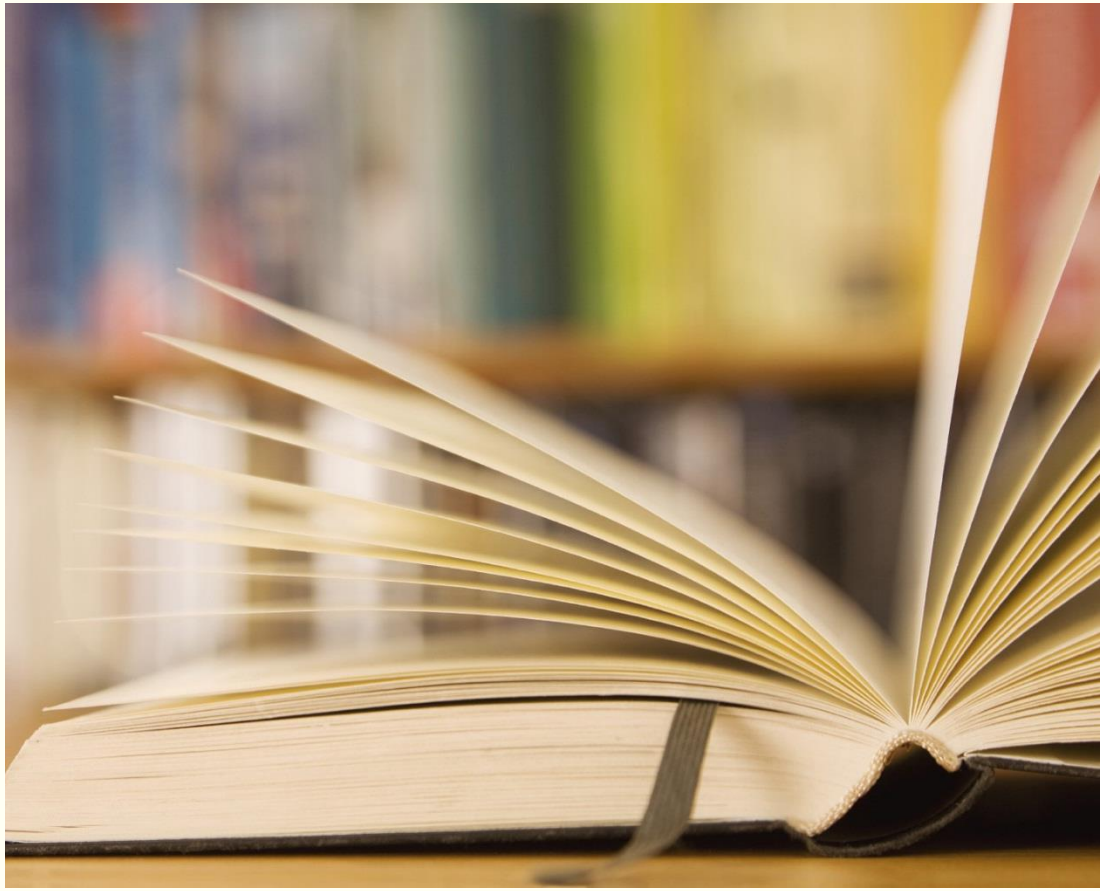
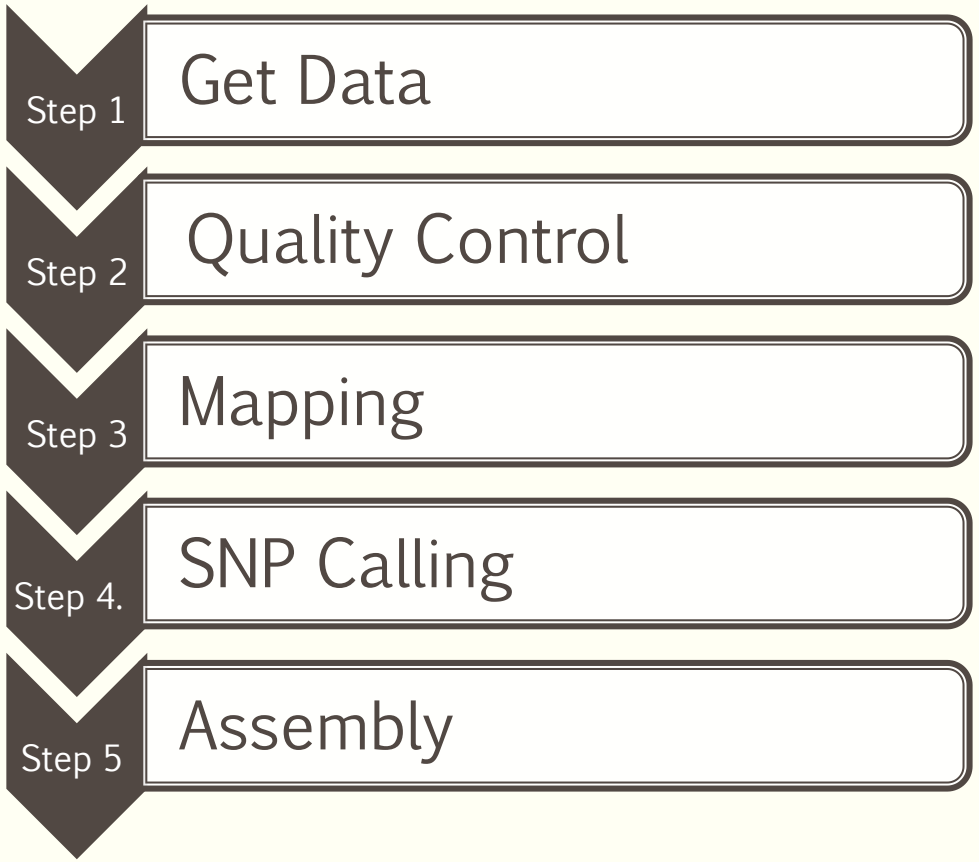
NGS DATA ANALYSIS OF BYMV (BEAN YELLOW MOSAIC VIRUS)

Meysam Zarei



Workflow

Input: SRX7118692



- **Data Retrieval:**
  The dataset was acquired from the designated source and verified to meet the project's criteria.
- **Customized Workflow:**
  A tailored analysis workflow was applied, as outlined in the subsequent slides.
- **Deliverables:**
  All results and this report are provided, with details included in the accompanying README file.

Workflow

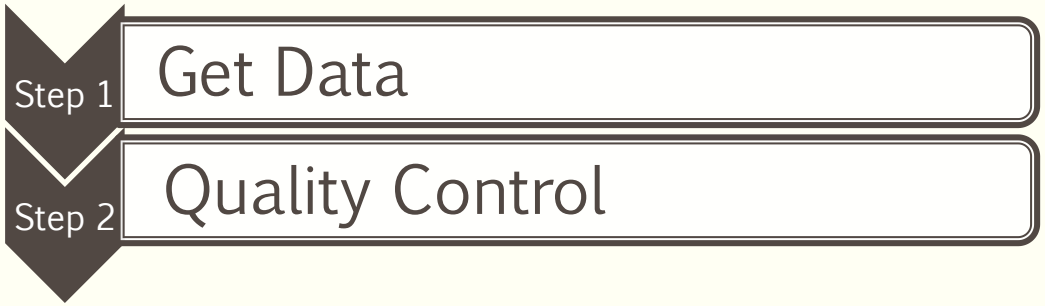
Step 1 Get Data

Input: SRX7118692

The data is available in SRA

Input: SRX7118692

Workflow

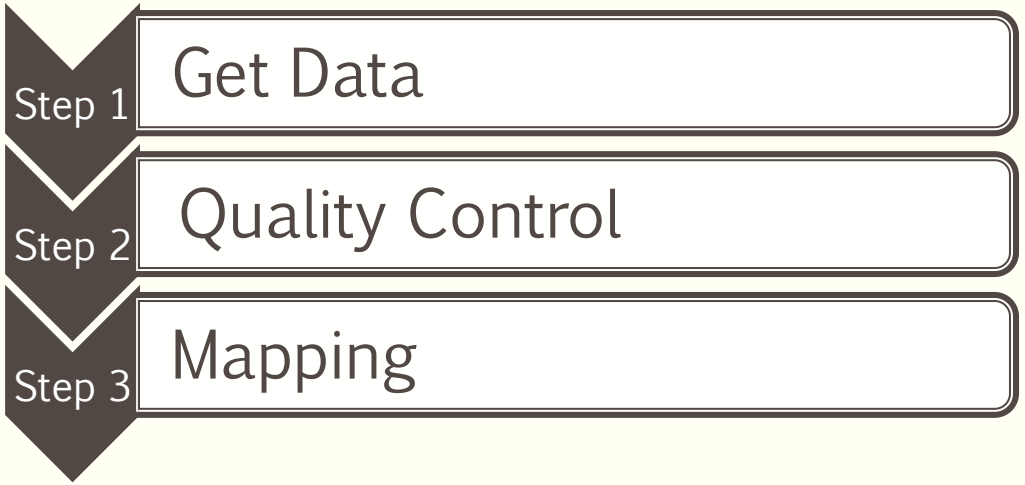


The raw sequencing data underwent quality assessment using FastQC (version 0.12.2) and MultiQC (version 1.27). Subsequent trimming was performed with Trim Galore! (version 0.6.10).

Trimming Parameters:

- ✓ **Minimum Read Length:** Reads shorter than 50 base pairs were discarded.
- ✓ **Quality Threshold:** Reads with an average quality score < 30 were removed.
- ✓ **Adapter Removal:** Sequencing adapters were identified and trimmed.

Workflow

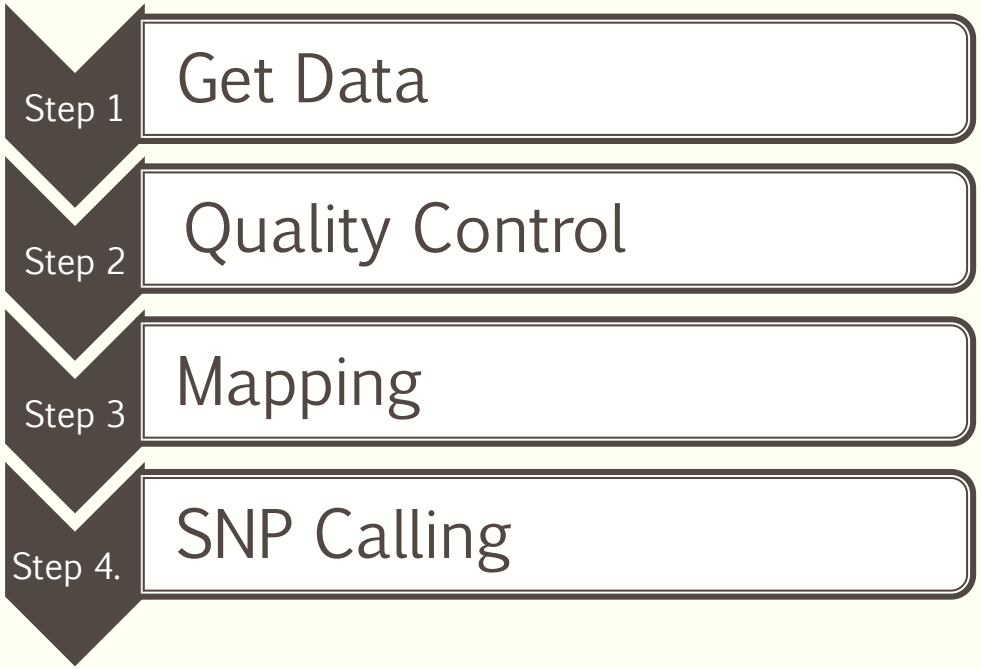


The sequencing reads were aligned to the BYMV reference genome (NC_003492.1) obtained from NCBI. Alignment was performed using BWA-MEM2 (version 2.2.1) and the resulting alignments were processed and saved in BAM format using SAMtools (version 1.21).

Quality Assessment

Alignment quality was assessed using SAMtools flagstat for summary statistics and IGV for detailed visual inspection of read alignments.

Workflow



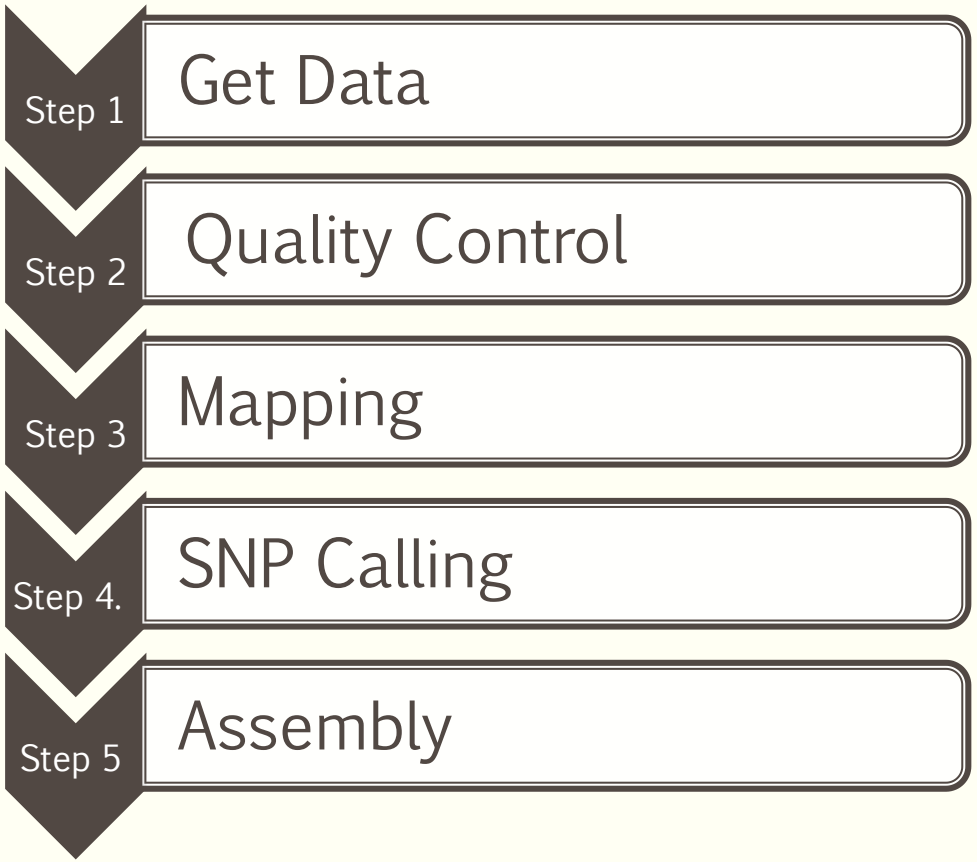
To identify variations relative to the reference sequence, we performed SNP calling using LoFreq (Version 2.1.5).

Initially, duplicate reads were removed with samtools markdup (version 1.21).

Lofreq Parameters:

- ✓ Minimum Coverage: 50
- ✓ Minimum Base Quality for any site: 30
- ✓ Minimum Base Quality for Alternate Allele: 30

Workflow



For the de novo assembly, we utilized velvet (version 1.2.10.2)

Post-assembly, we employed megaBLAST (version 1.2) to align the assembled sequences against known viral genomes, facilitating the identification of homologous regions and assessment of assembly accuracy. Additionally, we used QUAST (version 5.3.0) to evaluate the quality of the assembly.

Results: Pre-Processing
FastQC & MultiQC Results

Sample Name	% Dups	% GC	Average Read Length	% Failed	M Seqs
SRR10420664_forward	57.2 %	42 %	259 bp	30 %	1.8 M
SRR10420664_reverse	53.5 %	42 %	261 bp	20 %	1.8 M
Trim Galore_F_fastq	57.4 %	42 %	253 bp	30 %	1.7 M
Trim Galore_R_fastq	54.1 %	42 %	252 bp	10 %	1.7 M

Read Counts

Before Trimming :

Forward: 1,794,367
Reverse: 1,794,367
Total: 3,588,734 reads

After Trimming :

Forward: 1,742,089
Reverse: 1,742,089
Total: 3,484,178 reads

Reads Retained : ~97.08%

Reads Discarded : ~2.92%

Trimming slightly reduced the number of reads and average read length, while improving data quality—especially visible in the reduced failure rate of reverse reads (from 20% → 10%). GC content and duplication remained stable, indicating consistent base composition and sequencing complexity.

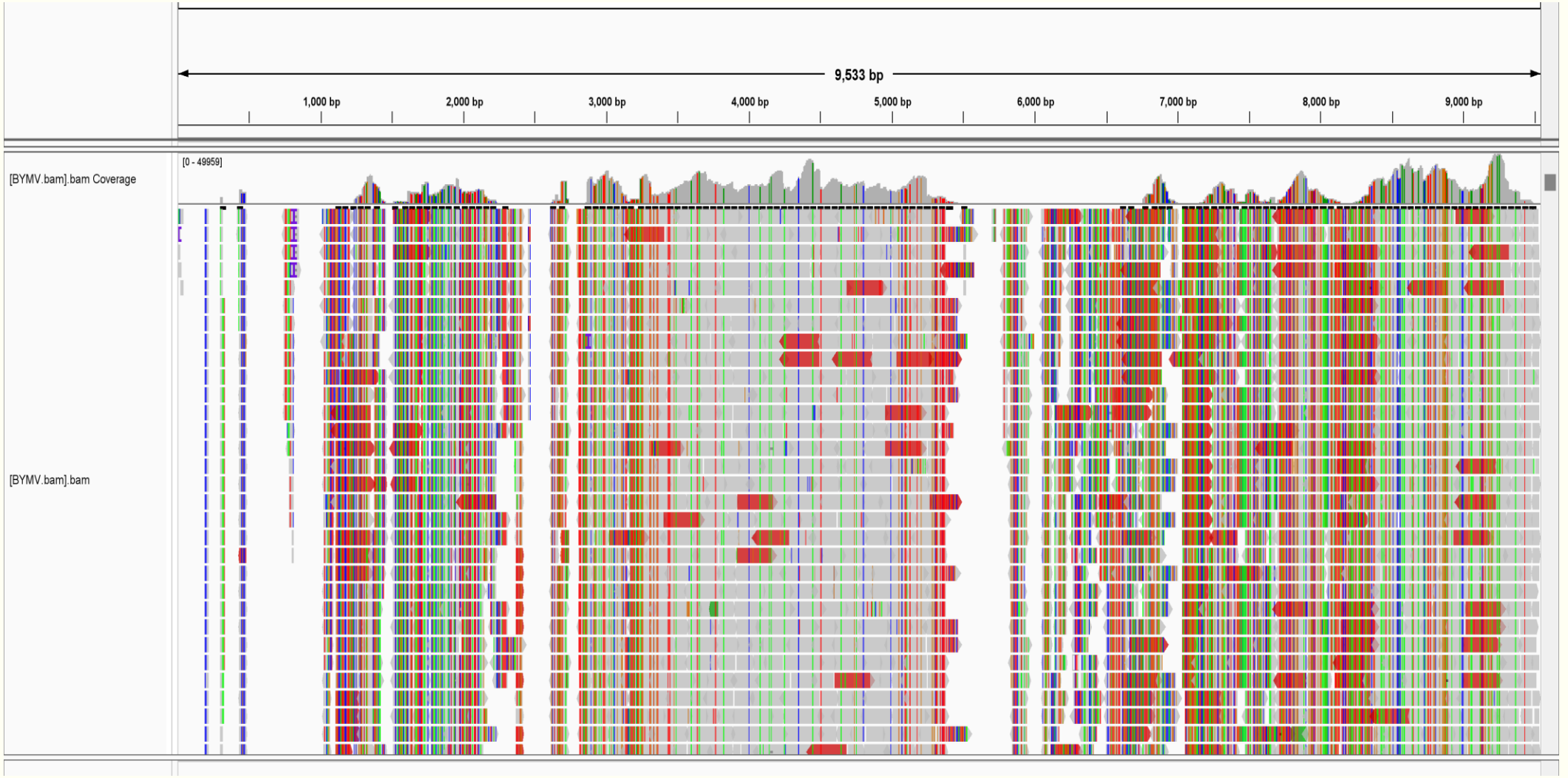
Pre-processing Overview

✅ **Improved Data Quality**
Quality control and trimming procedures successfully enhanced the overall dataset by eliminating low-quality regions and adapter contamination.

🔁 **High Duplication Rate**
A notable percentage of duplicate reads was identified, which may reflect PCR amplification artifacts or excessive sequencing depth.

🧬 **Possible Contamination**
Variations in GC content suggest the potential presence of non-target sequences, indicating a risk of sample contamination.

Results: Mapping
Initial alignment: IGV



Conclusion:
Initial mapping

Upon aligning the sequencing reads to the BYMV reference genome, 26.98% of the reads successfully mapped. Visualization in the IGV revealed uneven coverage across different regions, with some areas exhibiting gaps.

Recommended Actions:

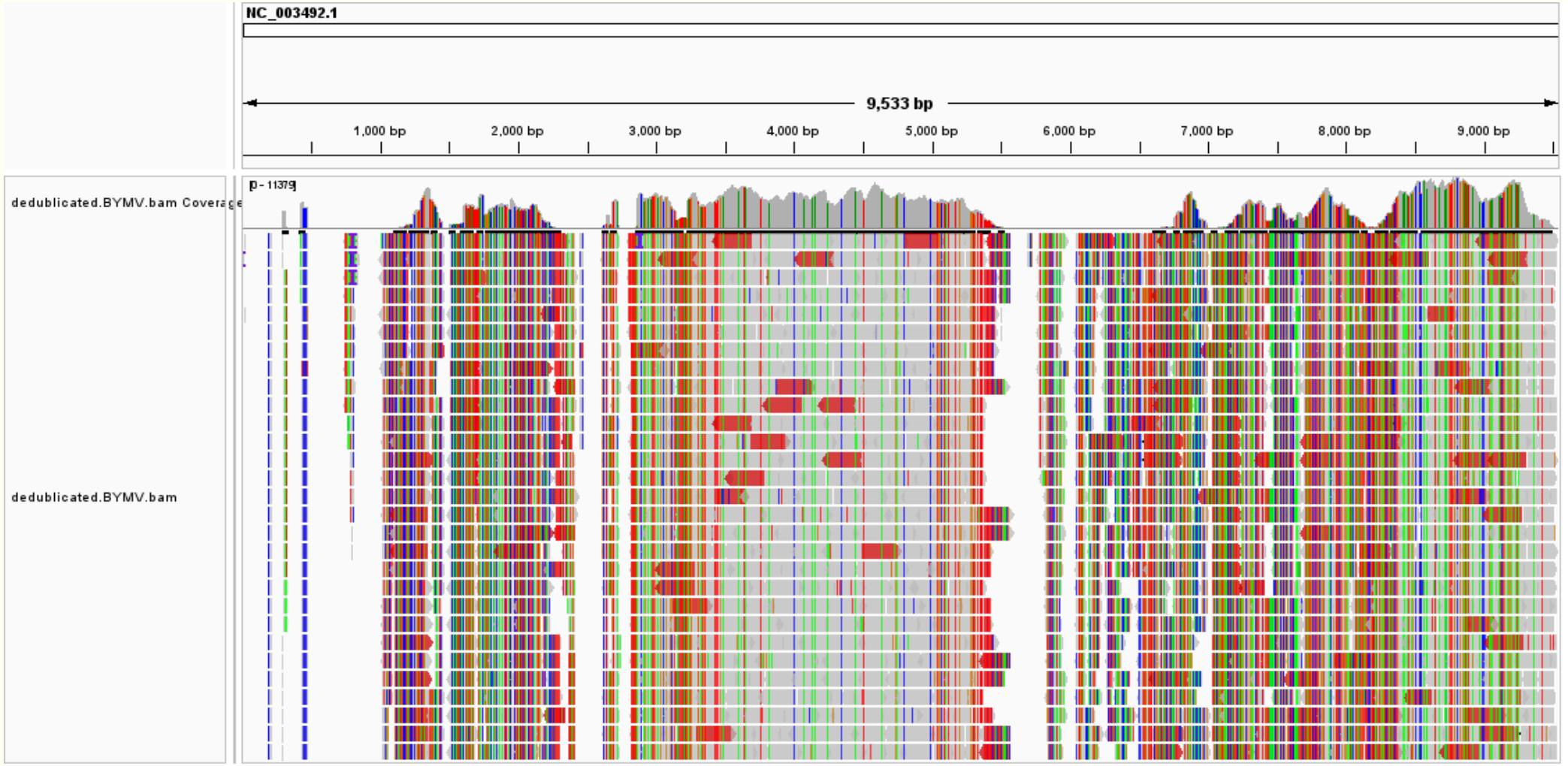
✓ **Remove Duplicate Reads**:
Eliminate duplicate reads to achieve a more accurate coverage profile.

✓ **Generate a Consensus Sequence**:
Create a consensus sequence from the mapped reads to serve as a more representative reference, enhancing mapping efficiency.

✓ **Perform *De Novo* Assembly**:
Assemble the viral genome without relying on the reference to identify novel sequences and structural variations.

✓ **Conduct SNP Calling**:
Identify SNPs and other variants to understand genetic differences from the reference genome.

Results: Mapping
Post duplicate removal: IGV



Results: Mapping
samtools markdup report

```
COMMAND: samtools markdup -@ 0 -r -m t -s -f
/data/jwd05e/main/085/098/85098240/outputs/dataset_7245c7da-2f7d-41ee-a281-8f02b7e818de.dat -O
BAM coordsort.sam /data/jwd05e/main/085/098/85098240/outputs/dataset_bcb669f5-6c77-4aee-ba7d-
3c5d0571b250.dat
READ: 3485044
WRITTEN: 2832207
EXCLUDED: 2545131
EXAMINED: 939913
PAIRED: 939496
SINGLE: 417
DUPLICATE PAIR: 652422
DUPLICATE SINGLE: 415
DUPLICATE PAIR OPTICAL: 0
DUPLICATE SINGLE OPTICAL: 0
DUPLICATE NON PRIMARY: 0
DUPLICATE NON PRIMARY OPTICAL: 0
DUPLICATE PRIMARY TOTAL: 652837
DUPLICATE TOTAL: 652837
ESTIMATED_LIBRARY_SIZE: 150102
```

Out of 3,485,044 total sequencing reads,

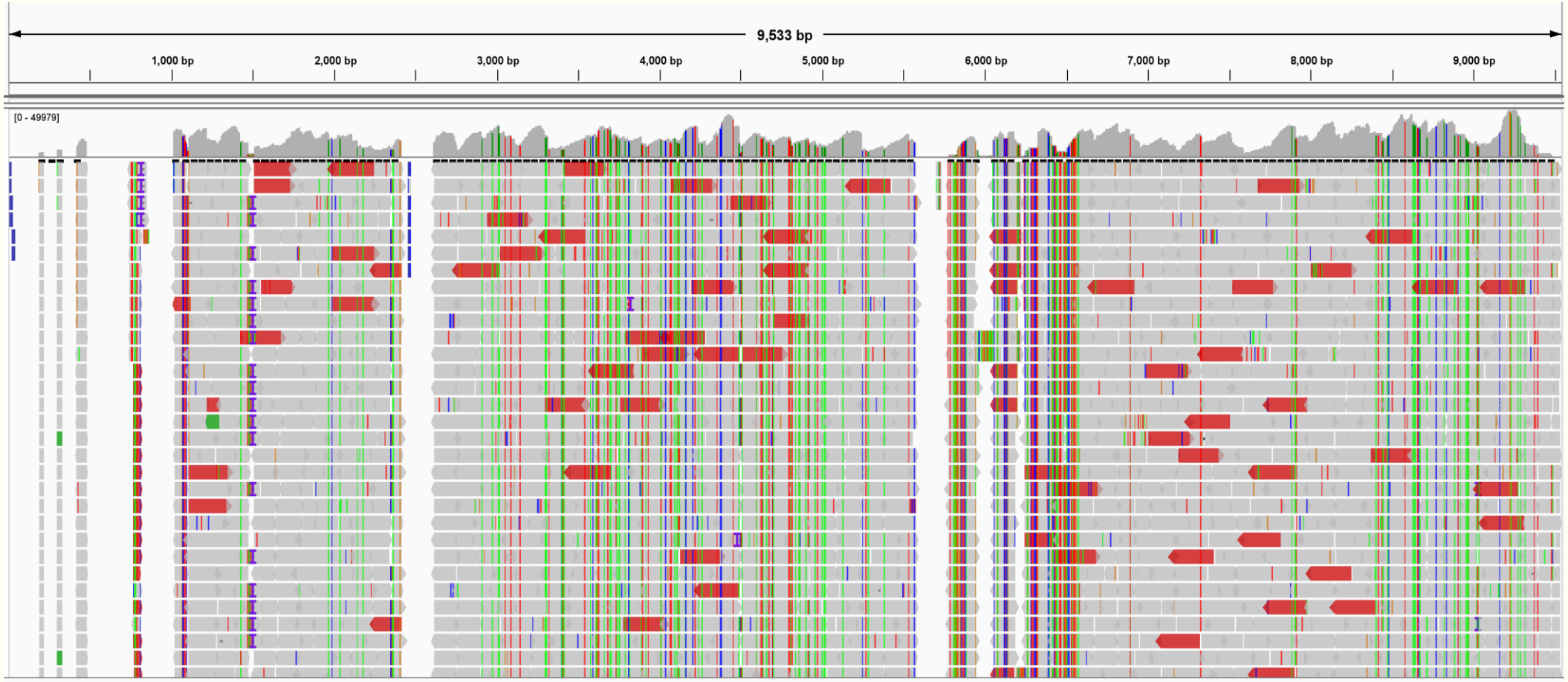
652,837 (18,73%) were duplicated.

Results: Mapping

Initial alignment: samtools flagstat report

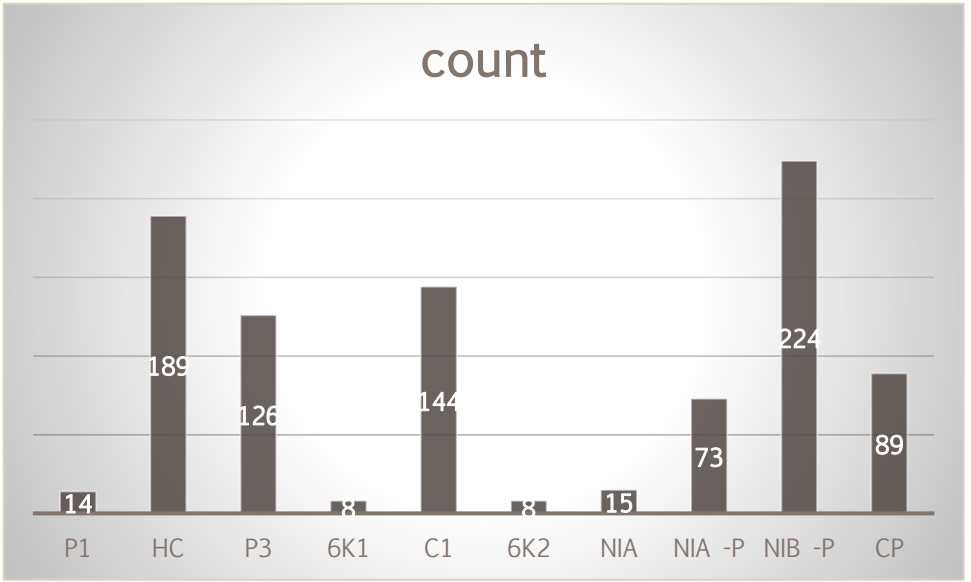
```
3485046 + 0 in total (QC-passed reads + QC-failed reads)
3484180 + 0 primary
0 + 0 secondary
866 + 0 supplementary
0 + 0 duplicates
0 + 0 primary duplicates
940779 + 0 mapped (26.99% : N/A)
939913 + 0 primary mapped (26.98% : N/A)
3484180 + 0 paired in sequencing
1742090 + 0 read1
1742090 + 0 read2
935442 + 0 properly paired (26.85% : N/A)
939496 + 0 with itself and mate mapped
417 + 0 singletons (0.01% : N/A)
0 + 0 with mate mapped to a different chr
0 + 0 with mate mapped to a different chr (mapQ>=5)
```

Alignment to consensus sequence: IGV



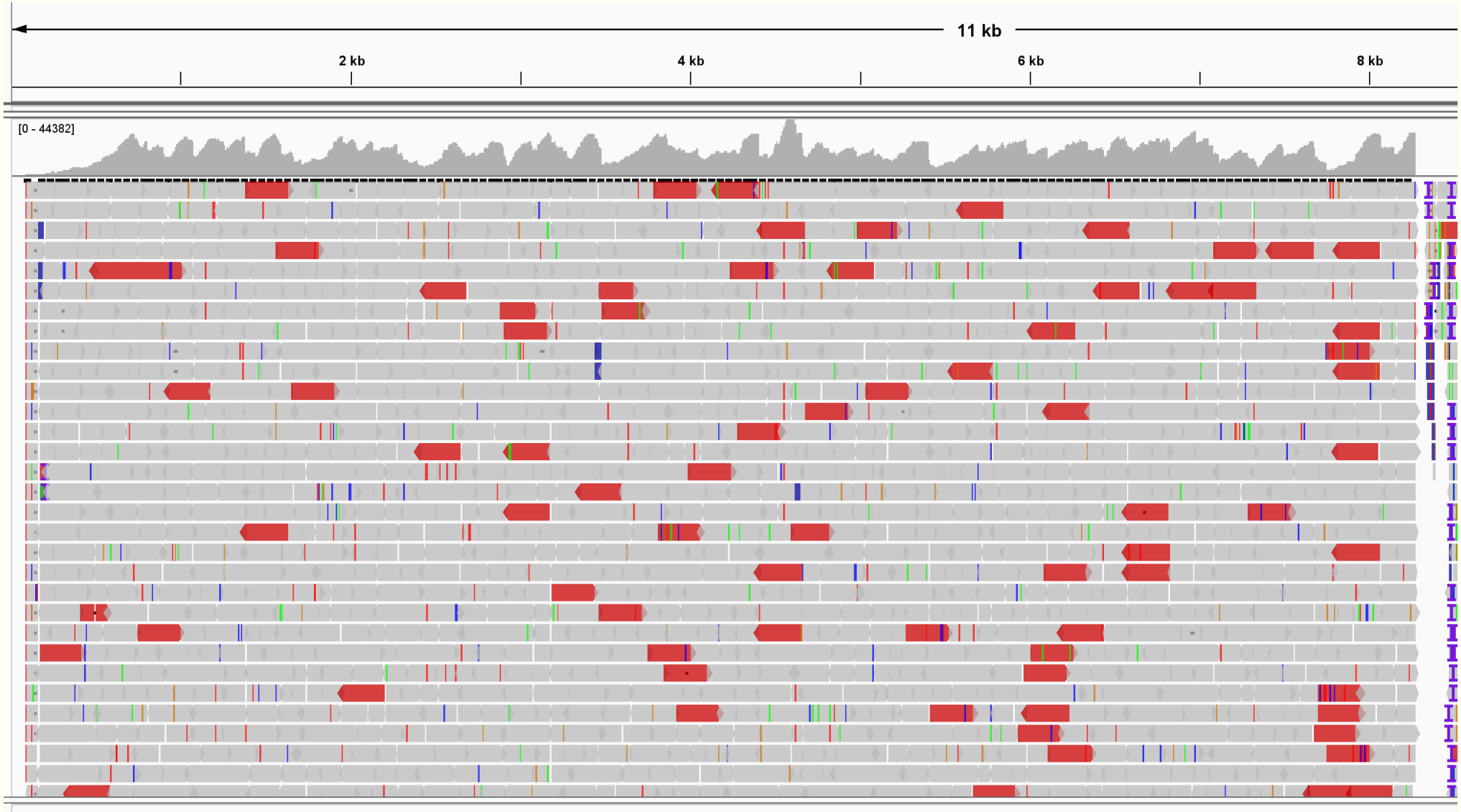
Workflow

191- 1042	P1
1043 - 2413	HC
2414 - 3457	P3
3458 - 3616	6K1
3617 - 5521	CI
5522 - 5680	6K2
5681 - 6253	NIa
6254 - 6982	Nia- p
6983 – 8539	Nib-p
8540 - 9358	CP



SNPs	count
P1	14
HC	189
P3	126
6K1	8
C1	144
6K2	8
Nia	15
Nia -p	73
Nib -p	224
CP	89

Alignment to a spades contig: IGV



Alignment to a spades contig: samtools flagstat report

```
3484959 + 0 in total (QC-passed reads + QC-failed reads)
3484178 + 0 primary
0 + 0 secondary
781 + 0 supplementary
0 + 0 duplicates
0 + 0 primary duplicates
1310052 + 0 mapped (37.59% : N/A)
1309271 + 0 primary mapped (37.58% : N/A)
3484178 + 0 paired in sequencing
1742089 + 0 read1
1742089 + 0 read2
1298664 + 0 properly paired (37.27% : N/A)
1307144 + 0 with itself and mate mapped
2127 + 0 singletons (0.06% : N/A)
0 + 0 with mate mapped to a different chr
0 + 0 with mate mapped to a different chr (mapQ>=5)
```

QUAST- Report

QUAST

Quality Assessment Tool for Genome Assemblies

25 June 2025, Wednesday, 17:25:20

View in Icarus contig browser

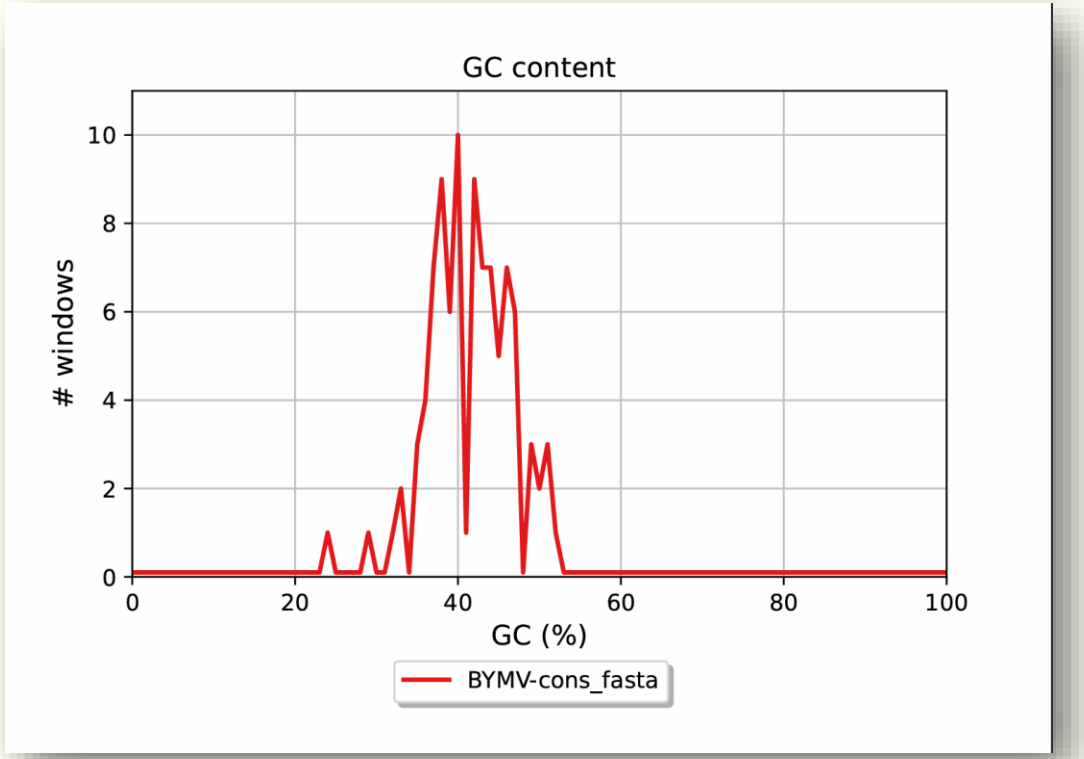
All statistics are based on contigs of size >= 500 bp, unless otherwise noted (e.g., "# contigs (>= 0 "Total length (>= 0 bp)" include all contigs).

Statistics without reference	Contig-BYMV_fa
# contigs	1247
# contigs (>= 0 bp)	433 675
# contigs (>= 1000 bp)	49
Largest contig	1807
Total length	804 160
Total length (>= 0 bp)	43 182 547
Total length (>= 1000 bp)	60 057
N50	619
N90	513
auN	688
L50	513
L90	1089
GC (%)	45.51
Per base quality	
# N's per 100 kbp	0
# N's	0

QUAST- Report

Report

	BYMV-cons_fasta
# contigs (>= 0 bp)	1
# contigs (>= 1000 bp)	1
Total length (>= 0 bp)	9532
Total length (>= 1000 bp)	9532
# contigs	1
Largest contig	9532
Total length	9532
GC (%)	41.57
N50	9532
N90	9532
auN	9532.0
L50	1
L90	1
# N's per 100 kbp	0.00



All statistics are based on contigs of size >= 500 bp, unless otherwise noted (e.g., "# contigs (>= 0 bp)" and "Total length (>= 0 bp)" include all contigs).

Final Results: Mapping

Initial Alignment:

Mapped Reads: 26.98%

Post Duplicate Removal:

Duplicates Removed: 18.37%

Mapped Reads: 37.59%

Alignment to Consensus Sequence:

Mapped Reads:

Alignment to SPAdes Contig:

Mapped Reads: 37.59%

Conclusion

- The initial quality assessment revealed a high duplication rate and possible contamination, as evidenced by FastQC and MultiQC reports. Trimming significantly improved overall data quality by reducing adapter contamination and low-quality reads, with over 97% of the reads retained.
- Upon aligning the reads to the BYMV reference genome, only 26.98% successfully mapped, and coverage was uneven, with gaps in several regions—suggesting divergence between the sample and reference genome. After removing duplicate reads and optimizing the alignment strategy, mapping efficiency improved to 37.59%, and more uniform coverage was observed.
- De novo assembly followed by QUAST and BLAST analyses provided further insights into genome structure, confirming homology with known viral sequences. However, the presence of coverage gaps and variable GC content highlights potential sequence novelty or contamination.
- Further analyses, including SNP calling and RNA-Seq, would be essential to characterize genetic variation and assess viral gene expression comprehensively.

WE ACKNOWLEDGE THE USE OF GALAXY EUROPE FOR PROVIDING THE COMPUTATIONAL INFRASTRUCTURE USED IN THIS ANALYSIS.

July 2025