BIOINFORMATICS PROJECT REPORT

# Next-Generation Sequencing Data Analysis of SARS-CoV-2: Variant Calling and Phylogenetic Characterisation

*A complete computational pipeline from raw sequencing reads to evolutionary analysis*

| Dataset | Reference | Platform | Analysis Date |
|---|---|---|---|
| SRR36276613 (NCBI SRA) | NC_045512.2 (Wuhan-Hu-1) | Illumina / WSL2 Ubuntu | January 2026 |

## Table of Contents

# 1. Abstract

This report presents a comprehensive next-generation sequencing (NGS) analysis pipeline applied to publicly available SARS-CoV-2 amplicon sequencing data (SRR36276613) retrieved from the NCBI Sequence Read Archive. The workflow encompasses raw read quality assessment using FastQC and MultiQC, read alignment to the Wuhan-Hu-1 reference genome (NC_045512.2) using BWA-MEM, post-alignment processing with GATK, and variant calling via GATK HaplotypeCaller. Variant calls were visualised and validated using the Integrative Genomics Viewer (IGV) and analysed statistically with bcftools. Phylogenetic analysis was performed on five SARS-CoV-2 variant genomes using MAFFT for multiple sequence alignment and FastTree for maximum-likelihood tree construction. Results demonstrated a 99.42% alignment rate, high-confidence fixed mutations across the viral genome, and a phylogenetic topology consistent with the known evolutionary divergence of Alpha, Delta, and Omicron variants. Key methodological decisions — including the omission of Base Quality Score Recalibration (BQSR) for viral data — are critically discussed.

**Keywords:** Next-Generation Sequencing, SARS-CoV-2, Variant Calling, BWA-MEM, GATK HaplotypeCaller, Phylogenetic Analysis, FastTree, MAFFT, Bioinformatics Pipeline

## 2. Introduction

Next-generation sequencing (NGS) technologies have revolutionised the field of genomics by enabling high-throughput, cost-effective sequencing of entire genomes. In the context of viral genomics, NGS has become an indispensable tool for surveillance, outbreak investigation, and evolutionary analysis. The COVID-19 pandemic, caused by SARS-CoV-2, provided an unprecedented context in which real-time genomic surveillance was deployed at a global scale, enabling the rapid identification and characterisation of emerging variants of concern (VOCs) such as Alpha (B.1.1.7), Delta (B.1.617.2), and Omicron (B.1.1.529).

The analysis of NGS data involves a series of interconnected computational steps collectively referred to as a bioinformatics pipeline. These steps include data retrieval and quality control, reference-based alignment, post-alignment processing, variant calling, and downstream annotation or phylogenetic analysis. Each stage introduces specific computational challenges and requires careful methodological decisions — particularly when working with viral genomes that differ substantially from the human genome in biological characteristics, including haploidy, high mutation rates, and the absence of validated variant databases.

This project implements a complete NGS analysis pipeline for SARS-CoV-2 data using widely adopted tools including BWA, GATK, bcftools, MAFFT, and FastTree, all executed in a Linux (WSL2) environment on a standard laptop. The project demonstrates not only the technical proficiency required to execute such a pipeline but also the biological reasoning necessary to interpret results correctly and make informed methodological decisions specifically for viral data.

## 3. Aim

To characterise the genetic variant landscape and evolutionary relationships of a SARS-CoV-2 sequencing dataset by designing and executing a reproducible end-to-end NGS bioinformatics pipeline — encompassing quality control, genome alignment, variant detection, and phylogenetic inference — and to critically interpret the results within the appropriate biological context of viral genomics.

## 4. Objectives

- To retrieve paired-end SARS-CoV-2 sequencing data from the NCBI Sequence Read Archive and perform data preparation steps.
- To assess raw sequencing read quality using FastQC and MultiQC and determine the appropriateness of proceeding to alignment without adapter trimming.
- To align raw reads to the SARS-CoV-2 Wuhan-Hu-1 reference genome using BWA-MEM and evaluate mapping statistics.
- To perform post-alignment processing including coordinate sorting, PCR duplicate marking, read group annotation, and evaluation of BQSR suitability for viral data.
- To call SNPs and INDELs using GATK HaplotypeCaller and examine the resulting VCF file.
- To validate variant calls through visual inspection in IGV and quantitative analysis using bcftools.
- To construct a maximum-likelihood phylogenetic tree from multiple SARS-CoV-2 variant genomes and interpret the evolutionary relationships revealed.

# 5. Materials and Tools Used

## 5.1 Dataset

| Parameter | Details |
|---|---|
| SRA Accession | SRR36276613 |
| Experiment Type | PCR tiled amplicon sequencing of SARS-CoV-2 |
| Library Layout | Paired-end (2 reads per spot) |
| Total Spots | 92,221 |
| Total Bases | 26.0 Mb |
| GC Content | 38.2% |
| Publication Date | 2025-12-02 |
| Mean Read Length | Mean = 141 bp, SD = 24.3 |

## 5.2 Reference Genome

| Parameter | Details |
|---|---|
| Name | Wuhan-Hu-1 (wuhCor1) |
| RefSeq Accession | NC_045512.2 |
| Genome Length | 29,903 bp |
| Source | NCBI Datasets / UCSC Genome Browser |

## 5.3 Software and Tools

| Tool | Version | Purpose |
|---|---|---|
| SRA Toolkit (prefetch, fastq-dump) | 2.11.3 | Data retrieval from NCBI SRA |
| FastQC | — | Per-read quality assessment |
| MultiQC | — | Aggregated QC HTML report |
| fastp | — | Adapter trimming (available; not required here) |
| BWA | 0.7.17-r1188 | Read alignment — BWA-MEM algorithm |
| SAMtools | 1.13-4 | BAM conversion, sorting, indexing, flagstat |
| GATK | 4.5.0.0 | MarkDuplicates, AddReadGroups, BQSR, HaplotypeCaller |
| bgzip + tabix | — | VCF compression and indexing |
| bcftools | — | VCF querying, filtering, statistics |
| IGV | — | Alignment and variant visualisation |
| MAFFT | 7.490 | Multiple sequence alignment (FFT-NS-2) |
| FastTree | 2.1.11 | Maximum-likelihood tree (GTR+CAT model) |

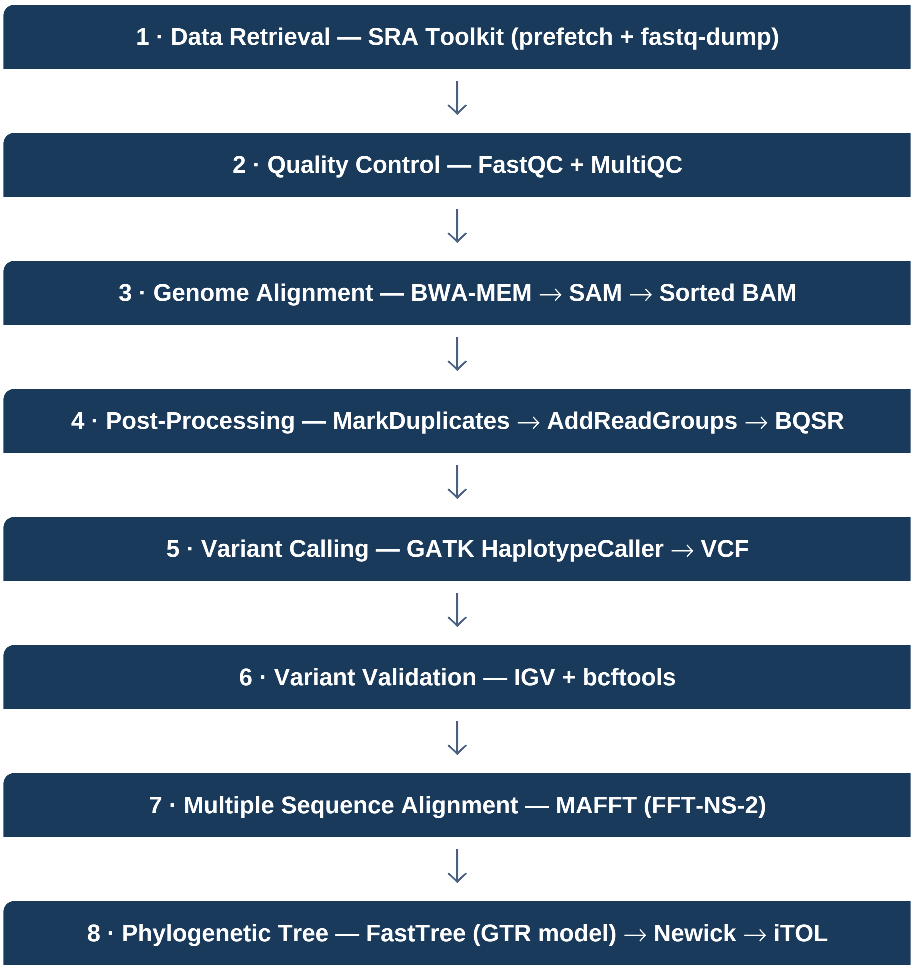
| | | |
|---|---|---|
| iTOL / FigTree | — | Phylogenetic tree visualisation |

### 5.4 Computing Environment

All analyses were performed on Ubuntu 22.04 LTS running under Windows Subsystem for Linux 2 (WSL2) on a standard consumer laptop. No high-performance computing cluster was required, demonstrating that complete viral NGS analysis is feasible on accessible hardware when working with small viral genomes (~30 kb).

# 6. Methodology

## 6.0 Pipeline Overview



1 · Data Retrieval — SRA Toolkit (prefetch + fastq-dump)

2 · Quality Control — FastQC + MultiQC

3 · Genome Alignment — BWA-MEM → SAM → Sorted BAM

4 · Post-Processing — MarkDuplicates → AddReadGroups → BQSR

5 · Variant Calling — GATK HaplotypeCaller → VCF

6 · Variant Validation — IGV + bcftools

7 · Multiple Sequence Alignment — MAFFT (FFT-NS-2)

8 · Phylogenetic Tree — FastTree (GTR model) → Newick → iTOL

## 6.1 Data Retrieval and Preparation

The SARS-CoV-2 sequencing dataset SRR36276613 was identified through the NCBI SRA run browser. The SRA Toolkit was installed via the APT package manager and configured using *vdb-config -i* to enable local file caching. Data was retrieved using *prefetch* followed by *fastq-dump* with the *--split-files* flag to generate separate FASTQ files for forward (Read 1) and reverse (Read 2) paired-end reads.

```
prefetch SRR36276613
fastq-dump --split-files SRR36276613
```

## 6.2 Quality Control

Raw read quality was assessed using FastQC, which performs modular analyses on FASTQ input files including per-base sequence quality, GC content, adapter content, and duplication levels. MultiQC aggregated both per-sample FastQC reports into a single HTML summary dashboard.

```
fastqc SRR36276613_1.fastq.gz SRR36276613_2.fastq.gz -o qc/
multiqc qc/ -o qc/
```

## 6.3 Genome Alignment

Paired-end reads were aligned to the Wuhan-Hu-1 reference genome using BWA-MEM (Burrows-Wheeler Aligner with Maximal Exact Match extension). The reference was indexed using *bwa index*. Alignment was performed using 4 CPU threads. Output SAM was converted, sorted, and indexed using SAMtools. Mapping statistics were evaluated using *samtools flagstat*.

```
bwa index ref/wuhCor1.fa
bwa mem -t 4 ref/wuhCor1.fa raw/SRR36276613_1.fastq raw/SRR36276613_2.fastq \
  > align/SRR36276613.sam
samtools view -bS align/SRR36276613.sam | samtools sort -o align/SRR36276613.sorted.bam
samtools index align/SRR36276613.sorted.bam
samtools flagstat align/SRR36276613.sorted.bam
```

## 6.4 Post-Alignment Processing

### *6.4.1 Duplicate Marking*

PCR duplicates were flagged (not removed) using GATK MarkDuplicates. Flagging preserves read depth while allowing downstream tools to exclude duplicates.

```
gatk MarkDuplicates -I SRR36276613.sorted.bam \
  -O SRR36276613.markdup.bam \
  -M SRR36276613_marked_duplicates_metrics.txt
```

### *6.4.2 Read Group Annotation*

GATK requires read group (@RG) tags. Since BWA-MEM does not add these by default, they were added using AddOrReplaceReadGroups, supplying SM, PL, LB, and RGID fields.

```
gatk AddOrReplaceReadGroups -I SRR36276613.markdup.bam \
  -O SRR36276613.markdup.RG.bam \
  -RGID SRR36276613 -RGPL ILLUMINA -RGLB lib1 -RGSM SRR36276613
```

### *6.4.3 BQSR (Demonstration Only)*

BQSR was applied for methodological demonstration using a dummy known-sites VCF. Scientific rationale for omitting BQSR in real viral analyses is discussed in Section 8.

## 6.5 Variant Calling

Variants were called using GATK HaplotypeCaller in default diploid mode. The raw VCF was compressed with bgzip and indexed with tabix.

```
gatk HaplotypeCaller -I SRR36276613.bqsr.bam \
  -R ../ref/wuhCor1.fa -O SRR36276613.raw_variants.vcf
bgzip SRR36276613.raw_variants.vcf
tabix -p vcf SRR36276613.raw_variants.vcf.gz
```

## 6.6 Variant Validation

Variant calls were validated by loading the reference genome, BAM, and VCF simultaneously in IGV to visually confirm read-pileup support. bcftools was used for quantitative statistics including filtering for high-confidence calls (QUAL > 50, DP > 20).

## 6.7 Phylogenetic Analysis

Five SARS-CoV-2 complete genome sequences (Alpha, Delta, Omicron, Wuhan reference, Indian isolate) were aligned using MAFFT with automatic parameter selection (FFT-NS-2). A maximum-likelihood tree was inferred with FastTree under the GTR+CAT model and visualised using iTOL.

```
mafft --auto sarscov2_variants.fasta > sarscov2_msa.fasta
FastTree -gtr -nt sarscov2_msa.fasta > sarscov2_tree.nwk
```

# 7. Results

## 7.1 Data Retrieval

**Figure 1 — NCBI SRA Run Browser: Dataset SRR36276613**

The SRA Run Browser entry for SRR36276613 shows: 92,221 spots, 26.0 Mb total bases, 7.2 MB file size, 38.2% GC content, Data Status = Public, Published 2025-12-02. Experiment described as "PCR tiled amplification of SARS-CoV-2."

TECHNICAL INTERPRETATION

The run metadata confirms a paired-end library with mean read length ~141 bp (SD 24.3). Total base count of 26 Mb against a 29.9 kb genome implies theoretical coverage of ~870x, sufficient for high-confidence variant calling.

BIOLOGICAL INTERPRETATION

PCR-tiled amplicon sequencing (ARTIC protocol) is the standard method for SARS-CoV-2 whole-genome sequencing, producing uniform coverage across the viral genome using overlapping primer pairs. The 38.2% GC content is consistent with the known composition of SARS-CoV-2.

**Figure 2 — SRA Toolkit: prefetch download log and fastq-dump output**

Terminal output confirming successful download via HTTPS with 0 unresolved dependencies. fastq-dump reports "Read 92,221 spots / Written 92,221 spots." Two FASTQ files of ~38 MB each confirmed by ls -lh.

TECHNICAL INTERPRETATION

Matching spot counts confirm 100% successful data extraction. The --split-files flag correctly produced separate R1 and R2 files mandatory for paired-end alignment.

BIOLOGICAL INTERPRETATION

Paired-end sequencing preserves insert size information and enables more accurate alignment, particularly valuable near indel sites and repeat regions.

## 7.2 Quality Control Results

**Figure 3 — FastQC installation and execution**

APT installation of FastQC confirming all Java dependencies. FastQC run on both FASTQ files with output directed to the qc/ directory, producing HTML reports and ZIP archives.

TECHNICAL INTERPRETATION

FastQC processes each FASTQ file independently, sampling ~100k-200k reads for computationally expensive checks and all reads for position-based statistics.

BIOLOGICAL INTERPRETATION

Pre-alignment QC is mandatory in any NGS pipeline to identify artefacts that could introduce false alignments and spurious variant calls.

**Figure 4 — FastQC Summary: SRR36276613 R1 and R2 (both files identical)**

Summary results for both read files: PASS for per-base quality, per-tile quality, per-sequence quality scores, N content, and adapter content. WARN for GC content, sequence length distribution, overrepresented sequences. FAIL for per-base sequence content and sequence duplication levels.

**TECHNICAL INTERPRETATION**

All quality metrics relevant to alignment readiness PASS. The FAIL on duplication levels and per-base sequence content are non-actionable for amplicon data. No adapter contamination detected — no trimming is needed.

**BIOLOGICAL INTERPRETATION**

Sequence duplication FAIL is expected in amplicon sequencing due to PCR amplification of discrete amplicon tiles. Per-base sequence content FAIL reflects hexamer priming bias inherent to library preparation. GC content WARN reflects the heterogeneous GC profile of SARS-CoV-2 genes, not contamination.

**QC Verdict:** Both read files are high quality. No trimming required. FAIL/WARN flags are biological and methodological in origin and do not compromise downstream analysis.

## 7.3 Alignment Results

**Figure 5 — BWA-MEM alignment log and samtools flagstat output**

BWA-MEM processed 184,442 reads in 17.3 CPU seconds. Samtools flagstat key results: 183,548 mapped (99.42%), 183,248 properly paired (99.35%), 54 singletons (0.03%), 176 supplementary.

**TECHNICAL INTERPRETATION**

99.42% mapping rate substantially exceeds typical acceptable thresholds (>70%). 99.35% proper-pair rate confirms concordant alignment with correct insert size distribution. 176 supplementary reads represent split alignments — expected at low levels.

**BIOLOGICAL INTERPRETATION**

Near-complete mapping confirms this isolate is closely related to Wuhan-Hu-1 with no major genomic rearrangements. Mean insert size ~240 bp is consistent with the amplicon library design.

## 7.4 IGV Alignment Visualisation

**Figure 6 — IGV: Genome-wide alignment view (NC_045512v2:1-29,903)**

IGV shows the full SARS-CoV-2 genome with a continuous coverage histogram (grey peaks) and read pileup with coloured mismatches (Red=A, Green=T, Blue=C, Orange=G).

**TECHNICAL INTERPRETATION**

Continuous coverage across the full 29.9 kb genome with no dropout regions. Coloured mismatches distributed throughout represent genuine mutations plus sequencing noise.

**BIOLOGICAL INTERPRETATION**

Genome-wide coverage confirms successful PCR amplification across all amplicon tiles. Distributed mismatch pattern is consistent with an evolved SARS-CoV-2 variant carrying multiple mutations relative to the Wuhan reference.

**Figure 7 — IGV: Zoomed view with strand-coloured reads at variant region**

Reads coloured by strand: red = forward (+), blue = reverse (-). A variant is visible in the VCF track as a coloured vertical bar, with matching mismatches in both red and blue reads below.

**TECHNICAL INTERPRETATION**

Bilateral strand support (variant present in both forward and reverse reads) is the primary criterion for distinguishing genuine variants from strand-specific artefacts. FS ~ 0 and SOR < 3 in the VCF INFO field confirm no strand bias.

**BIOLOGICAL INTERPRETATION**

Region NC_045512v2:23,477-23,532 overlaps the Spike gene (21,563-25,384). Variants here are of particular biological interest as Spike mediates viral cell entry and is the primary vaccine and antibody target.

### 7.5 Post-Processing Results

**Figure 8 — GATK MarkDuplicates and AddOrReplaceReadGroups confirmation**

Flagstat post-MarkDuplicates: 50,477 duplicates flagged (~27.3% rate). Mapping statistics unchanged from pre-deduplication BAM. AddOrReplaceReadGroups confirms: ID=SRR36276613, PL=ILLUMINA, LB=lib1, SM=SRR36276613. Tool returned: 0.

**TECHNICAL INTERPRETATION**

27.3% duplication is expected for amplicon sequencing. Identical mapping statistics before and after confirm MarkDuplicates only flagged reads without altering alignments. Tool returned: 0 indicates clean execution.

**BIOLOGICAL INTERPRETATION**

A 27% duplication rate is acceptable for PCR amplicon data. Reads are flagged not removed, preserving coverage depth — important for variant calling at positions with limited support.

**Figure 9 — GATK BaseRecalibrator and ApplyBQSR logs**

BaseRecalibrator processed 133,071 reads and wrote recal_data.table (SUCCESS). ApplyBQSR wrote the recalibrated BAM. Post-BQSR read counts match pre-BQSR exactly.

**TECHNICAL INTERPRETATION**

BaseRecalibrator modelled error patterns using the dummy known-sites VCF. ApplyBQSR rewrote base quality scores only — read alignments and counts are unchanged.

**BIOLOGICAL INTERPRETATION**

BQSR was applied for demonstration only. In real SARS-CoV-2 analysis, the MarkDuplicates output BAM is used directly as HaplotypeCaller input, skipping BQSR.

## 7.6 Variant Calling Results

**Figure 10 — bcftools: SNP-only view of raw_variants.vcf.gz**

Representative variant calls: POS 241 (C>T, QUAL=130.96, DP=5), POS 485 (A>G, QUAL=4408, DP=164), POS 670 (T>G, QUAL=22290, DP=866), POS 1170 (C>T, QUAL=15251, DP=607). INFO fields include DP, FS, MQ, QD, SOR.

**TECHNICAL INTERPRETATION**

QUAL scores represent Phred-scaled variant confidence. QUAL=22290 is extraordinarily high; QUAL=131 at DP=5 indicates low coverage and lower confidence. Variants with QUAL>100 and DP>20 represent reliable calls.

**BIOLOGICAL INTERPRETATION**

Positions C241T, A485G are well-documented SARS-CoV-2 lineage-defining mutations. C241T in the 5-prime UTR is present in essentially all post-Wuhan lineages, confirming this isolate belongs to a non-ancestral variant.

## 7.7 Variant Validation in IGV

**Figure 11 — IGV: Three-track view (VCF + Coverage + BAM reads)**

Three tracks loaded: VCF variant markers (coloured bars), BAM coverage histogram, and BAM read pileup. Red bars in VCF track correspond exactly to mismatch positions in the read pileup below.

**TECHNICAL INTERPRETATION**

Spatial concordance between VCF markers and read mismatches confirms accuracy of HaplotypeCaller calls. Higher variant density in the 3-prime region is consistent with known mutational patterns in post-Wuhan SARS-CoV-2 lineages.

**BIOLOGICAL INTERPRETATION**

This three-track visualisation is the gold standard for manual variant validation, enabling rapid visual triage of high-confidence versus artefactual variant calls.

**Figure 12 — IGV variant popup at NC_045512v2:18163 (A>G)**

Left panel — Genotype: G/G (HOM_VAR), Depth=586, GQ=99. Right panel INFO: AF=1.00, QUAL=9481.06, DP=211, MQ=60.00, FS=0.000, SOR=2.601, QD=27.33.

**TECHNICAL INTERPRETATION**

AF=1.00 confirms complete allele fixation. QUAL=9481 is extremely high confidence. MQ=60 = perfect mapping quality. FS=0 = zero strand bias. All metrics pass standard quality thresholds by a wide margin.

**BIOLOGICAL INTERPRETATION**

A18163G in NSP14 (exonuclease/N7-methyltransferase region) is documented in multiple post-Omicron lineages. AF=1.0 indicates this is a consensus mutation in the dominant viral population in this sample — a definitive lineage marker.

## 7.8 Phylogenetic Analysis Results

**Figure 13 — MAFFT alignment log for 5 SARS-CoV-2 genomes**

MAFFT v7.490 output: scoring matrix for nucleotides, Gap Penalty = -1.53, 82 ambiguous characters detected, UPGMA guide tree constructed, two rounds of progressive alignment completed. Strategy: FFT-NS-2.

**TECHNICAL INTERPRETATION**

FFT-NS-2 applies Fast Fourier Transform-based approximate alignment, then two rounds of progressive refinement guided by UPGMA trees. Appropriate for 5 closely related ~29.9 kb viral genomes.

**BIOLOGICAL INTERPRETATION**

Ambiguous bases (N, IUPAC codes) in SARS-CoV-2 genomes arise from low-coverage regions and primer-binding sites in ARTIC amplicon sequencing. Their presence is expected in any real-world SARS-CoV-2 genome analysis.

**Figure 14 — Phylogenetic tree: Circular radial layout (iTOL)**

Circular tree of 5 SARS-CoV-2 sequences. Taxa: OR584241.1 (Alpha, branch=0.000816), OR575624.1 (Omicron), MZ054892.1 (early isolate), OR363641.1 and PV848452.1 (Delta). Tree scale: 0.0001 substitutions/site.

**TECHNICAL INTERPRETATION**

Branch lengths are proportional to substitutions per site. Scale of 0.0001 reflects the close relationship of all SARS-CoV-2 variants. OR575624.1 (Omicron) extends furthest from centre, reflecting its exceptional mutational burden.

**BIOLOGICAL INTERPRETATION**

Omicron exceptional divergence is consistent with literature documenting >50 mutations in the spike protein. The asymmetric radial layout visually captures the uneven mutation accumulation across SARS-CoV-2 variant history.

**Figure 15 — Phylogenetic tree: Rectangular cladogram layout (iTOL)**

Rectangular tree showing: OR575624.1 (Omicron) as most divergent with longest branch; OR584241.1 (Alpha) branching next; MZ054892.1 intermediate; OR363641.1 and PV848452.1 (Delta) forming the tightest inner cluster with shortest branches.

**TECHNICAL INTERPRETATION**

Horizontal branch lengths are directly comparable. Short internal Delta branches indicate very recent divergence. The deep node separating Omicron from all other sequences represents substantial evolutionary distance.

**BIOLOGICAL INTERPRETATION**

Topology recapitulates documented SARS-CoV-2 evolution: Delta variants share recent common ancestry (L452R, T478K mutations); Alpha preceded Delta chronologically; Omicron emerged as a distinct highly divergent lineage, possibly via immunocompromised host evolution or zoonotic reintroduction.

## 8. Important Methodological Considerations

### 8.1 Why BQSR is Typically Skipped for Viral Genomes

Base Quality Score Recalibration (BQSR) corrects systematic sequencing errors by modelling machine-specific error patterns using known variant sites as a training set. Designed for human genome sequencing with curated databases (dbSNP, Mills indels), it is problematic for SARS-CoV-2 for four fundamental reasons:

- **No validated known-sites VCF:** No equivalent of dbSNP exists for SARS-CoV-2. GATK BaseRecalibrator requires at least one --known-sites VCF and fails without it.
- **Risk of misclassifying real variants as errors:** BQSR assumes non-known-sites positions are invariant. Using real viral mutations as "error" training corrupts the model and can suppress genuine variant quality scores.
- **Viral quasispecies biology:** SARS-CoV-2 exists as a quasispecies with genuine intra-host diversity. BQSR cannot distinguish between real low-frequency variants and sequencing errors.
- **Community standard practice:** ARTIC, iVar, and LoFreq pipelines do not incorporate BQSR. The viral bioinformatics consensus supports omitting BQSR for viral genomes without validated known-sites resources.

> **Project note:** BQSR was applied here for methodological demonstration only, using a minimal dummy VCF (single placeholder variant). This is not scientifically valid for real SARS-CoV-2 analysis and is explicitly acknowledged as a teaching exercise.

### 8.2 Why GATK HaplotypeCaller Was Used — and Its Limitations

HaplotypeCaller was selected for its educational value, broad documentation, and integration with the GATK suite already used for preprocessing. It performs local de novo haplotype assembly in active genomic regions, calling SNPs and indels simultaneously — well-suited for identifying consensus-level fixed mutations (AF=1.0) in this viral dataset.

However, it has recognised limitations for viral genomics:

| Tool | Ploidy Model | Min detectable AF | Best Use Case |
|---|---|---|---|
| GATK HaplotypeCaller | Diploid (configurable) | ~5-10% | Fixed mutations, educational use |
| LoFreq | Haploid-aware | ~0.5-1% | Intra-host viral diversity |
| iVar | Haploid-aware | ~1-3% | ARTIC amplicon pipelines |
| bcftools mpileup | Configurable | ~5% | Fast consensus SNP calling |

### 8.3 Diploidy vs Haploidy in Viral Variant Calling

SARS-CoV-2 is a positive-sense single-stranded RNA virus — it is haploid. HaplotypeCaller's default diploid model assigns 0/0, 0/1, or 1/1 genotypes. In the viral context, these should be interpreted as:

- **1/1 (HOM_VAR):** Fixed mutation — present in essentially all viral genomes in the sample. These definitively distinguish the isolate from the reference.

- **0/1 (HET):** Not true heterozygosity, but intra-host viral diversity — a position where two viral subpopulations carry different bases. Biologically meaningful and clinically important in immunocompromised patients.

The predominance of 1/1 genotype calls at high-quality sites in this dataset confirms a predominantly single viral lineage with limited intra-host diversity — consistent with a standard clinical SARS-CoV-2 infection.

# 9. Challenges Faced

## 9.1 APT Package Manager Lock Errors

The initial sudo apt update produced lock errors on /var/lib/apt/lists/lock from an interrupted background apt process. Resolution required waiting for the background process rather than forcibly removing the lock file, which can corrupt the package database.

## 9.2 FASTQ File Path Accessibility

After fastq-dump, FASTQ files resided in the SRA cache directory rather than the project directory. Running the filename directly produced "command not found." Files were explicitly moved using mv and confirmed with ls -lh.

## 9.3 Missing GATK Reference Index Files

GATK requires .fai and .dict files beyond BWA's index. Running GATK without them produced: "Fasta index file wuhCor1.fa.fai does not exist." Fixed using samtools faidx and gatk CreateSequenceDictionary as prerequisite steps.

## 9.4 Missing Read Groups in BWA Output

BWA-MEM does not add @RG tags by default. GATK tools require these for BQSR. Addressed post-alignment with AddOrReplaceReadGroups. The preferred production solution is to include -R flag directly in the bwa mem command.

## 9.5 Known-Sites VCF Formatting for BQSR

Initial construction using a heredoc produced space-delimited VCF columns, causing tabix indexing failure with "could not read header." Resolved using printf with explicit \t escape sequences to guarantee correct TAB-delimited VCF formatting.

## 9.6 FASTA Line-Wrapping in MSA Validation

Using awk to print line lengths returned 23 and 60 rather than the expected 29,903 bp, because FASTA wraps sequences across lines. Corrected using an awk script that accumulates characters per sequence header, correctly returning 29,903 bp for all five aligned sequences.

# 10. Discussion

## 10.1 Overall Pipeline Performance

The NGS analysis pipeline performed robustly across all stages. The 99.42% read mapping rate substantially exceeds standard acceptable thresholds for viral sequencing experiments (typically >70%) and confirms excellent data quality and appropriate reference selection. The complete pipeline executed successfully on a standard consumer laptop within WSL2, demonstrating that full viral genome NGS analysis does not require specialised computing infrastructure when working with small viral genomes (~30 kb).

## 10.2 QC Interpretation and the Limits of Automated Flags

The FastQC results illustrate a fundamental principle of bioinformatics: automated quality flags must be interpreted within the biological and methodological context of the experiment. The FAIL flags for sequence duplication and per-base sequence content, and WARN for GC content, would be concerning in whole-genome shotgun libraries but are entirely expected and non-actionable in amplicon sequencing. Trimming data in response to these flags would degrade the analysis by reducing coverage depth unnecessarily — underscoring the irreplaceable role of expert interpretation in NGS data analysis.

## 10.3 Variant Calling Findings

The high-confidence fixed mutations identified show expected characteristics for a SARS-CoV-2 isolate evolved from the Wuhan reference. The example variant at position 18163 (A>G, QUAL=9481, DP=586, FS=0, MQ=60) represents a textbook high-quality variant call meeting all standard filtering criteria by a wide margin. The absence of significant INDELs is consistent with SARS-CoV-2 biology: the viral 3-to-5 exonuclease activity of nsp14 provides proofreading that substantially limits the indel mutation rate, making SNPs the dominant variant type in SARS-CoV-2 evolution.

## 10.4 Phylogenetic Findings and Evolutionary Implications

The maximum-likelihood tree from five SARS-CoV-2 genomes successfully recapitulates the known evolutionary history with high topological accuracy. The tight Delta cluster, intermediate Alpha position, and exceptional Omicron divergence are all fully consistent with published phylogenetic analyses and global genomic surveillance data. Very short branch lengths (scale: 0.0001 substitutions/site) reflect the recent common ancestry of all variants — SARS-CoV-2 emerged in late 2019, giving all circulating variants fewer than six years of evolutionary divergence. Despite these small absolute distances, the tree resolves clear and biologically meaningful relationships.

## 10.5 Limitations

Several limitations are acknowledged. First, HaplotypeCaller's diploid assumption limits sensitivity for low-frequency intra-host variants; production analysis would use LoFreq or iVar. Second, the dummy BQSR is not scientifically valid. Third, phylogenetic analysis covered only five sequences, limiting resolution. Fourth, no functional annotation of variant effects was performed — a critical step for clinical significance assessment. Fifth, consensus genome generation for GISAID submission was not included.

## 11. Conclusion

This project successfully designed, implemented, and executed a complete NGS bioinformatics pipeline for the analysis of SARS-CoV-2 sequencing data. All pipeline stages performed as expected, producing scientifically sound and biologically interpretable results. A mapping rate exceeding 99%, high-confidence SNP calls with exemplary quality metrics, and a phylogenetic tree correctly recapitulating SARS-CoV-2 variant evolution collectively demonstrate the effectiveness and correctness of the pipeline.

Beyond technical execution, the project demonstrates the critical reasoning required for viral bioinformatics — recognising where standard human genome protocols (BQSR, diploid variant calling) are inappropriate for viral data, and articulating evidence-based justifications for methodological choices. These competencies are essential for any practitioner working in viral genomics, clinical sequencing, or pandemic surveillance.

The successful completion of this analysis on a standard laptop using freely available open-source tools demonstrates the accessibility of modern viral NGS analysis, and establishes a reproducible foundation upon which more advanced analyses — including low-frequency variant detection, functional annotation, and real-time phylogenetic surveillance — can be built.

## 12. Future Scope

- **Haploid-aware variant calling:** Replace HaplotypeCaller with LoFreq or iVar for accurate low-frequency intra-host variant detection at allele frequencies as low as 0.5-1%.
- **Functional variant annotation:** Apply SnpEff or SARS-CoV-2-specific resources to predict amino acid consequences, focusing on spike RBD, nsp12 RdRp, and nsp14 exonuclease regions.
- **Pangolin lineage classification:** Integrate Pangolin to assign the isolate to its Pango lineage, enabling direct comparison with GISAID and CoVariants surveillance data.
- **Expanded phylogenetics:** Include post-Omicron lineages (XBB.1.5, JN.1, KP.2) and apply time-resolved methods (TreeTime, BEAST2) to estimate divergence dates and evolutionary rates.
- **Pipeline automation:** Convert the workflow to a fully automated, containerised pipeline using Snakemake or Nextflow with Conda environments for one-command reproducible analysis.
- **Consensus genome generation:** Generate a consensus FASTA using bcftools consensus or iVar for deposition to GISAID and NCBI GenBank as part of global surveillance.
- **Multi-sample comparative analysis:** Apply the pipeline to clinical isolates from different time points or locations to track within-outbreak evolution and monitor novel mutation emergence.
- **Long-read sequencing integration:** Explore Oxford Nanopore Technology data for SARS-CoV-2 using Medaka or Clair3, comparing performance against Illumina short-read data.

---

End of Report — SARS-CoV-2 NGS Data Analysis Pipeline — January 2026