-BVSTVK8Q:~/NGS/variant_calling/5_Alignment$ bwa mem /home/ibab/NGS/variant_calling/4_Reference/GCF_000146045.2_R64_genomic.fna /home/ibab/NGS/va
riant_calling/3_trimming/SRR22300007_1_val_1.fq.gz /home/ibab/NGS/variant_calling/3_trimming/SRR22300007_2_val_2.fq.gz -o SRR22300007.sam
[M::bwa_idx_load_from_disk] read 0 ALT contigs
[M::process] read 71500 sequences (10000083 bp)...
[M::process] read 71508 sequences (10000275 bp)...
[M::mem_pestat] # candidate unique pairs for (FF, FR, RF, RR): (3, 28816, 0, 0)
[M::mem_pestat] skip orientation FF as there are not enough pairs
[M::mem_pestat] analyzing insert size distribution for orientation FR...
[M::mem_pestat] (25, 50, 75) percentile: (154, 216, 294)
[M::mem_pestat] low and high boundaries for computing mean and std.dev: (1, 574)
[M::mem_pestat] mean and std.dev: (230.71, 101.20)
[M::mem_pestat] low and high boundaries for proper pairs: (1, 714)
[M::mem_pestat] skip orientation RF as there are not enough pairs
[M::mem_pestat] skip orientation RR as there are not enough pairs
[M::mem_process_seqs] Processed 71500 reads in 6.363 CPU sec, 6.190 real sec
```

less -S SRR22300007.sam

```
@SQ     SN:NC_001133.9  LN:230218
@SQ     SN:NC_001134.8  LN:813184
@SQ     SN:NC_001135.5  LN:316620
@SQ     SN:NC_001136.10 LN:1531933
@SQ     SN:NC_001137.3  LN:576874
@SQ     SN:NC_001138.5  LN:270161
@SQ     SN:NC_001139.9  LN:1090940
@SQ     SN:NC_001140.6  LN:562643
@SQ     SN:NC_001141.2  LN:439888
@SQ     SN:NC_001142.9  LN:745751
@SQ     SN:NC_001143.9  LN:666816
@SQ     SN:NC_001144.5  LN:1078177
@SQ     SN:NC_001145.3  LN:924431
@SQ     SN:NC_001146.8  LN:784333
@SQ     SN:NC_001147.6  LN:1091291
@SQ     SN:NC_001148.4  LN:948066
@SQ     SN:NC_001224.1  LN:85779
@PG     ID:bwa  PN:bwa  VN:0.7.17-r1188 CL:bwa mem /home/ibab/NGS/variant_calling/4_Reference/GCF_000146045.2_R64_genomic.fna /home/ibab/NGS/variant_callin>
SRR22300007.1   99      NC_001136.10    142543  60      150M    =       142726  332     CNTTGTAATGAACCTTCCCACTGATTACATTCAGTATAAGAACATGTACAACCCCATGTTTCCAGAT>
SRR22300007.1   147     NC_001136.10    142726  60      149M    =       142543  -332    CCTTTTGCCTATAGAACCATCCCATTTAAGATCGTCACCCGTGGTATCTTCGTTTTGCAATATGTAT>
SRR22300007.2   83      NC_001140.6     181815  60      150M    =       181635  -330    AAGCCAACTCCGAAGCCGACAAAAATTGGATAGGCCTACTATATAAGGGTGACAGGACAGCTGACTT>
SRR22300007.2   163     NC_001140.6     181635  60      150M    =       181815  330     AAGAGCTACTATGGCAATGGCATCAAAACATATCCCTTGCCATCTTCACCGAAATCGCCGACATCAG>
SRR22300007.3   99      NC_001137.3     134153  60      150M    =       134308  305     TTGACGTTTATCAGTTTGATGGGTTTAGATTCGATGGTGTCACATCAATGTTATACGTTCATCATGG>
SRR22300007.3   147     NC_001137.3     134308  60      150M    =       134153  -305    TTAATGTTAGCCAATGATTTGGTTCACGAAATGTTGCCAAATCTGGCTGTAACTGTTGCAGAAGATG>
SRR22300007.4   83      NC_001145.3     774268  60      150M    =       774044  -374    CGCACTATGACTAAATGGTGCTGGACATCTCCATGGCTGTGACTTGTGTGTATCTCACAGTGGTAAC>
SRR22300007.4   163     NC_001145.3     774044  60      149M    =       774268  374     CACACACATCACATAGACTGCGTCATAAAAATACACTACGGAAAAACCATAAAGAGCAAAGCGATAC>
SRR22300007.5   83      NC_001148.4     684477  60      150M    =       684290  -337    AATATGAAGAAATTTCCCAAGCAAAGCTCAATAAAATTATTTAGCTAAACAAACATATGCATATATA>
SRR22300007.5   163     NC_001148.4     684290  60      150M    =       684477  337     TACATGGGAGTAAATCTGCTACTTGAGACTGAAATATTGACGTTGATATGACCAATTTCGACATAGC>
```

Step-5: Convert .sam into .bam file and then, sort it.

samtools view -bS SRR22300007.sam -o SRR22300007.bam

samtools sort -o SRR22300007_sorted.bam SRR22300007.bam

```
ibab@LAPTOP-BVSTVK8Q:~/NGS/variant_calling/5_Alignment$ ls -lh
total 4.1G
-rw-r--r-- 1 ibab ibab 725M Aug 28 10:53 SRR22300007.bam
-rw-r--r-- 1 ibab ibab 3.0G Aug 28 10:50 SRR22300007.sam
-rw-r--r-- 1 ibab ibab 428M Aug 28 10:57 SRR22300007_sorted.bam
```

samtools view SRR22300007_sorted.bam

```
TTTTGTAATTGGTCATCGTAAATTTGTGCCTTGTAAGTTGCTTGGAACAAAGTACCTTTAGACAATTCCTTGTGGAAGCACTTACGTTCA      JJJJJJJJJJJJJJJJJFFFF<<<<JJFF7JFJFJJJJAJJJJFFJJFJJJJJJJJJJJ
JJJJJJJJJJFFJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJFJJJJJFJJJJJJJJJJJJJJJJJJJJJJJJJJJJJFFFAA      NM:i:0  MD:Z:150        MC:Z:149M       AS:i:150        XS:i
:0
SRR22300007.788016      83      NC_001133.9     154488  60      150M    =       154393  -245    AAAAGTTTCCTCAATATCAATCAAGACACCAAAATCTTGCGCACCAGCGTCTCTGTAATT
TTGTAATTGGTCATCGTAAATTTGTGCCTTGTAAGTTGCTTGGAACAAAGTACCTTTAGACAATTCCTTGTGGAAGCACTTACGTTCAGC      FJJFJJJJFAFJJFJJJ<JAA<F<)7J7FA-<A<F-JF<AFA<JAJFJJJJJJJJJFJJJ
JJJJJFJJJJFJJJJFJJJFJJJJJFJJJJJJJJJJJJFFFJFJJAJJJJJJFJJJJJJJJJJJJJJJJJFJFFFFJ<JJJJJJJFAAAA      NM:i:0  MD:Z:150        MC:Z:150M       AS:i:150        XS:i
:0
SRR22300007.3368427     83      NC_001133.9     154488  60      150M    =       154427  -211    AAAAGTTTCCTCAATATCAATCAAGACACCAAAATCTTGCGCACCAGCGTCTCTGTAATT
TTGTAATTGGTCATCGTAAATTTGTGCCTTGTAAGTTGCTTGGAACAAAGTACCTTTAGACAATTCCTTGTGGAAGCACTTACGTTCAGC      JJFAJJJFFJJJJJJJJJJJFJFJJFJFJJJJJJJJFJJJJJJFAJJJJJJJJJJJJJ
JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJFJJJJJJJJJJJJJJJJJJJJFFFAA      NM:i:0  MD:Z:150        MC:Z:150M       AS:i:150        XS:i
:0
SRR22300007.2293325     163     NC_001133.9     154489  60      94M     =       154489  94      AAAGTTTCCTCAATATCAATCAAGACACCAAAATCTTGCGCACCAGCGTCTCTGTAATTT
TGTAATTGGTCATCGTAAATTTGTGCCTTGTAAG      AAFFFJJJFJJFJJ<JJJJJJJJJJFJAAFFAJ<JJJJJJJJJJFJJFAFJJJJJJ<FJJJJJ--FFJJFJJJJJJ7J7AJJJJFJJJJJJJJJ  NM:i:0  MD:Z:94 MC:Z
:94M    AS:i:94 XS:i:0
SRR22300007.1717285     83      NC_001133.9     154489  60      150M    =       154326  -313    AAAGTTTCCTCAATATCAATCAAGACACCAAAATCTTGCGCACCAGCGTCTCTGTAATTT
TGTAATTGGTCATCGTAAATTTGTGCCTTGTAAGTTGCTTGGAACAAAGTACCTTTAGACAATTCCTTGTGGAAGCACTTACGTTCAGCA      JJJJJJAAFJJJJJJJJJJJJJJJJJFFJJJJJJJJJJJJFFJJ<JJJJJJJJJJJJJJ
JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJFJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJFFFAA      NM:i:0  MD:Z:150        MC:Z:150M       AS:i:150        XS:i
:0
```

samtools flagstat SRR22300007_sorted.bam

```
ibab@LAPTOP-BVSTVK8Q:~/NGS/variant_calling/5_Alignment$ samtools flagstat SRR22300007_sorted.bam
8007148 + 0 in total (QC-passed reads + QC-failed reads)
7981872 + 0 primary
0 + 0 secondary
25276 + 0 supplementary
0 + 0 duplicates
0 + 0 primary duplicates
7956653 + 0 mapped (99.37% : N/A)
7931377 + 0 primary mapped (99.37% : N/A)
7981872 + 0 paired in sequencing
3990936 + 0 read1
3990936 + 0 read2
7859762 + 0 properly paired (98.47% : N/A)
7924198 + 0 with itself and mate mapped
7179 + 0 singletons (0.09% : N/A)
51664 + 0 with mate mapped to a different chr
40526 + 0 with mate mapped to a different chr (mapQ>=5)
```

bamtools stats -in SRR22300007_sorted.bam -insert

```
ibab@LAPTOP-BVSTVK8Q:~/NGS/variant_calling/5_Alignment$ bamtools stats -in SRR22300007_sorted.bam -insert

*********************************************
Stats for BAM file(s):
*********************************************

Total reads:       8007148
Mapped reads:      7956653      (99.3694%)
Forward strand:    4027369      (50.2972%)
Reverse strand:    3979779      (49.7028%)
Failed QC:         0    (0%)
Duplicates:        0    (0%)
Paired-end reads:  8007148      (100%)
'Proper-pairs':    7875102      (98.3509%)
Both pairs mapped: 7948964      (99.2733%)
Read 1:            4003184
Read 2:            4003964
Singletons:        7689 (0.0960267%)
Average insert size (absolute value): 426.419
Median insert size (absolute value): 218
```

Since, Picard tool was not working in my laptop, so I switched to IBAB's PC to proceed with the remaining following steps.

Step-6: Mark duplicates

picard MarkDuplicates I=SRR22300007_sorted.bam O=SRR22300007_markdup.bam M=marked_dup_metrics.txt

```
(base) ibab@IBAB-MScBDB2-Comp007:~/NGS/variant_calling/5_Alignment$ picard MarkDuplicates I=SRR22300007_sorted.bam O=SRR22300007_markdup.bam M=marked_
dup_metrics.txt
INFO    2025-08-28 11:06:26     MarkDuplicates

********** NOTE: Picard's command line syntax is changing.
**********
********** For more information, please see:
********** https://github.com/broadinstitute/picard/wiki/Command-Line-Syntax-Transition-For-Users-(Pre-Transition)
**********
********** The command line looks like this in the new syntax:
**********
**********    MarkDuplicates -I SRR22300007_sorted.bam -O SRR22300007_markdup.bam -M marked_dup_metrics.txt
**********
```

```
(base) ibab@IBAB-MScBDB2-Comp007:~/NGS/variant_calling/5_Alignment$ ls -lh
total 4.5G
-rw-rw-r-- 1 ibab ibab 3.3K Aug 28 11:07 marked_dup_metrics.txt
-rw-rw-r-- 1 ibab ibab 719M Aug 28 10:49 SRR22300007.bam
-rw-rw-r-- 1 ibab ibab 446M Aug 28 11:07 SRR22300007_markdup.bam
-rw-rw-r-- 1 ibab ibab 3.0G Aug 28 10:48 SRR22300007.sam
```

cat marked_dup_metrics.txt

```
(base) ibab@LAPTOP-BVSTVK8Q:~/NGS/variant_calling/5_Alignment$ cat marked_dup_metrics.txt
## htsjdk.samtools.metrics.StringHeader
# MarkDuplicates INPUT=[SRR22300007_sorted.bam] OUTPUT=SRR22300007_markdup.bam METRICS_FILE=marked_dup_metrics.txt    MAX_SEQUENCES_FOR_DISK_READ_ENDS_MAP=5
0000 MAX_FILE_HANDLES_FOR_READ_ENDS_MAP=8000 SORTING_COLLECTION_SIZE_RATIO=0.25 TAG_DUPLICATE_SET_MEMBERS=false REMOVE_SEQUENCING_DUPLICATES=false TAGGING_P
OLICY=DontTag CLEAR_DT=true DUPLEX_UMI=false ADD_PG_TAG_TO_READS=true REMOVE_DUPLICATES=false ASSUME_SORTED=false DUPLICATE_SCORING_STRATEGY=SUM_OF_BASE_QUA
LITIES PROGRAM_RECORD_ID=MarkDuplicates PROGRAM_GROUP_NAME=MarkDuplicates READ_NAME_REGEX=<optimized capture of last three ':' separated fields as numeric v
alues> OPTICAL_DUPLICATE_PIXEL_DISTANCE=100 MAX_OPTICAL_DUPLICATE_SET_SIZE=300000 VERBOSITY=INFO QUIET=false VALIDATION_STRINGENCY=STRICT COMPRESSION_LEVEL=
5 MAX_RECORDS_IN_RAM=500000 CREATE_INDEX=false CREATE_MD5_FILE=false GA4GH_CLIENT_SECRETS=client_secrets.json USE_JDK_DEFLATER=false USE_JDK_INFLATER=false
## htsjdk.samtools.metrics.StringHeader
# Started on: Thu Aug 28 11:06:26 IST 2025

## METRICS CLASS        picard.sam.DuplicationMetrics
LIBRARY UNPAIRED_READS_EXAMINED READ_PAIRS_EXAMINED     SECONDARY_OR_SUPPLEMENTARY_RDS  UNMAPPED_READS  UNPAIRED_READ_DUPLICATES        READ_PAIR_DUPLICATES
READ_PAIR_OPTICAL_DUPLICATES    PERCENT_DUPLICATION     ESTIMATED_LIBRARY_SIZE
Unknown Library 7176    3962068 25263   50492   3246    256185  0       0.06501 29302547

## HISTOGRAM    java.lang.Double
BIN     CoverageMult    all_sets        non_optical_sets
1.0     1       3466067 3466067
2.0     1.87353 224731  224731
3.0     2.636586        13952   13952
4.0     3.303138        1007    1007
5.0     3.885391        110     110
6.0     4.394007        12      12
7.0     4.838298        1       1
8.0     5.2264  2       2
9.0     5.565419        0       0
10.0    5.861563        1       1
11.0    6.120253        0       0
12.0    6.346227        0       0
```

samtools view SRR22300007_markdup.bam

```
SRR22300007.129176      83      NC_001133.9     129692  60      150M    =       129531  -311    GGACGAAGGTGATAATGAGGAGGAAAACAACGAGGAGGATAATGAAAACGAAAACGAGCA
TACAGCACCGCCTGCGCTGGTGATGCCCTCCCCCATCGAAATGGAGGAACAGAGGATGACTGCGCTGAAGGAAATCACCGACATCGAGTA      JJF7FJJJJAJJJJJJJJJJJJJJJJJJJJFFJFJJJJJJJJJJJJJJJFJJJJFJJJJJJF
JJJFJJJJJJJJFJJJJJJJJJJJJJJJFJJJJJJFJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJFFFAA      MC:Z:150M       MD:Z:150        PG:Z:MarkDuplicates     NM:i
:0      AS:i:150        XS:i:0
SRR22300007.3195770     83      NC_001133.9     129692  60      150M    =       129618  -224    GGACGAAGGTGATAATGAGGAGGAAAACAACGAGGAGGATAATGAAAACGAAAACGAGCA
TACAGCACCGCCTGCGCTGGTGATGCCCTCCCCCATCGAAATGGAGGAACAGAGGATGACTGCGCTGAAGGAAATCACCGACATCGAGTA      JJAAF<JJA<FJJFJFJAF<JFFF<AJFFFJJ7<AFJAA-JJJJA---JJA<FJJAJFJJ
JJJJJJJJJJJJJJJJJJJJJAA7JJJA<JJJJFFJJJJFFJJFJFJFFJJJJJFFJJJJJJJJJJJJJJJJJJJFFJJJAJAJJFFAAA      MC:Z:150M       MD:Z:150        PG:Z:MarkDuplicates     NM:i
:0      AS:i:150        XS:i:0
```

Step-7: Add read group ID

picard AddOrReplaceReadGroups I=SRR22300007_markdup.bam O=SRR22300007_grpadded.bam RGID=4 RGLB=LIB1 RGPL=illumina RGPU=unit1 RGSM=sample_name

```
(base) ibab@IBAB-MScBDB2-Comp007:~/NGS/variant_calling/5_Alignment$ picard AddOrReplaceReadGroups I=SRR22300007_markdup.bam O=SRR22300007_grpadded.bam
 RGID=4 RGLB=LIB1 RGPL=illumina RGPU=unit1 RGSM=sample_name
INFO    2025-08-28 11:48:57     AddOrReplaceReadGroups

********** NOTE: Picard's command line syntax is changing.
**********
********** For more information, please see:
********** https://github.com/broadinstitute/picard/wiki/Command-Line-Syntax-Transition-For-Users-(Pre-Transition)
**********
********** The command line looks like this in the new syntax:
**********
**********    AddOrReplaceReadGroups -I SRR22300007_markdup.bam -O SRR22300007_grpadded.bam -RGID 4 -RGLB LIB1 -RGPL illumina -RGPU unit1 -RGSM sample
_name
**********
```

```
(base) ibab@IBAB-MScBDB2-Comp007:~/NGS/variant_calling/5_Alignment$ ls -lh
total 5.0G
-rw-rw-r-- 1 ibab ibab 3.3K Aug 28 11:07 marked_dup_metrics.txt
-rw-rw-r-- 1 ibab ibab 719M Aug 28 10:49 SRR22300007.bam
-rw-rw-r-- 1 ibab ibab 447M Aug 28 11:49 SRR22300007_grpadded.bam
-rw-rw-r-- 1 ibab ibab 446M Aug 28 11:07 SRR22300007_markdup.bam
-rw-rw-r-- 1 ibab ibab 3.0G Aug 28 10:48 SRR22300007.sam
-rw-rw-r-- 1 ibab ibab 422M Aug 28 10:50 SRR22300007_sorted.bam
```

samtools index SRR22300007_grpadded.bam

```
(base) ibab@IBAB-MScBDB2-Comp007:~/NGS/variant_calling/5_Alignment$ samtools index SRR22300007_grpadded.bam
(base) ibab@IBAB-MScBDB2-Comp007:~/NGS/variant_calling/5_Alignment$ ls
marked_dup_metrics.txt  SRR22300007_grpadded.bam      SRR22300007_markdup.bam  SRR22300007_sorted.bam
SRR22300007.bam         SRR22300007_grpadded.bam.bai  SRR22300007.sam
```

samtools faidx /home/ibab/NGS/variant_calling/4_Reference/GCF_000146045.2_R64_genomic.fna

```
(base) ibab@IBAB-MScBDB2-Comp007:~/NGS/variant_calling/6_variant_call$ samtools faidx /home/ibab/NGS/variant_calling/4_Reference/GCF_000146045.2_R64_g
enomic.fna
(base) ibab@IBAB-MScBDB2-Comp007:~/NGS/variant_calling/6_variant_call$ ls
(base) ibab@IBAB-MScBDB2-Comp007:~/NGS/variant_calling/6_variant_call$ cd ..
(base) ibab@IBAB-MScBDB2-Comp007:~/NGS/variant_calling$ cd 4_Reference/
(base) ibab@IBAB-MScBDB2-Comp007:~/NGS/variant_calling/4_Reference$ ls
GCF_000146045.2_R64_genomic.fna      GCF_000146045.2_R64_genomic.fna.ann  GCF_000146045.2_R64_genomic.fna.fai  GCF_000146045.2_R64_genomic.fna.sa
GCF_000146045.2_R64_genomic.fna.amb  GCF_000146045.2_R64_genomic.fna.bwt  GCF_000146045.2_R64_genomic.fna.pac
```

cd 4_Reference

gatk CreateSequenceDictionary -R GCF_000146045.2_R64_genomic.fna -O GCF_000146045.2_R64_genomic.dict

```
(base) ibab@IBAB-MScBDB2-Comp007:~/NGS/variant_calling/4_Reference$ gatk CreateSequenceDictionary -R GCF_000146045.2_R64_genomic.fna -O GCF_000146045.
2_R64_genomic.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.compre
ssion_level=2 -jar /home/ibab/miniconda3/share/gatk4-4.3.0.0-0/gatk-package-4.3.0.0-local.jar CreateSequenceDictionary -R GCF_000146045.2_R64_genomic.
fna -O GCF_000146045.2_R64_genomic.dict
13:08:12.158 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 28 13:08:12 IST 2025] CreateSequenceDictionary --OUTPUT GCF_000146045.2_R64_genomic.dict --REFERENCE GCF_000146045.2_R64_genomic.fna --TRUNCA
TE_NAMES_AT_WHITESPACE true --NUM_SEQUENCES 2147483647 --VERBOSITY INFO --QUIET false --VALIDATION_STRINGENCY STRICT --COMPRESSION_LEVEL 2 --MAX_RECOR
DS_IN_RAM 500000 --CREATE_INDEX false --CREATE_MD5_FILE false --GA4GH_CLIENT_SECRETS client_secrets.json --help false --version false --showHidden fal
se --USE_JDK_DEFLATER false --USE_JDK_INFLATER false
[Thu Aug 28 13:08:12 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 28 13:08:12 IST 2025] picard.sam.CreateSequenceDictionary done. Elapsed time: 0.01 minutes.
Runtime.totalMemory()=257949696
Tool returned:
0
```

```
(base) ibab@IBAB-MScBDB2-Comp007:~/NGS/variant_calling/4_Reference$ ls -lh
total 33M
-rw-rw-r-- 1 ibab ibab 2.5K Aug 28 13:08 GCF_000146045.2_R64_genomic.dict
-rw-rw-r-- 1 ibab ibab  12M Aug 28 08:39 GCF_000146045.2_R64_genomic.fna
-rw-rw-r-- 1 ibab ibab   14 Aug 28 10:28 GCF_000146045.2_R64_genomic.fna.amb
-rw-rw-r-- 1 ibab ibab 1.6K Aug 28 10:28 GCF_000146045.2_R64_genomic.fna.ann
-rw-rw-r-- 1 ibab ibab  12M Aug 28 10:28 GCF_000146045.2_R64_genomic.fna.bwt
-rw-rw-r-- 1 ibab ibab  562 Aug 28 13:05 GCF_000146045.2_R64_genomic.fna.fai
-rw-rw-r-- 1 ibab ibab 2.9M Aug 28 10:28 GCF_000146045.2_R64_genomic.fna.pac
-rw-rw-r-- 1 ibab ibab 5.8M Aug 28 10:28 GCF_000146045.2_R64_genomic.fna.sa
```

Step-8: Call the first set of variants using HaplotypeCaller

cd 6_variant_call

gatk HaplotypeCaller -I /home/ibab/NGS/variant_calling/5_Alignment/SRR22300007_grpadded.bam -R

/home/ibab/NGS/variant_calling/4_Reference/GCF_000146045.2_R64_genomic.fna -O SRR22300007_raw_variants.vcf

```
(base) ibab@IBAB-MScBDB2-Comp007:~/NGS/variant_calling/6_variant_call$ gatk HaplotypeCaller -I /home/ibab/NGS/variant_calling/5_Alignment/SRR22300007_
grpadded.bam -R /home/ibab/NGS/variant_calling/4_Reference/GCF_000146045.2_R64_genomic.fna -O SRR22300007_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.compre
ssion_level=2 -jar /home/ibab/miniconda3/share/gatk4-4.3.0.0-0/gatk-package-4.3.0.0-local.jar HaplotypeCaller -I /home/ibab/NGS/variant_calling/5_Alig
nment/SRR22300007_grpadded.bam -R /home/ibab/NGS/variant_calling/4_Reference/GCF_000146045.2_R64_genomic.fna -O SRR22300007_raw_variants.vcf
13:15:15.300 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
13:15:15.366 INFO  HaplotypeCaller - ------------------------------------------------------------
13:15:15.366 INFO  HaplotypeCaller - The Genome Analysis Toolkit (GATK) v4.3.0.0
```

```
(base) ibab@IBAB-MScBDB2-Comp007:~/NGS/variant_calling/6_variant_call$ ls -lh
total 2.3M
-rw-rw-r-- 1 ibab ibab 2.3M Aug 28 13:19 SRR22300007_raw_variants.vcf
-rw-rw-r-- 1 ibab ibab  16K Aug 28 13:19 SRR22300007_raw_variants.vcf.idx
```

less SRR22300007_raw_variants.vcf

```
##source=HaplotypeCaller
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  sample_name
NC_001133.9     1179    .       C       T       346.64  .       AC=1;AF=0.500;AN=2;BaseQRankSum=0.137;DP=93;ExcessHet=0.0000;FS=5.911;MLEAC=1;MLEAF=0.500;MQ
=56.98;MQRankSum=-8.169;QD=3.77;ReadPosRankSum=1.260;SOR=1.522  GT:AD:DP:GQ:PL  0/1:78,14:92:99:354,0,2453
NC_001133.9     1193    .       A       T       507.64  .       AC=1;AF=0.500;AN=2;BaseQRankSum=-0.946;DP=101;ExcessHet=0.0000;FS=2.184;MLEAC=1;MLEAF=0.500;
MQ=56.47;MQRankSum=-7.206;QD=5.18;ReadPosRankSum=0.959;SOR=1.048        GT:AD:DP:GQ:PL  0/1:80,18:98:99:515,0,3223
NC_001133.9     1197    .       G       T       596.64  .       AC=1;AF=0.500;AN=2;BaseQRankSum=-2.051;DP=102;ExcessHet=0.0000;FS=0.921;MLEAC=1;MLEAF=0.500;
MQ=56.26;MQRankSum=-6.503;QD=5.97;ReadPosRankSum=-0.762;SOR=0.841       GT:AD:DP:GQ:PL  0/1:79,21:100:99:604,0,3182
NC_001133.9     1217    .       C       T       859.64  .       AC=1;AF=0.500;AN=2;BaseQRankSum=-2.696;DP=110;ExcessHet=0.0000;FS=2.906;MLEAC=1;MLEAF=0.500;
MQ=55.61;MQRankSum=-6.539;QD=8.11;ReadPosRankSum=-2.730;SOR=1.080       GT:AD:DP:GQ:PL  0/1:77,29:106:99:867,0,2442
NC_001133.9     1230    .       T       A       1008.64 .       AC=1;AF=0.500;AN=2;BaseQRankSum=-1.885;DP=114;ExcessHet=0.0000;FS=1.753;MLEAC=1;MLEAF=0.500;
MQ=55.19;MQRankSum=-5.269;QD=9.09;ReadPosRankSum=-2.562;SOR=0.985       GT:AD:DP:GQ:PL  0/1:81,30:111:99:1016,0,3260
```

Step-9: Base Quality Score Recalibration (BQSR)

gatk BaseRecalibrator -I /home/ibab/NGS/variant_calling/5_Alignment/SRR22300007_grpadded.bam -R /home/ibab/NGS/variant_calling/4_Reference/GCF_000146045.2_R64_genomic.fna --known-sites SRR22300007_raw_variants.vcf -O recalibration_table.table

```
(base) ibab@IBAB-MScBDB2-Comp007:~/NGS/variant_calling/6_variant_call$ gatk BaseRecalibrator -I /home/ibab/NGS/variant_calling/5_Alignment/SRR22300007
_grpadded.bam -R /home/ibab/NGS/variant_calling/4_Reference/GCF_000146045.2_R64_genomic.fna --known-sites SRR22300007_raw_variants.vcf -O recalibratio
n_table.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 /home/ibab/NGS/variant_calling/5_Ali
gnment/SRR22300007_grpadded.bam -R /home/ibab/NGS/variant_calling/4_Reference/GCF_000146045.2_R64_genomic.fna --known-sites SRR22300007_raw_variants.v
cf -O recalibration_table.table
13:35:39.879 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
13:35:39.943 INFO  BaseRecalibrator - ------------------------------------------------------------
```

less recalibration_table.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
           9         0              93
          10    686943              10
          11         0              11
          12  18199301              12
          13         0              93
          14         0              93
:
```

```
#:GATKTable:6:1:%s:%s:%.4f:%.4f:%d:%.2f:;
#:GATKTable:RecalTable0:
ReadGroup  EventType  EmpiricalQuality  EstimatedQReported  Observations  Errors
unit1      M                 27.0000             27.6856     869818354  1601665.00

#:GATKTable:6:7:%s:%d:%s:%.4f:%d:%.2f:;
#:GATKTable:RecalTable1:
ReadGroup  QualityScore  EventType  EmpiricalQuality  Observations  Errors
unit1                8  M                  10.0000        686943    68192.00
unit1               12  M                  12.0000      18199301  1211217.00
unit1               22  M                  23.0000      12824770    65656.00
unit1               27  M                  27.0000      20280435    43334.00
unit1               32  M                  31.0000      47542768    34064.00
unit1               37  M                  34.0000     102644943    39666.00
unit1               41  M                  37.0000     667639194   139536.00

#:GATKTable:8:2202:%s:%d:%s:%s:%s:%.4f:%d:%.2f:;
#:GATKTable:RecalTable2:
ReadGroup  QualityScore  CovariateValue  CovariateName  EventType  EmpiricalQuality  Observations  Errors
unit1                8  -1              Cycle          M                  8.0000            11       0.00
unit1                8  -10             Cycle          M                  8.0000            16       0.00
unit1                8  -100            Cycle          M                  9.0000           618      66.00
unit1                8  -101            Cycle          M                  9.0000           652      50.00
unit1                8  -102            Cycle          M                  9.0000           608      63.00
unit1                8  -103            Cycle          M                  9.0000           623      65.00
unit1                8  -104            Cycle          M                  9.0000           681      71.00
unit1                8  -105            Cycle          M                  9.0000           814      74.00
unit1                8  -106            Cycle          M                 10.0000          1031      85.00
unit1                8  -107            Cycle          M                  9.0000          1174     141.00
unit1                8  -108            Cycle          M                  9.0000          1705     174.00
unit1                8  -109            Cycle          M                  9.0000          2813     305.00
```

Step-10: Apply the corrected base quality scores to the reads

gatk ApplyBQSR -I /home/ibab/NGS/variant_calling/5_Alignment/SRR22300007_grpadded.bam -R /home/ibab/NGS/variant_calling/4_Reference/GCF_000146045.2_R64_genomic.fna -bqsr recalibration_table.table -O SRR22300007_recalibrated_reads.bam

```
(base) ibab@IBAB-MScBDB2-Comp007:~/NGS/variant_calling/6_variant_call$ gatk ApplyBQSR -I /home/ibab/NGS/variant_calling/5_Alignment/SRR22300007_grpadd
ed.bam -R /home/ibab/NGS/variant_calling/4_Reference/GCF_000146045.2_R64_genomic.fna -bqsr recalibration_table.table -O SRR22300007_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 /home/ibab/NGS/variant_calling/5_Alignment/
SRR22300007_grpadded.bam -R /home/ibab/NGS/variant_calling/4_Reference/GCF_000146045.2_R64_genomic.fna -bqsr recalibration_table.table -O SRR22300007_
recalibrated_reads.bam
13:38:57.572 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
13:38:57.637 INFO  ApplyBQSR - ------------------------------------------------------------
```

```
ibab@LAPTOP-BVSTVK8Q:~/NGS/variant_calling/6_variant_call$ ls -lh
total 757M
-rw-r--r-- 1 ibab ibab 2.3M Aug 28 14:55 SRR22300007_raw_variants.vcf
-rw-r--r-- 1 ibab ibab  16K Aug 28 14:53 SRR22300007_raw_variants.vcf.idx
-rw-r--r-- 1 ibab ibab  39K Aug 28 15:31 SRR22300007_recalibrated_reads.bai
-rw-r--r-- 1 ibab ibab 754M Aug 28 15:31 SRR22300007_recalibrated_reads.bam
-rw-r--r-- 1 ibab ibab 242K Aug 28 15:23 recalibration_table.table
```

samtools view SRR22300007_recalibrated_reads.bam

```
ACAAGTGGCTGATCTTTTGTTAACTATCTGACATGTCCTCGCT     BGEDCD+;<,B;DB@F+-A;-BCC6-BD-)=>D-)*8=A?G++-*(-=?+-+-+(-)3).+75@G-:+(::DB+:?<<@C@+8>C@BG.EDG+@G.6+.6D:F MC:Z
:103M   MD:Z:51T29T21   PG:Z:MarkDuplicates     RG:Z:4  NM:i:2  AS:i:93 XS:i:0
SRR22300007.3711027     83      NC_001133.9     104945  60      103M    =       104945  -103    CTGCATATGATGACGGGTAGTCTACGCCGTTCATTAGTGTGGATACTAGAGTTACTTTTG
ACAAGTGGCTGATCTTTTGTTACTATCTGACATGTCCTCGCT     GB+EFCD-GEE==EECGDFGDEB=1@=@+:CEFJ8=@=DC@=FADDDF=0D/.DGJJJED?A<B?EEGGED=;1FBFDGIG<DGBBECCB4DCB@?BE3616C MC:Z
:103M   MD:Z:28T23G50   PG:Z:MarkDuplicates     RG:Z:4  NM:i:2  AS:i:93 XS:i:0
SRR22300007.3237823     99      NC_001133.9     104946  60      84M     =       104946  84      TGCATATGATGACGGGTAGTCTACGCCTTTCATTAGTGTGGATACTAGAGTGACTTTTGA
CAAGTGGCTGATCTTTTGTTTACT        CHA>GCEEDEGEGDFGBDGDGEDFEFEEEEDDEEDBDEDGGEEDGEBGEDDGE;B<@EEDGCCDDEGGEA?FGE7EBFDEE@;E    MC:Z:84M        MD:Z:84 PG:Z:MarkDup
licates RG:Z:4  NM:i:0  AS:i:84 XS:i:0
SRR22300007.3237823     147     NC_001133.9     104946  60      84M     =       104946  -84     TGCATATGATGACGGGTAGTCTACGCCTTTCATTAGTGTGGATACTAGAGTGACTTTTGA
CAAGTGGCTGATCTTTTGTTTACT        EGGFDFEGFEGDEEEGDEGEGDDEGGGJJEGFJDEGEGDDGECDGCDFDGDFCFIJIDFCFDDGDDFFDFEDFIIICFHFBCEB    MC:Z:84M        MD:Z:84 PG:Z:MarkDup
licates RG:Z:4  NM:i:0  AS:i:84 XS:i:0
SRR22300007.1588645     147     NC_001133.9     104950  60      149M    =       104746  -353    TATGATGACGGGTAGTCTACGCCTTTCATTAGTGTGGATACTAGAGTGACTTTTGACAAG
TGGCTGATCTTTTGTTTACTATCTGACATGTCCTCGCTTTCATTATCAACTGAAGGTTCTTTCGCTATTCGGTGCTTCGAGTAAACATT      <EEHFEGDDECGCEDD9DD/*:C,KCGDGDEGE*DEEBCDGDED;DEGCGJGJED?E@EG
E>DGDDEEGJJJEGJJDCGDDE=ECDD9*2@CGEEGCJJEGFJCEEFDDB54)=EF>AFC+D?B-@CDB8*79=.ED@FDFB@*C@DEB       MC:Z:150M       MD:Z:149        PG:Z:MarkDuplicates     RG:Z
:4      NM:i:0  AS:i:149        XS:i:0
```

Step-11: Call the final set of variants

gatk HaplotypeCaller -R /home/ibab/NGS/variant_calling/4_Reference/GCF_000146045.2_R64_genomic.fna -I SRR22300007_recalibrated_reads.bam -O raw_final_variants.vcf

```
ibab@LAPTOP-BVSTVK8Q:~/NGS/variant_calling/6_variant_call$ /home/ibab/NGS/Packages/gatk-4.6.2.0/gatk HaplotypeCaller -R /home/ibab/NGS/variant_calling/4_Ref
erence/GCF_000146045.2_R64_genomic.fna -I SRR22300007_recalibrated_reads.bam -O raw_final_variants.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 /home/ibab/NGS/variant_calling/4_Reference/GCF_000146045
.2_R64_genomic.fna -I SRR22300007_recalibrated_reads.bam -O raw_final_variants.vcf
15:41:38.394 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
15:41:38.770 INFO  HaplotypeCaller - ------------------------------------------------------------
15:41:38.777 INFO  HaplotypeCaller - The Genome Analysis Toolkit (GATK) v4.6.2.0
```

less raw_final_variants.vcf

```
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  sample_name
NC_001133.9     1179    .       C       T       343.64  .       AC=1;AF=0.500;AN=2;BaseQRankSum=-0.305;DP=93;ExcessHet=0.0000;FS=5.911;MLEAC=1;MLEAF=0.500;
NC_001133.9     1193    .       A       T       507.64  .       AC=1;AF=0.500;AN=2;BaseQRankSum=-0.227;DP=101;ExcessHet=0.0000;FS=2.184;MLEAC=1;MLEAF=0.500
NC_001133.9     1197    .       G       T       596.64  .       AC=1;AF=0.500;AN=2;BaseQRankSum=-2.275;DP=102;ExcessHet=0.0000;FS=0.921;MLEAC=1;MLEAF=0.500
NC_001133.9     1217    .       C       T       846.64  .       AC=1;AF=0.500;AN=2;BaseQRankSum=-2.070;DP=110;ExcessHet=0.0000;FS=2.906;MLEAC=1;MLEAF=0.500
NC_001133.9     1230    .       T       A       1008.64 .       AC=1;AF=0.500;AN=2;BaseQRankSum=-1.490;DP=114;ExcessHet=0.0000;FS=1.753;MLEAC=1;MLEAF=0.500
NC_001133.9     1238    .       C       T       999.64  .       AC=1;AF=0.500;AN=2;BaseQRankSum=-3.191;DP=119;ExcessHet=0.0000;FS=1.742;MLEAC=1;MLEAF=0.500
NC_001133.9     1263    .       C       T       1146.64 .       AC=1;AF=0.500;AN=2;BaseQRankSum=-1.347;DP=118;ExcessHet=0.0000;FS=11.344;MLEAC=1;MLEAF=0.50
```

Step-12: Extract SNPs and Indels respectively

gatk SelectVariants -R /home/ibab/NGS/variant_calling/4_Reference/GCF_000146045.2_R64_genomic.fna -V raw_final_variants.vcf -select-type SNP -O SRR22300007.snps.vcf

gatk SelectVariants -R /home/ibab/NGS/variant_calling/4_Reference/GCF_000146045.2_R64_genomic.fna -V raw_final_variants.vcf -select-type INDEL -O SRR22300007.indels.vcf

```
ibab@LAPTOP-BVSTVK8Q:~/NGS/variant_calling/6_variant_call$ ls
SRR22300007.indels.vcf      SRR22300007_raw_variants.vcf        raw_final_variants.vcf
SRR22300007.indels.vcf.idx  SRR22300007_raw_variants.vcf.idx    raw_final_variants.vcf.idx
SRR22300007.snps.vcf        SRR22300007_recalibrated_reads.bai  recalibration_table.table
SRR22300007.snps.vcf.idx    SRR22300007_recalibrated_reads.bam
ibab@LAPTOP-BVSTVK8Q:~/NGS/variant_calling/6_variant_call$ less SRR22300007.snps.vcf | wc
  11434  114808 2067310
ibab@LAPTOP-BVSTVK8Q:~/NGS/variant_calling/6_variant_call$ less SRR22300007.indels.vcf | wc
   1511   15578  310197
```

Step-13: Hard Filtering

VQSR is only practical for humans because it needs large variant datasets and gold-standard training sets. Since yeast lacks such resources, we use hard filters instead.

gatk VariantFiltration -R /home/ibab/NGS/variant_calling/4_Reference/GCF_000146045.2_R64_genomic.fna -V SRR22300007.snps.vcf -O SRR22300007.snps.filtered.vcf --filter-name "LowQD" --filter-expression "QD < 2.0" --filter-name "HighFS" --filter-expression "FS > 60.0" --filter-name "LowMQ" --filter-expression "MQ < 40.0" --filter-name "LowMQRankSum" --filter-expression "MQRankSum < -12.5" --filter-name "LowReadPosRankSum" --filter-expression "ReadPosRankSum < -8.0"

gatk VariantFiltration -R /home/ibab/NGS/variant_calling/4_Reference/GCF_000146045.2_R64_genomic.fna -V SRR22300007.indels.vcf -O SRR22300007.indels.filtered.vcf --filter-name "LowQD" --filter-expression "QD < 2.0" --filter-name "HighFS" --filter-expression "FS > 200.0" --filter-name "LowReadPosRankSum" --filter-expression "ReadPosRankSum < -20.0"

```
ibab@LAPTOP-BVSTVK8Q:~/NGS/variant_calling/6_variant_call$ ls -lh
total 764M
-rw-r--r-- 1 ibab ibab 310K Aug 28 16:35 SRR22300007.indels.filtered.vcf
-rw-r--r-- 1 ibab ibab 1.7K Aug 28 16:35 SRR22300007.indels.filtered.vcf.idx
-rw-r--r-- 1 ibab ibab 303K Aug 28 16:04 SRR22300007.indels.vcf
-rw-r--r-- 1 ibab ibab 1.7K Aug 28 16:04 SRR22300007.indels.vcf.idx
-rw-r--r-- 1 ibab ibab 2.1M Aug 28 16:30 SRR22300007.snps.filtered.vcf
-rw-r--r-- 1 ibab ibab  13K Aug 28 16:30 SRR22300007.snps.filtered.vcf.idx
-rw-r--r-- 1 ibab ibab 2.0M Aug 28 15:59 SRR22300007.snps.vcf
-rw-r--r-- 1 ibab ibab  13K Aug 28 15:59 SRR22300007.snps.vcf.idx
-rw-r--r-- 1 ibab ibab 2.3M Aug 28 14:55 SRR22300007_raw_variants.vcf
-rw-r--r-- 1 ibab ibab  16K Aug 28 14:53 SRR22300007_raw_variants.vcf.idx
-rw-r--r-- 1 ibab ibab  39K Aug 28 15:31 SRR22300007_recalibrated_reads.bai
-rw-r--r-- 1 ibab ibab 754M Aug 28 15:31 SRR22300007_recalibrated_reads.bam
-rw-r--r-- 1 ibab ibab 2.3M Aug 28 15:48 raw_final_variants.vcf
-rw-r--r-- 1 ibab ibab  15K Aug 28 15:48 raw_final_variants.vcf.idx
-rw-r--r-- 1 ibab ibab 242K Aug 28 15:23 recalibration_table.table
```

bcftools view -f PASS SRR22300007.snps.filtered.vcf > SRR22300007.snps.pass.vcf

bcftools view -f PASS SRR22300007.indels.filtered.vcf > SRR22300007.indels.pass.vcf

```
ibab@LAPTOP-BVSTVK8Q:~/NGS/variant_calling/6_variant_call$ bcftools view -H -v snps SRR22300007.snps.pass.vcf | wc -l
10699
ibab@LAPTOP-BVSTVK8Q:~/NGS/variant_calling/6_variant_call$ bcftools view -H -v indels SRR22300007.indels.pass.vcf | wc -l
1360
```

Step-14: Annotate the filtered variants

ANNOVAR is mainly human-focused, while for yeast, snpEff is the most commonly used. Since, I didn't have snpEff installed before beginning the test, so I couldn't do annotation.

Fill in the table below:

Step	Result	
% reads mapped	**99.37**	
Mean coverage (×)	**90.93**	Mean coverage: samtools depth -a SRR22300007_sorted.bam | awk '{sum += $3; n++} END {print sum/n}'
Duplicate reads (%)	**0.06**	
Raw SNP count	**11389**	
Raw indel count	**1466**	grep -v "^#" SRR22300007.indels.vcf | wc -l
Filtered SNP count	**10699**	
Filtered indel count	**1360**	grep -v "^#" SRR22300007.indels.pass.vcf | wc -l
Filtering thresholds	**QD<2.0, FS>60.0, MQ<40.0 (for SNPs) and QD<2.0, FS>200.0 (for indels)**	
HIGH-impact variants (#)	_____	
MODERATE-impact variants (#)	_____	
LOW-impact variants (#)	_____	
Example missense variant (gene + change)	_____	